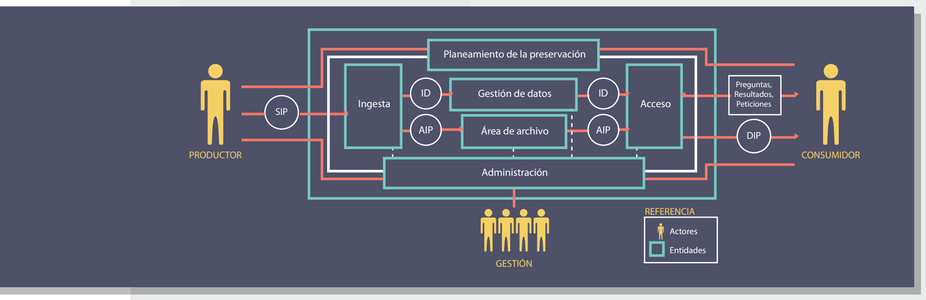
Trabajo de Tesis para la obtención del grado de Doctor en Ciencias Informáticas

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

Objetivos

Esta tesis apunta a proponer una metodología de evaluación de repositorios digitales con el fin de asegurar la **preservación**, el **acceso** y la **comprensión** de los objetos digitales a largo plazo.



MODELO OAIS ISO 14721: 2012: ISO Reference Model of an Open Archival Information System (OAIS)

> *"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".*

Amenazas a los objetos digitales

1. Su propia naturaleza los hace efímeros dado que no tiene representación en el mundo real
2. Obsolescencia tecnológica.
3. Pérdidas por desastres (naturales, accidentes, técnicos, etc)
4. Barreras de acceso en los archivos: claves, cifrado, formatos cerrados.
5. Ausencia de planes de contingencia y recuperación
6. Falta de recursos económicos para preservación a largo plazo.
7. Problemas legales por falta de permisos.
8. Descripción inadecuada que deriva en imposibilidad de recuperación.
9. Pérdida de información sobre el contexto, es decir, sobre las tecnologías utilizadas para accederlos.

Propuesta metodológica

Analizar la presencia de los distintos elementos del paquete de información (IP) de la norma del modelo OAIS ISO 14721 en los items del repositorio.

Si los contenidos están bien estructurados en cuanto a los elementos propuestos para el IP significa que el repositorio:

- tiene la funcionalidad recomendada por la norma,
- puede asegurar la preservación y
- asegura el acceso a los contenidos en el tiempo.

EL PAQUETE DE INFORMACIÓN EN EL MODELO OAIS

Las acciones propuestas están vinculadas a los elementos constitutivos del paquete de información:

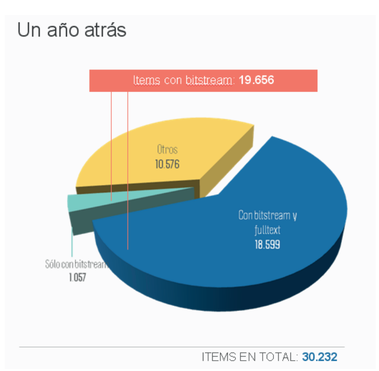
- la información de contenido (CDO),
- la información sobre la representación de ese contenido (RI),
- **la información descriptiva de preservación (PDI)** y
- la información descriptiva (DI).

Experimento

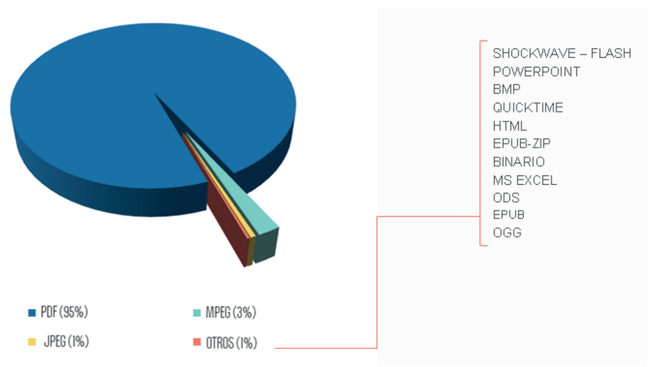
- Perfilamiento automatizado de los objetos del repositorio utilizando herramientas libres (DROID, PRONOM, entre otras): esto involucra al objeto de contenido (CDO) con sus propiedades significativas y a la información de representación de ese objeto (RI).
- Revisión de los metadatos de preservación que acompañan a los objetos digitales del repositorio y contraste con los metadatos de la PDI a través de una herramienta de validación propia.
- Revisión de la información descriptiva: contraste con las directrices DRIVER 2.0. y algunos metadatos extras.

Primer experimento *Sobre CDO*

Se realizó un primer experimento sobre alrededor de 90000 objetos considerando todos los archivos de los items.



LOCALIZACIÓN DE ARCHIVOS



FORMATOS-MIME TYPES

Algunos casos encontrados

- Archivos duplicados en diferentes ítems por carga errónea de archivos.
- Ítems sin OCR.
- Formatos de archivos desconocidos u obsoletos.
- Recursos sin archivos ni enlaces al exterior.

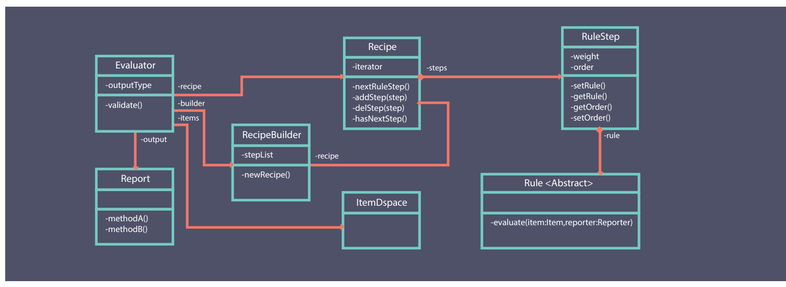
Recomendaciones

- Migrar documentos existentes a PDF/A1-a, o en su defecto PDF/A1-b.
- Generar versiones en PDF/A a partir de formatos originales: DOC, DOCX, ODT, RTF.
- Evitar la manipulación de documentos una vez depositados en el resposito-rio.
- Validar que todos los documentos tengan OCR
- Utilizar estándares abiertos siempre que sea posible.
- Almacenar y preservar por lo menos tres versiones de cada uno de los archivos ingresados al repositorio: la versión original tal y como ha sido subida, un nuevo formato normalizado y, opcionalmente, una adaptación a formatos abiertos.
- Extraer automáticamente los metadatos técnicos de los archivos subidos y guardarlos en la base de datos.

Validaciones de la PDI *Sobre PDI*

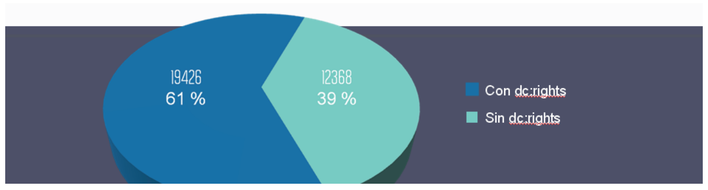
- **Referencia**: se evalúa validando los identificadores persistentes; para el caso de DSpace se evalúa el handle.
- **Integridad**: se evalúa utilizando el checksum de cada archivo.
- **Procedencia:** análisis de la trazabilidad del ítem.
- **Contexto**: revisión de la vinculación del ítem con su entorno.
- **Acceso**: se evalúan los permisos a partir de las licencias.

Validador



- Define recetas en función de reglas de validación
- Evalúa un ítem a partir de una receta
- Genera un reporte de acuerdo a la validación realizada

Ejemplo de resultados y acciones



Chequeo y resultados

- Todos los ítems del repositorio cumplen con Driver.
  Existen 8 metadatos vinculados a Subject en SEDICI: El único obligatorio es materias. Sólo un archivo no contaba con ninguno.
- En relación al metadato Description, existen dos metadatos en SEDICI: "Notas" y "Resumen" (opcionales). Resumen sólo es obligatorio en el autoarchivo de tesis.
- En la consulta del metadato resumen, se identificaron 4887 ítems sin resumen.

Análisis de la información descriptiva *Sobre DI*

- Todos los contenidos del repositorio SEDICI fueron evaluados: colecciones de revistas, artículos, tesis, imágenes, etcétera.
- La visión integral de los contenidos ha permitido detectar ítems duplicados, items sin localización física o electrónica, ítems sin resúmenes que dieran idea de su contenido, ítems sin licencia, ítems con licencias erróneas, etcétera.
- Se ha generado más de 100 tareas de revisión y corrección manual en el repositorio (algunos involucran cientos y hasta miles de ítems).

Trabajos futuros

- Análisis de factibilidad de migraciones masivas: simplicidad vs. riesgo de pérdida
- Selección de herramientas de migración.
- Registro de los eventos de migración y sus responsables de modo de asegurar la trazabilidad en el ciclo de vida (evento y agente en PREMIS).
- Definición de reglas abstractas a nivel de repositorio que permitan evaluar las características de preservación en cualquier repositorio.
- Extensión de la herramienta de validación: Implementación de un lenguaje específico de dominio que permita evaluar y transformar los ítems del repositorio a través de una sintaxis sencilla.