FACULTAD DE INFORMÁTICA
UNIVERSIDAD NACIONAL DE LA PLATA

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

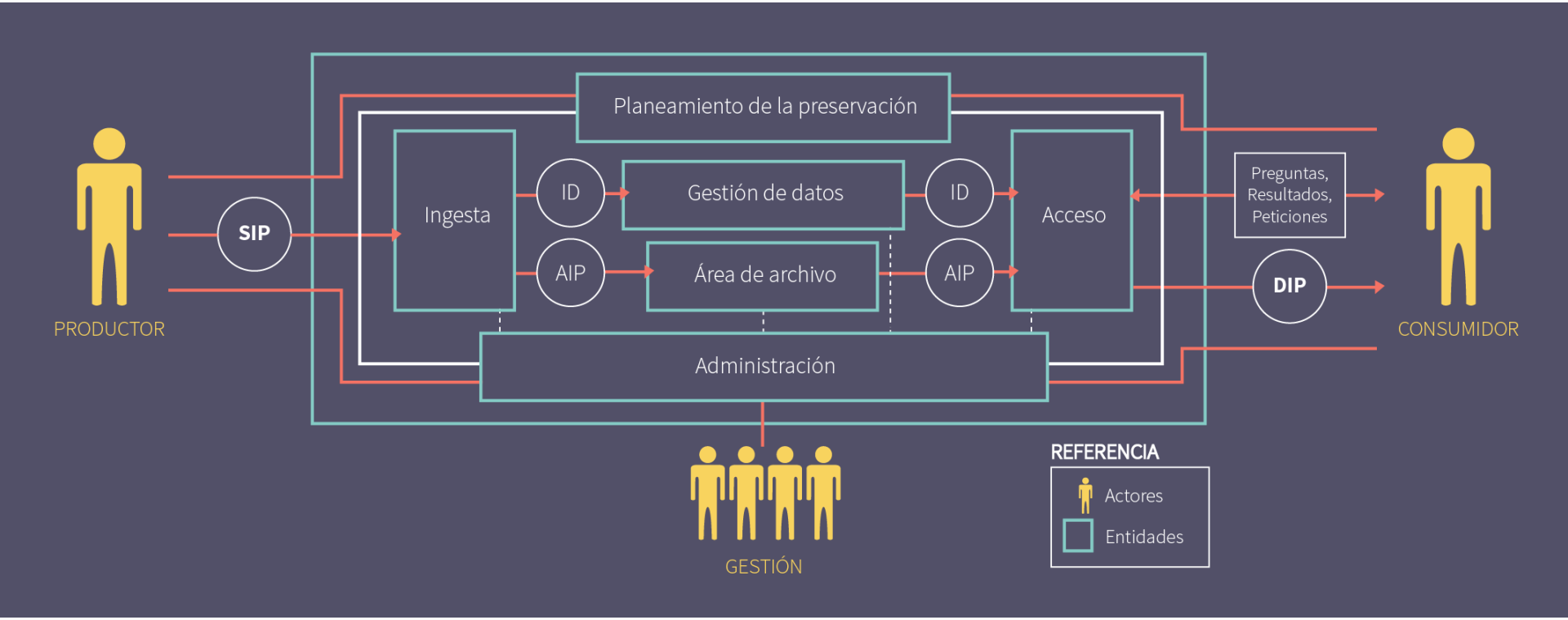
Índice

Objetivos de la tesis

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.

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

El modelo del repositorio



Preguntas y criterios de evaluación

- Saracevic-Covi: se plantean interrogantes: ¿por qué hay que evaluarlas? ¿con qué objeto? ¿qué se debe evaluar, a qué nivel y con qué criterios?
- La evaluación significa una valoración del desempeño o del funcionamiento de un sistema, o parte del mismo, en relación a cierto objetivo.
- El desempeño puede ser evaluado por ejemplo en relación a la Efectividad: ¿cuán bien desempeña un sistema (o cualquiera de sus partes) aquello para lo que fue designado?
- Fuhr: Su propuesta incluye evaluar todos los aspectos de las BD: colecciones, tecnología y usuarios y realizan una propuesta de criterios de evaluación en cada uno de los aspectos mencionados.

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)

Amenazas a los Objetos Digitales

1. Su propia naturaleza los hace efímeros.
2. Siempre están mediados por la tecnología. Obsolescencia.
3. Pérdidas por desastres.
4. Barreras de acceso en los archivos: claves, cifrado, formatos cerrados.
5. Desconocimiento de las responsabilidades por la curatela.
6. Falta de planificación, y de recursos para preservación.
7. Problemas legales: permisos para transformar las obras.
8. Descripción inadecuada: imposibilidad de recuperación.
9. Pérdida de información sobre el contexto, esto es sobre las tecnologías utilizadas para accederlos.

Objeto digital:

Acciones en su ciclo de vida para mantener el acceso

OD Y METADATOS DE PRESERVACIÓN

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

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

Deben guardarse las relaciones que vinculen al objeto con otros

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

Debe conocerse su **creador**

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

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

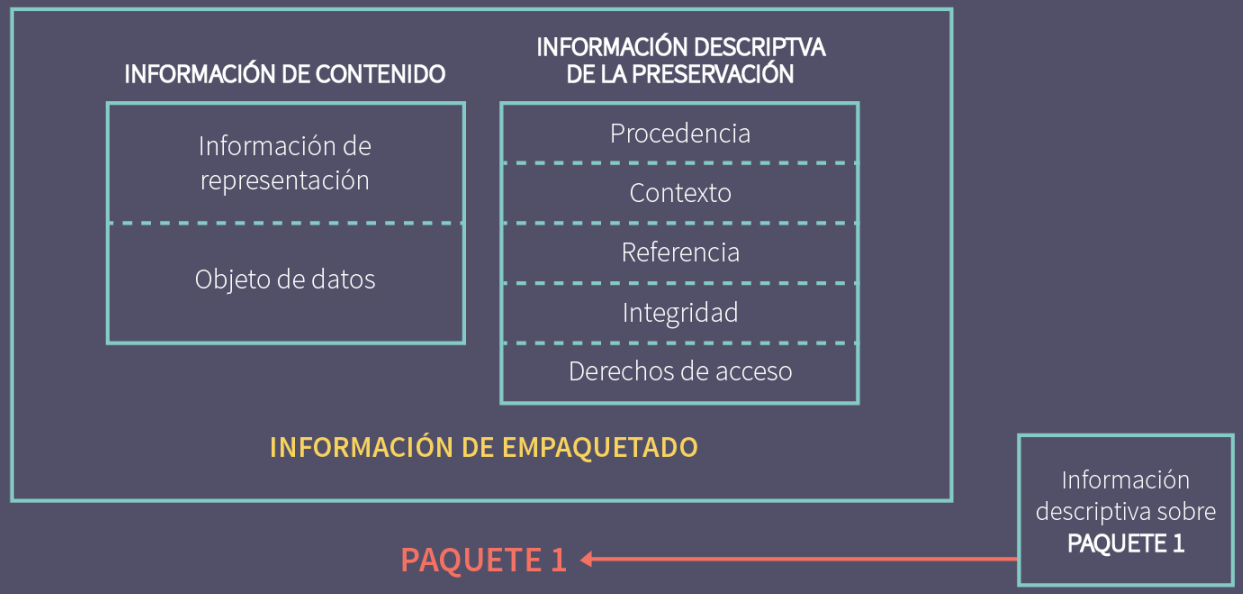
Las estrategias de **emulación** y **migración** requieren datos sobre los objetos originales y sus entornos

Autenticidad
Mediante la documentación de su procedencia

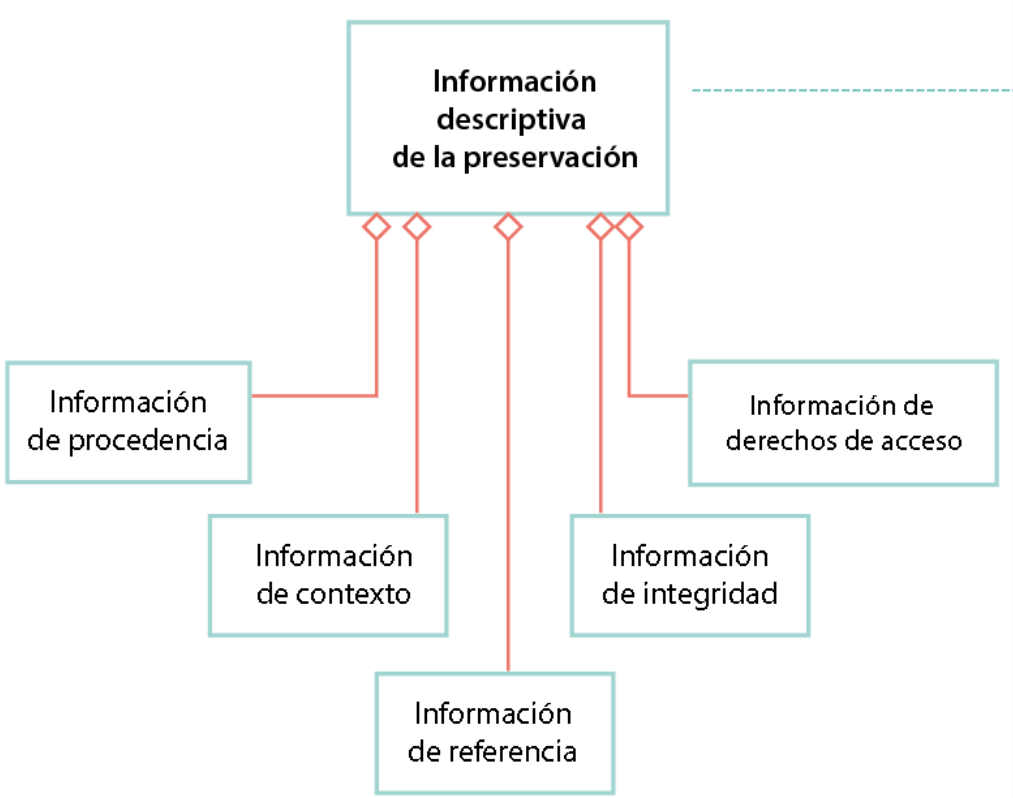
La propuesta metodológica

- Tomar como referencia central el Modelo OAIS de la ISO 14721.
- Utilizar la noción de paquete de información del Modelo OAIS y analizar la presencia de los distintos elementos del IP en los contenidos (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 y por lo tanto puede asegurar la preservación y el acceso a los contenidos en el tiempo.

El paquete de información en el OAIS



Acciones propuestas

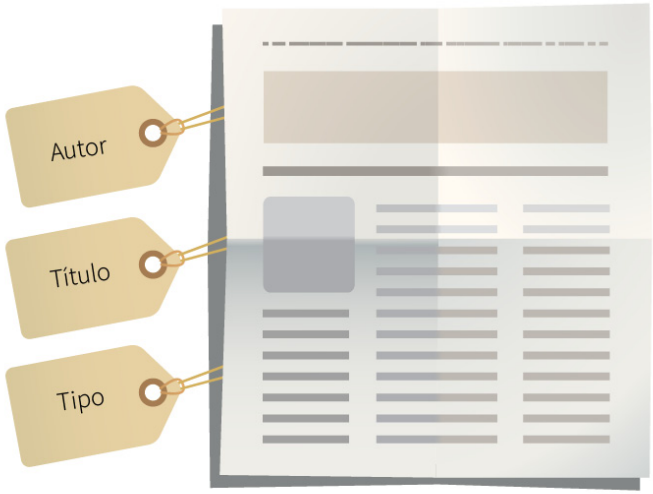


Las acciones propuestas están vinculadas justamente a los elementos constitutivos del paquete de información: la información de contenido (CDO), la información sobre la representación de ese contenido (RI), **la información descriptiva de preservación (PDI)** y la información descriptiva (DI).

¿Qué significan las acciones propuestas?

Revisar los objetos digitales y comprobar la existencia de muchos metadatos

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)

Acciones

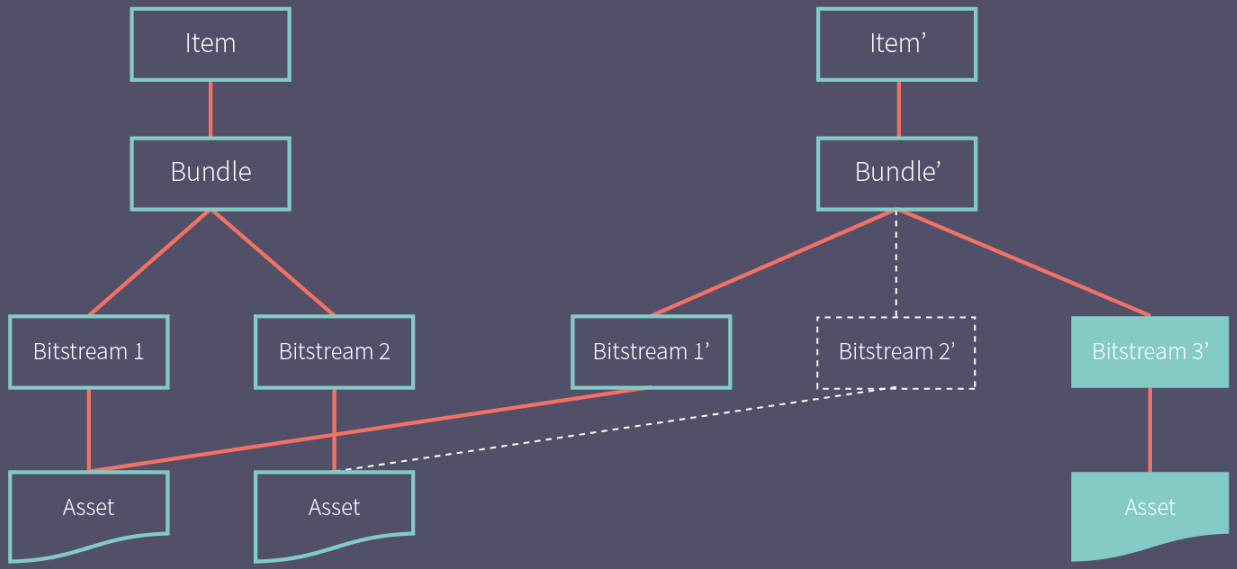
- Evaluar los objetos del repositorio como paquetes de información.
- Generar un reporte sobre el estado de los objetos del repositorio, considerándolos como paquetes de información.
- Informar si cuentan con todos los elementos que la norma define para el paquete de información.
    - Si los paquetes de información en el repositorio se adecúan a los criterios establecidos, los objetos digitales del repositorio y por tanto el repositorio mismo “pasan” la evaluación porque si el PI está bien formado el repositorio es funcional en sus procesos (ingesta, preservación y entrega).

Herramientas para llevar a cabo el experimento

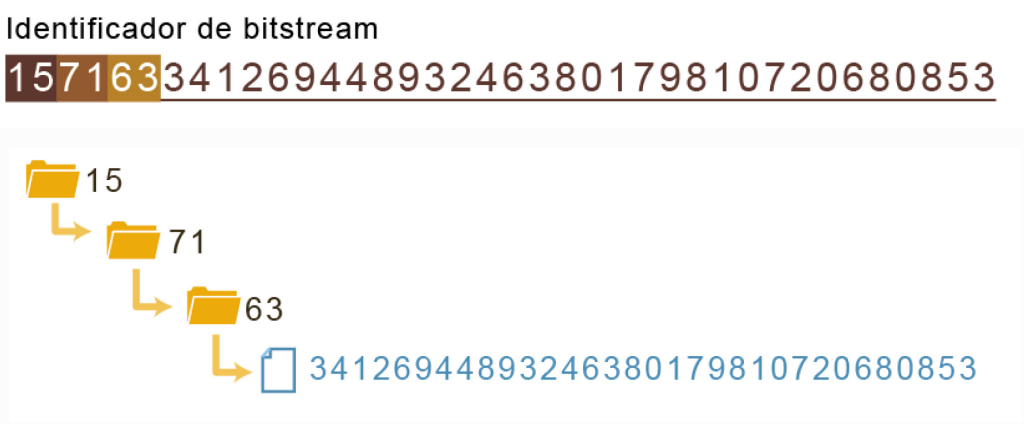
- **Perfilamiento automatizado de los objetos del repositorio:** esto involucra al objeto de contenido (CDO) con sus propiedades significativas y a la información de representación de ese objeto (RI). Realizar el perfil con DROID que contrasta con el registro PRONOM y brinda un reporte.
- **Revisión de los metadatos de preservación** que acompañan a los objetos digitales del repositorio y contraste con los metadatos de la PDI de OAIS a través de una herramienta de validación propia.
- **Revisión de la información descriptiva:** revisar los metadatos descriptivos y contrastarlos con las directrices DRIVER 2.0. y algunos metadatos extras.

Organización de los contenidos en DSPACE

Un item en DSPACE está constituído por metadatos, bundles y bitstreams

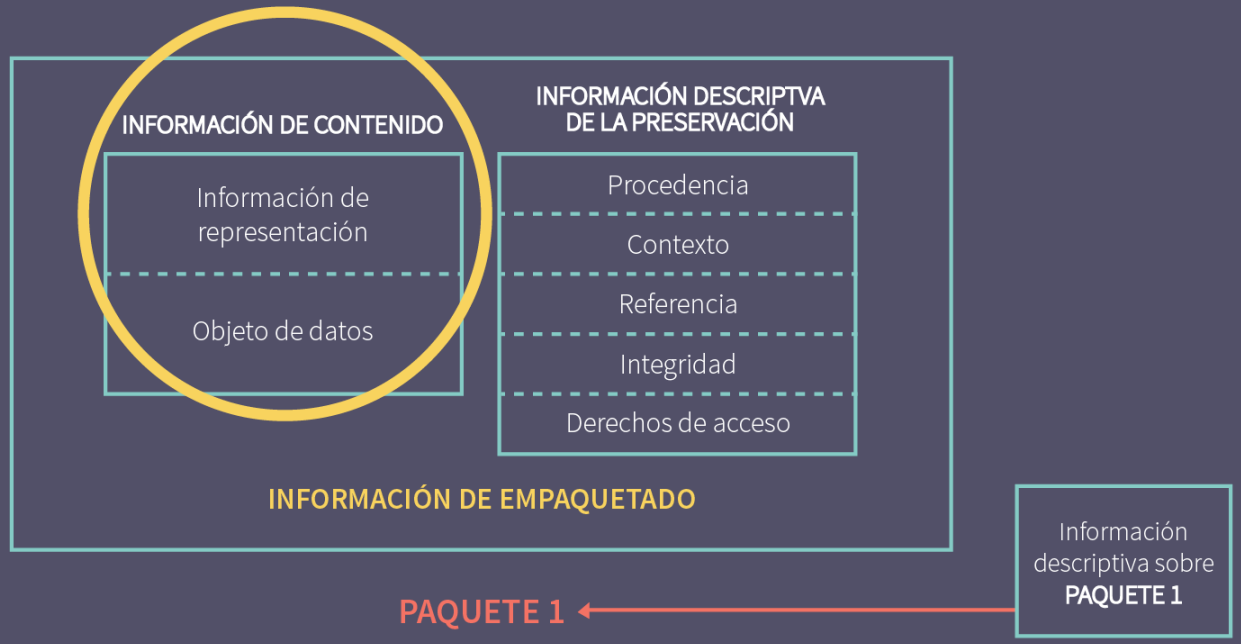


Identificador de bitstream y organización de carpetas y archivos del assetstore en DSpace



Mientras que el modelo de datos de Dspace (metadatos, workflows, estructura del repositorio, usuarios..) está soportado por la base de datos Oracle o Postgresql, los contenidos de los ítems se almacenan en el sistema de archivos denominado assetstore. Cada objeto digital se almacena con un identificador de 38 dígitos. La carpeta de almacenamiento se denomina *assetstore*.

El paquete de información en el OAIS



Primer Experimento

Información de representación:
Item: http://sedici.unlp.edu.ar/handle/10915/25088

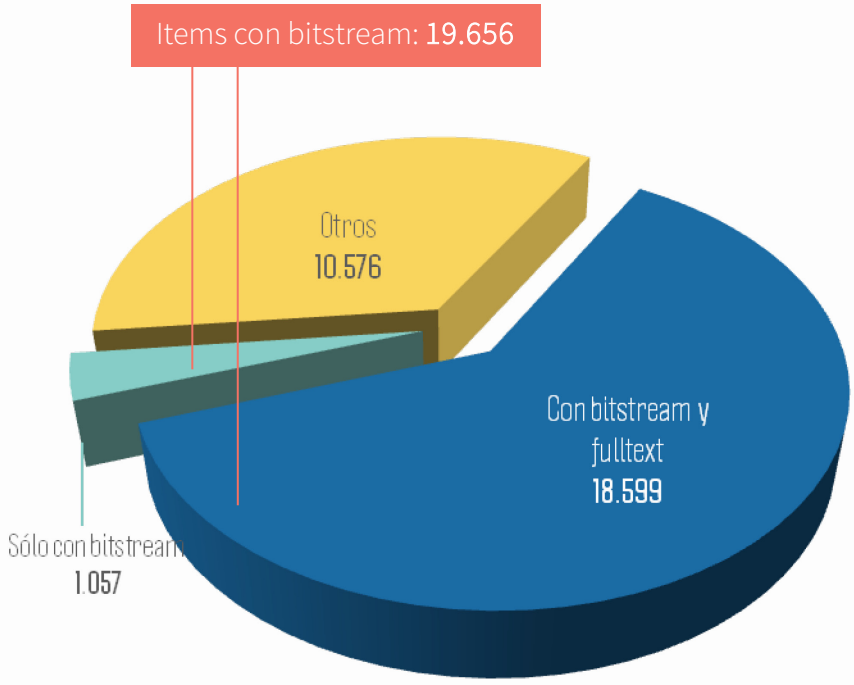
Se realizó un primer experimento con toda la estructura de carpetas del *assetstore*, alrededor de 90000 objetos, considerando todos los bitstreams de los distintos bundles de un item.

Falsas repeticiones detectadas debido al bundle TEXT en los casos de imágenes (extracción idéntica del media filter y mismo MD5.

Se decidió excluir el bundle TEXT para el próximo experimento y Corregir el problema de las falsas repeticiones (MD5 idénticos) haciendo el OCR de muchos archivos que eran PDF de sólo imágenes.

El caso de estudio SEDICI en números

Un año atrás



ITEMS EN TOTAL: 30.232

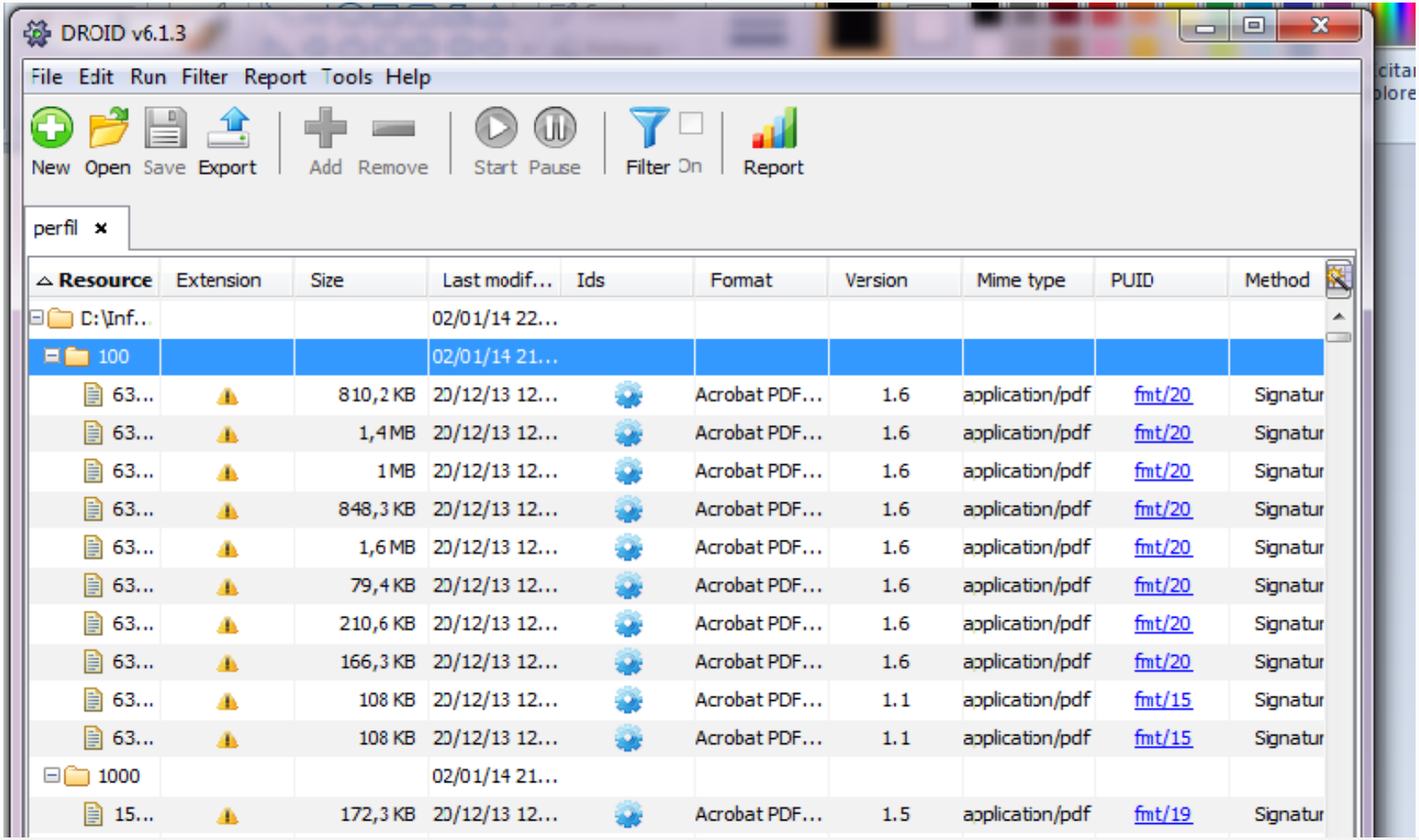
- ITEMS EN EL REPOSITORIO: **30.323**
- ITEMS CON BITSTREAM: **19.656**
- ITEMS CON BITSTREAM y FULLTEXT: **18.599**

Experimento final

- De la estructura del assetstore sólo la carpeta donde se encuentra el archivo.
- Sólo el bundle ORIGINAL de cada ítem.
- De los items sólo los que tienen al menos 1 bitstream en el bundle ORIGINAL
- De un assetstore completo de Diciembre de 2013 se analizaron casi 19000 archivos, se caracterizan los archivos por su formato PUID (Persistent Unique Identifier).
- Se analizaron los riesgos de esos formatos.
- Como tarea colateral se verificaron las repeticiones de checksum (MD5).
- Con los casos sospechosos se generan las tareas de análisis, revisión y corrección para que los administradores de SEDICI resuelvan el problema.

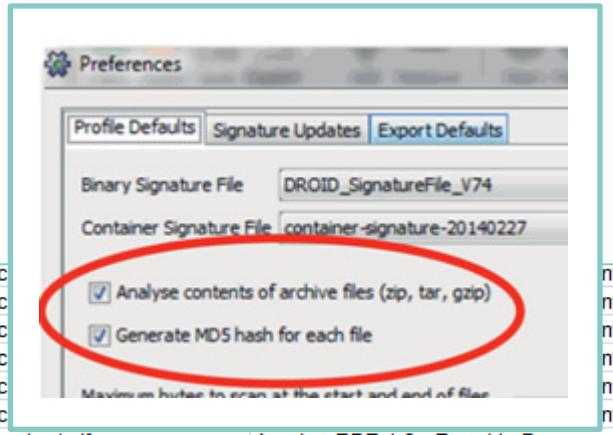
Perfil en DROID

A partir del perfil se revisó el registro PRONOM para ver los riesgos de los formatos



Resource	Extension	Size	Last modif...	Ids	Format	Version	Mime type	PUID	Method
C:\Inf...			02/01/14 22...						
100			02/01/14 21...						
63...		810,2 KB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		1,4 MB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		1 MB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		848,3 KB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		1,6 MB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		79,4 KB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		210,6 KB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		166,3 KB	20/12/13 12...		Acrobat PDF...	1.6	application/pdf	fmt/20	Signatur
63...		108 KB	20/12/13 12...		Acrobat PDF...	1.1	application/pdf	fmt/15	Signatur
63...		108 KB	20/12/13 12...		Acrobat PDF...	1.1	application/pdf	fmt/15	Signatur
1000			02/01/14 21...						
15...		172,3 KB	20/12/13 12...		Acrobat PDF...	1.5	application/pdf	fmt/19	Signatur

Perfil exportado con MD5
para detección de duplicados



1801	15757_	3328885	eaf803a90d5b190fd4b4935e6a852fc3	1	fmt/17	applic	nt Format	1.3
1802	6355_	175675	4564d6e8bc2f282435429c9560e841b1	1	fmt/17	applic	nt Format	1.3
1803	6354_	180887	cc0601b8518aa7d0addc22bab2925e1	1	fmt/17	applic	nt Format	1.3
1804	6356_	260190	6402659c609ce5bb4d22d585a596517	1	fmt/17	applic	nt Format	1.3
1805	6361_	105889	6b4f05561604f71946f3814ed2a313ee	1	fmt/17	applic	nt Format	1.3
1806	6359_	105889	6b4f05561604f71946f3814ed2a313ee	1	fmt/17	applic	nt Format	1.3
1807	6360_	105889	6b4f05561604f71946f3814ed2a313ee	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1808	15834_	5596903	0d02c816f3b71e08c270dc332fc580a9	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1809	6362_	196539	c45cc1c6dd997d059e318859ae77ee8	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1810	6363_	169421	6aa3fe9e89bcce079595fad164a99424	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1811	6364_	165625	c25df5f2549f1de3f8082152e5f2c8bc	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1812	6367_	120563	9c1db44a3a75b067398aacec160c247	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1813	6365_	491302	b8769fc53e969d10e37a3a40af15f672	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1814	6366_	1089885	ae988ddc448cfe7b526ca79a0cab8bd0	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1815	6372_	71355	283535028f1300c050d62da17d520372	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1816	15874_	503792	d2adf4cf28a2331e3fc8034a241251a0	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1817	15875_	877958	1d2c9ecf09361cee408def76b4cf2a4d	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1818	15877_	825731	c1c6d475af3dd9c6ae887741b44ba6cc	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1819	15880_	218787	a59e7f27ba35490150c4b530d25a509c	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1820	15888_	631271	c1c1e03084b4bdf06a91c19d56e68fcd	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1821	15892_	151251	c1788870df8b8e0defddea522a065357	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3
1822	15891_	246622	28a30b03071aed0478ba12e8a6e0a68	1	fmt/17	application/pdf	Acrobat PDF 1.3 - Portable Document Format	1.3

File count and sizes

Report field	Grouping fields
FILE_SIZE	

Filter fields:

Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

Profile	Count	Sum	Min	Max	Average
perfil	18522	26863674031	1120	701906423	1450365
Profile totals	**18522**	**26863674031**	**1120**	**701906423**	**1450365**

File sizes per extension

Report field	Grouping fields
FILE_SIZE	FILE_EXTENSION

Filter fields:

Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

Profile	Count	Sum	Min	Max	Average
perfil	18522	26863674031	1120	701906423	1450365
Profile totals	**18522**	**26863674031**	**1120**	**701906423**	**1450365**

Group totals

Count	Sum	Min	Max	Average
18522	26863674031	1120	701906423	1450365

File sizes per PUID

Report field	Grouping fields			
FILE_SIZE	PUID	FILE_FORMAT	FORMAT_VERSION	MIME_TYPE

Reporte en DROID exportado

Interpretación del reporte de DROID:

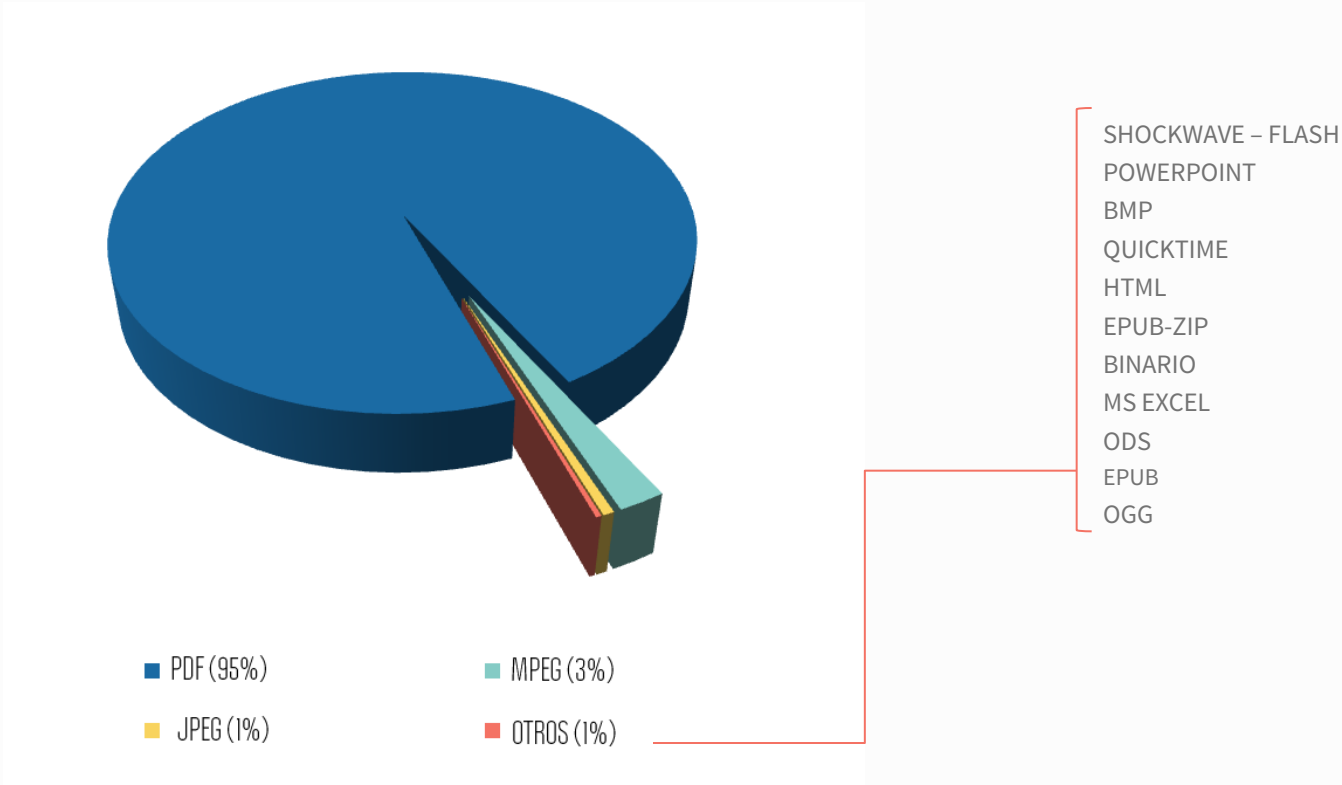
SÓLO ALGUNOS ARCHIVOS

El perfil elaborado por DROID queda plasmado en el reporte adjunto en la tesis denominado "Comprehensive Breakdown", una síntesis del mismo muestra:

- 47 archivos Macromedia Flash 5, formato que se corresponde con el PUID fmt/108 de PRONOM.
- 4 archivos Windows Bitmap 3.0, formato que se corresponde con el PUID fmt/116 de PRONOM.
- 5 archivos MS PPT 1997-2002, correspondiente al PUID fmt/126 de PRONOM.
- 583 archivos MPEG ½ Audio Layer 3, correspondiente al PUID fmt/134 de PRONOM.
- 17660 archivos PDF, de los cuales:
  - 1 archivo Acrobat PDF 1.0, formato que se corresponde con el PUID fmt/14 de PRONOM...hasta Acrobat PDF 1.7 que se corresponde con el PUID...
- 188 archivos JPEG, formato que se corresponde con el PUID …

Formatos-Mime Types

El gráfico representa de manera simplificada el perfil elaborado por DROID



Generación de tickets para mejora de calidad

Tickets para items sin bitstream ni LE

Generación de tickets para mejora de calidad

SeDiCI » SeDiCI-Docs

Búsqueda:

Vistazo | Actividad | Planificación | Peticiones | Nueva petición | Calendario | Noticias | Wiki | Foros

Tarea #2786

Modificar | Monitorizar | Copiar

Migración de item con PDFs de versiones 1.0 hasta 1.3 inclusive ---TMDG

« Anterior | 7/62 | Siguiente »

Añadido por Marisa Raquel De Giusti hace 6 meses. Actualizado hace 6 meses.

Estado:	Comentarios	**Fecha de inicio:**	21 Mayo 2014
Prioridad:	Normal	**Fecha fin:**	
Asignado a:	Carlos J. Nusch	**% Realizado:**	30%
Categoría:	archivos y formatos		
Versión prevista:	-		

Descripción

Citar

Dado que las versiones 1.0, 1.1, 1.2 y 1.3 de PDF no se pueden interpretados correctamente por versiones más nuevas.
Se plantea la conversión a PDF/A1-a a través de ABBYY e incluso con la verificación y validación de Acrobat.
Se sugiere la lectura del archivo: Conversión de archivos PDFs para realizar las acciones de manera correcta. El archivo fue compartido en DRIVE a todos los del SEDICI por Paula
Se adjunta el listado total de ítems que deben ser sometidos a este proceso.
Priorizando los que tienen fulltext del listado adjunto en EXCEL.

El handle es el numerito que está antes del _ y despues de la última/

Agrego al resto de la gente que va a hacer esta tarea

Archivos PDFs versiones viejas.xlsx (311,36 KB) Marisa Raquel De Giusti, 21 Mayo 2014 11:24 AM

Procedimiento de conversión de archivos a PDFA.pdf (1,112 MB) Analia Pinto, 23 Mayo 2014 09:19 AM

Análisis de los formatos existentes

- Portable Document Format (PDF) es el mayoritario de SEDICI.
- Se realizó un análisis de los subconjuntos estandarizados de PDF.
- Se realizaron pruebas con pdfaPilot para la conversión a PDF/A.
- Se perfiló también con otras herramientas abiertas como Jhove.

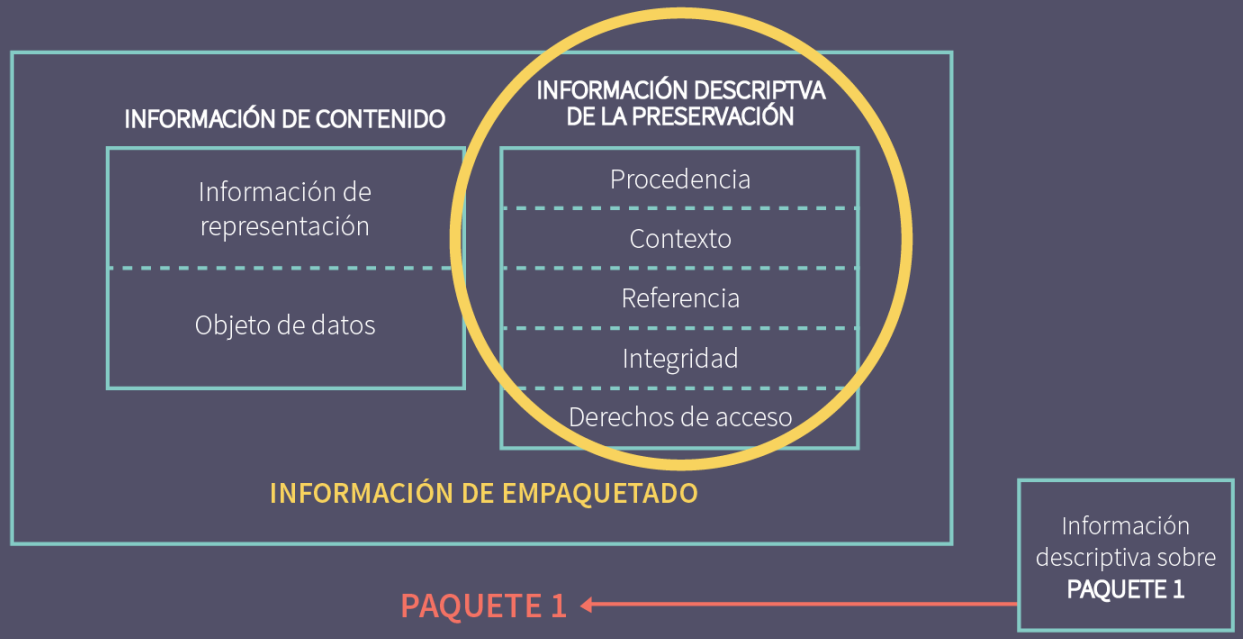
Recomendaciones

- Recomendación: migrar a PDF/A1-a en su defecto PDF/A1-b.
- En el caso de los libros, la opción de optimización es sumamente importante, ya que incide en numerosos aspectos (correcta recuperación del texto en búsquedas a texto completo, óptima visualización en web, por ejemplo), por lo que debe encontrarse la mejor manera de cumplimentar ambos objetivos: que el archivo cumpla con el estándar y que quede correctamente optimizado.
- Recomendaciones sobre qué hacer en el caso de tener que generar PDF a partir de documentos de texto: DOC,DOCX,ODT,RTF.

Recomendaciones

- Tras la digitalización y generación de documentos PDF/A con OCR evitar la edición mediante Acrobat Writer .
- Utilizar estándares abiertos siempre que sea posible.
- Aplicar transformaciones en archivos de imagen,texto, audio y video.
- 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 una posible migración a nuevas versiones, o a formatos abiertos.
- Generar un bundle de preservación.
- Extraer automáticamente los metadatos técnicos (FITS) y guardarlos en la base de datos acompañando al bitstream correspondiente.

El paquete de información en el OAIS



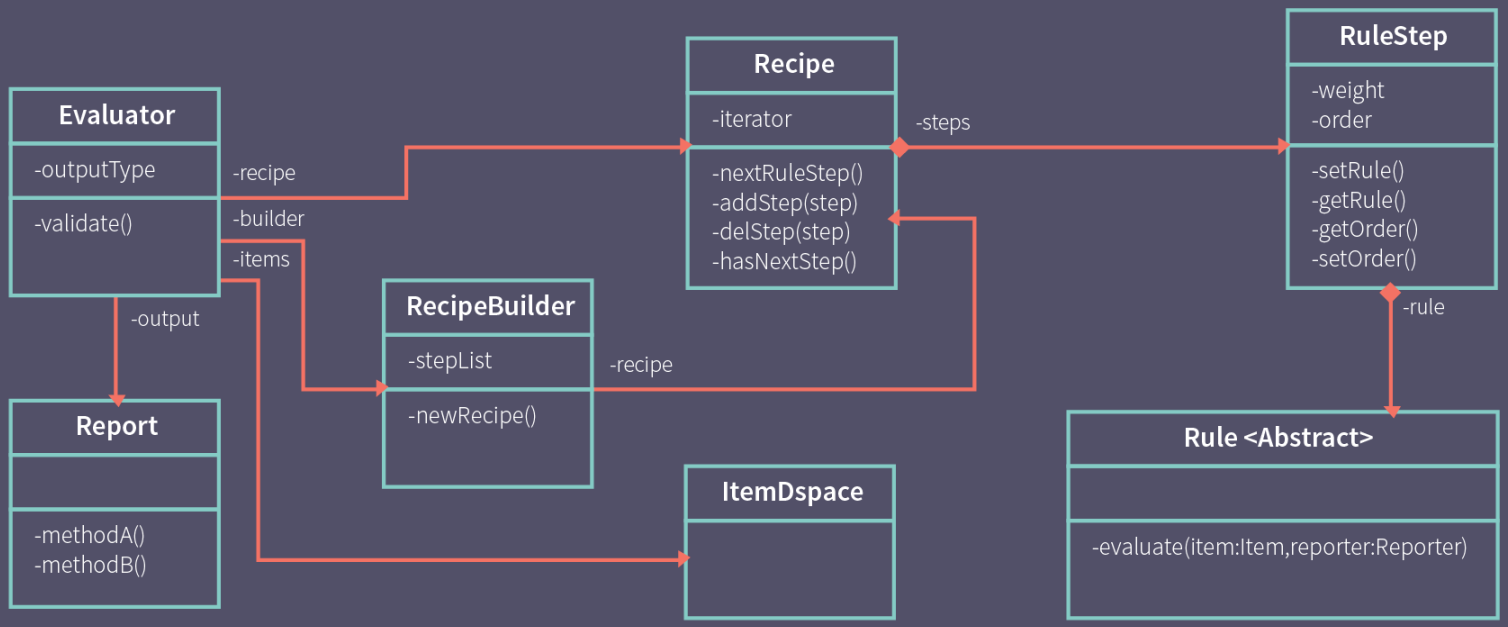
La Información descriptiva de la preservación

- La PDI de la norma OAIS está compuesta por información de procedencia, de contexto, de referencia, de integridad y de derechos.
- Esta información, de estar presente:
  - es generada por el software del repositorio (ej. procedencia)
  - se incorpora a través de tareas del flujo de trabajo de la administración.
- Para verificar la corrección de la PDI:
  - observar si estos metadatos están presentes
  - verificar que los valores son adecuados.
- Evidentemente es necesario automatizar esta tarea.

Metodología propuesta

- Desarrollo de un validador que tome la forma de una tarea de curación.
- Las tareas de curación añaden funcionalidad a DSpace.
- Se relacionan con la gestión de los objetos del repositorio, de ahí el término “curación” utilizado, homologable a “preservación”.
- Una tarea de curación en DSpace puede aplicarse a nivel ítem, colecciones o comunidades e incluso al repositorio completo.
- El propósito de las tareas de curación planteadas es efectuar el mantenimiento de ítems en el tiempo y a lo largo de todo su ciclo de vida.

Validador

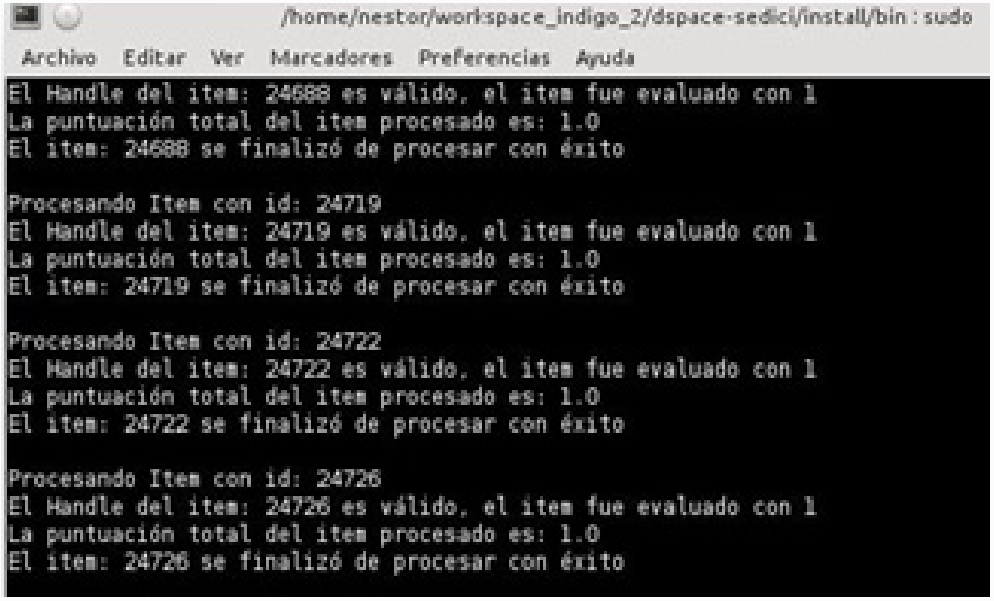

Evaluator
-outputType
-validate()
-recipe
-builder
-items
-output
Recipe
-iterator
-nextRuleStep()
-addStep(step)
-delStep(step)
-hasNextStep()
-steps
RuleStep
-weight
-order
-setRule()
-getRule()
-getOrder()
-setOrder()
-rule
RecipeBuilder
-stepList
-newRecipe()
-recipe
Report
-methodA()
-methodB()
ItemDspace
Rule <Abstract>
-evaluate(item:Item,reporter:Reporter)

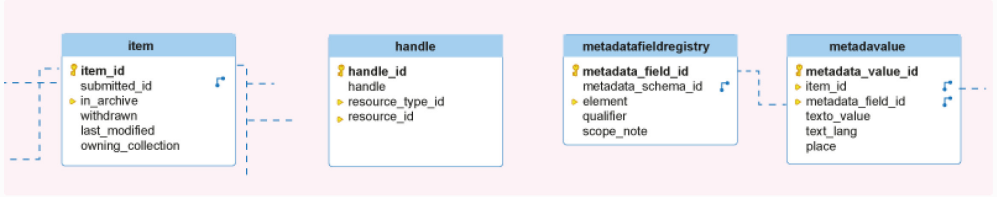
Validaciones de la PDI

- Referencia: se evalúa validando los identificadores persistentes; para el caso de DSpace se evalúa el handle.
- Integridad: se evalúa utilizando el checksum; para el caso de DSpace debe validarse que el algoritmo MD5 sea correcto y que no tenga el valor que está por defecto.
- Procedencia: se evalúa el metadato provenance, comenzando por ejecutar una consulta que muestre el contenido de ese metadato.
- Contexto: se evalúa el contexto según OAIS. La información de contexto tiene como objetivo principal documentar las relaciones de la información de contenido con su medioambiente (por qué fue creada esa información de contenido y su relación con otra información de contenido).
- Acceso: se evalúa a partir de las licencias.

Ejemplo de regla

Un año atrás





Enfoque: DSpace

Nombre: "Verificar validez de handle"

Restricciones: el handle siempre existe en DSpace, puede ser el handle predefinido

Regla: El ítem contiene un handle válido y no se trata del handle por defecto de (123456789)

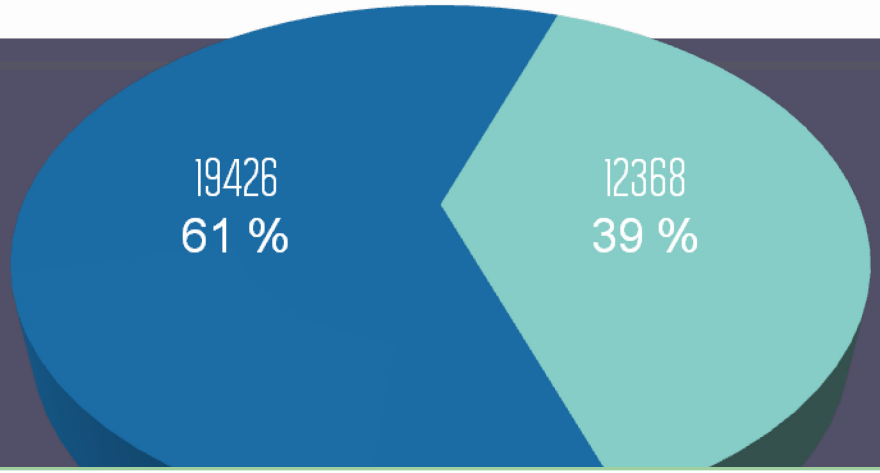
Metadato(s) asociado(s): dc.identifier.uri (sedici.identifier.handle)

Respuesta esperada:

✓ True → Válido

× False → Inválido

Ejemplo de resultados y acciones



Error #2767

Actualizar | Monitorizar | Copiar

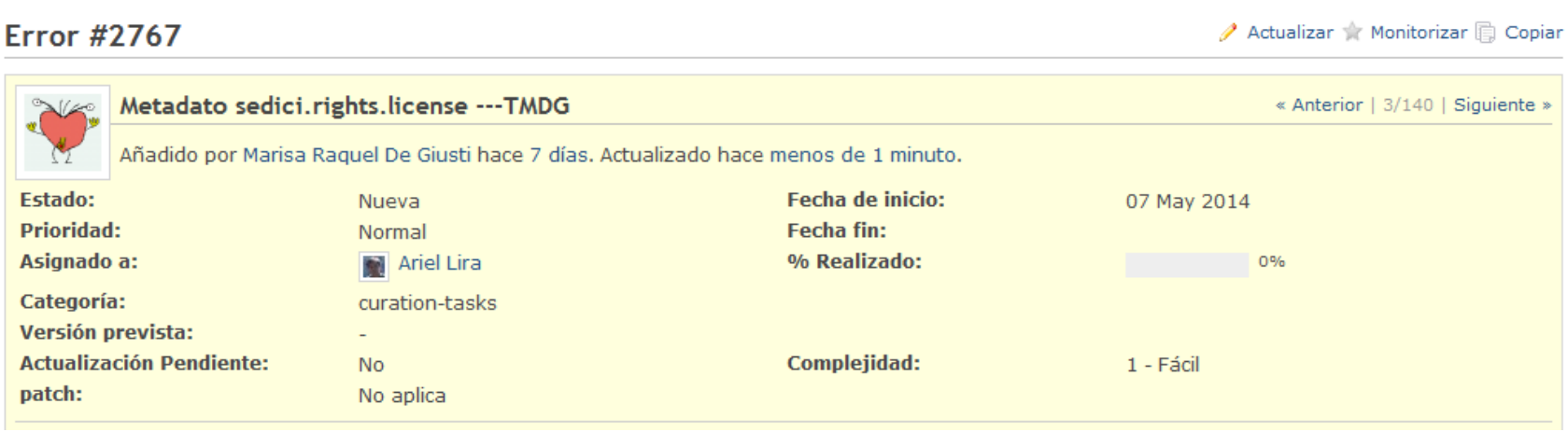
Metadato sedici.rights.license ---TMDG

« Anterior | 3/140 | Siguiente »

Añadido por Marisa Raquel De Giusti hace 7 días. Actualizado hace menos de 1 minuto.

Estado:	Nueva	**Fecha de inicio:**	07 May 2014
Prioridad:	Normal	**Fecha fin:**	
Asignado a:	Ariel Lira	**% Realizado:**	0%
Categoría:	curation-tasks		
Versión prevista:	-		
Actualización Pendiente:	No	**Complejidad:**	1 - Fácil
patch:	No aplica		

Ejemplo de resultados y acciones



Error #2767

Actualizar | Monitorizar | Copiar

Metadato sedici.rights.license ---TMDG

« Anterior | 3/140 | Siguiente »

Añadido por Marisa Raquel De Giusti hace 7 días. Actualizado hace menos de 1 minuto.

Estado:	Nueva	**Fecha de inicio:**	07 May 2014
Prioridad:	Normal	**Fecha fin:**	
Asignado a:	Ariel Lira	**% Realizado:**	0%
Categoría:	curation-tasks		
Versión prevista:	-		
Actualización Pendiente:	No	**Complejidad:**	1 - Fácil
patch:	No aplica		

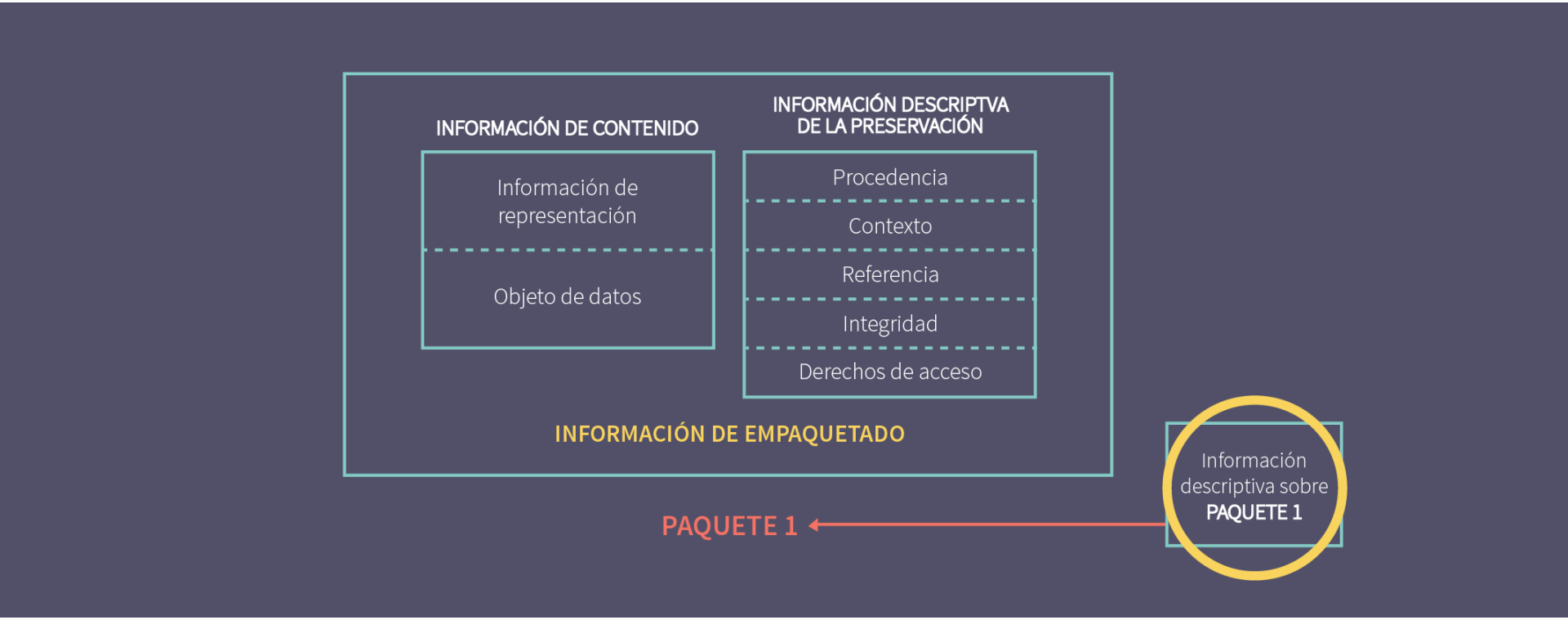
De los 19.946 ítems, no todos tienen el nombre correcto de la licencia, es decir, el metadato sedici.rights.license. Se encontraron 788 ítems con errores en el nombre de la licencia, problema que no se resuelve sólo con corregir esos ítems.

Items sin licencia

En relación a los 12.368 ítems sin licencia:

- Existen ítems que tienen asociada una licencia en papel (el archivo digital de éstas aún no fue realizado y/o no se ha incorporado al flujo de trabajo).
- Las licencias de distribución de SEDICI, previas a la versión 1.4 (de mayo de 2012), no obligaban a los autores a elegir licencias de uso Creative Commons, sino que solamente incluían la autorización del autor para la transformación de su obra con fines de preservación y el permiso de difusión.
- Como hoy en día se considera que todos los ítems deben tener una licencia de uso, se ha generado un ticket para corregir el problema.

Análisis de la información descriptiva



Directrices Driver

Basic element	Status	Encoding schemes
Title	M	None, free text
Creator	M	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name)
Subject	MA	Choice of keywords and classifications can be free text (preferably in English) and defined by an URI scheme (preferably info:eu-repo/classification)
Description	MA	None, free text. Recommended practice is to include an abstract in English. "Abstract" is the default interpretation to the value for dc:description
Publisher	R	None
Contributor	O	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name)
Date	M	Date | ISO 8601 W3C-DTF - "Published" is the default interpretation to the value for dc:date
Type	M	Publication type and Version type can be free text (preferably in English) and defined by an URI scheme (preferably info:eu-repo/semantics).
Format	R	IANA registered list of Internet Media Types (MIME types)
Identifier	M	URI scheme, linking to persistent identifier (URN, handle, DOI), full text document or human start page.
Source	O	Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata as in dcterms:bibliographicCitation
Language	R	ISO 639-3
Relation	O	None
Coverage	O	"Period" is the default interpretation to the value for dc:coverage Encoding: DCMI Period [http://dublincore.org/documents/2000/07/28/dcmi-period/] For more ncoding schemas see Chapter 5 Use of vocabularies and semantics.
Rights	R	None
Audience	O	None. "Eduction level" is the default value for dc:audience.

Verificaciones

item_export_2014_mar_19_1_2980.zip\1 - archivo ZIP, tamaño descomprimido 93.137 bytes

Nombre	Tamaño	Comprimido	Tipo	Modificado	CRC32
..			Carpeta de archivos		
metadata_sedici2003.xml	4.416	1.345	Documento XML	19/03/2014 02:...	8C53EBAC
metadata_sedici.xml	1.606	616	Documento XML	19/03/2014 02:...	F29F7C61
metadata_mods.xml	217	167	Documento XML	19/03/2014 02:...	0FC129B3
handle	12	14	Archivo	19/03/2014 02:...	62611031
dublin_core.xml	2.044	944	Documento XML	19/03/2014 02:...	23FB374D
contents	116	92	Archivo	19/03/2014 02:...	F1C588AA
all-0001.pdf.txt	21.653	7.132	Documento de texto	19/03/2014 02:...	FD6A39BC
all-0001.pdf	63.073	47.921	Documento Adob...	19/03/2014 02:...	749CAACD

De los metadatos descriptivos, según la sugerencia de DRIVER, son obligatorios: *Título, Creador, Fecha, Tipo e Identificador*. Los elementos *Description* que aparecen en el archivo dublin_core.xml y *Subject* en el archivo metadata_sedici.xml también serán verificados para conocer qué porcentaje de los objetos del repositorio cuentan con esta información.

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.

Ticket en relación al metadato resumen

Tarea #2758

Actualizar | Monitorizar | Copiar

Items sin resumen

« Anterior | 1/52 | Siguiente »

Añadido por Marisa Raquel De Giusti hace 31 minutos. Actualizado hace 28 minutos.

Estado:	En progreso	**Fecha de inicio:**	30 April 2014
Prioridad:	Normal	**Fecha fin:**	
Asignado a:	Analia Pinto	**% Realizado:**	30%
Categoría:	Mejora de calidad		
Versión prevista:	-		

Descripción Citar

En muchos archivos falta el metadato dc.description.abstract. Hay casi 4500 ítems que no lo tienen. Va un archivo adjunto.

Sin metadato resumen.csv (160 KB) Marisa Raquel De Giusti, 30 April 2014 08:34 AM

Evaluación realizada

- Centrada en el contenido, la representación del contenido (formato), la PDI y la información descriptiva.
- Verificación de muchos metadatos.
- Bajo el paquete de información en sus tres presentaciones (SIP, AIP y DIP) se han recorrido las tres dimensiones subyacentes en los modelos de evaluación de RIs: sistema, datos y usuarios (comunidad designada).

Experimento con SEDICI

- Se ha evaluado la totalidad de los contenidos de SEDICI afectando 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 con formatos antiguos, ítems sin licencia, ítems con licencias erróneas, etcétera.
- Se han generado 104 tickets en el gestor de incidencias de SEDICI (algunos involucran cientos y hasta miles de ítems).
- En SEDICI-Docs y en SEDICI-Dspace (p.e. migraciones).

Recomendaciones

Vinculadas a:

- Materiales nacidos digitales.
- Materiales digitalizados.
- Formatos de visualización y formatos de preservación.
- Conveniencia de formatos abiertos.
- Formas de trabajo que aseguran los formatos de preservación, por ejemplo PDF/A.
- Verificación de metadatos de la PDI y de la DI.
- Necesidad de trabajar con herramientas adicionales a DSPACE.
- Necesidad de chequeo y validaciones periódicas de los metadatos.

Trabajos futuros

- Análisis de las migraciones masivas y sus problemas para la administración.
- Selección de herramientas de migración.
- Reporte de los eventos de migración y sus responsables de modo de asegurar la trazabilidad en el ciclo de vida (evento y agente en PREMIS).

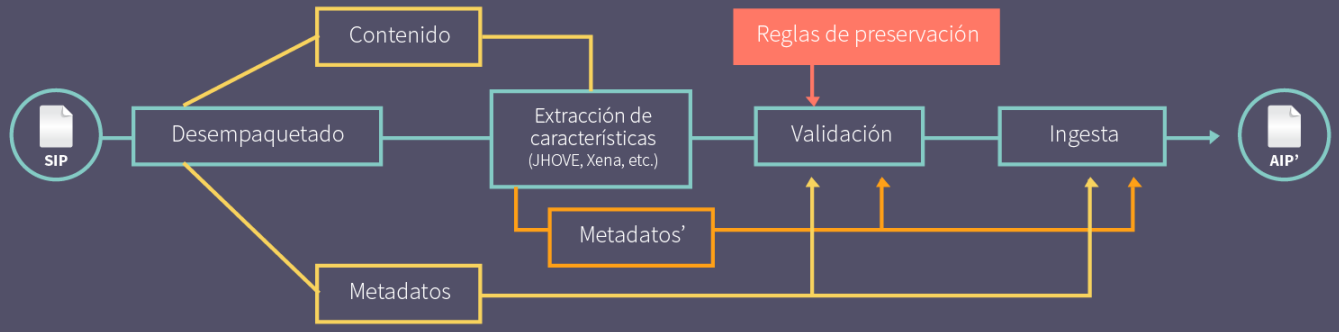
Trabajos futuros

- En relación a las posibles “capas” de las reglas, se plantean como trabajos futuros, a llevar adelante por otros integrantes de SEDICI, reglas más abstractas, pensadas más allá de que el repositorio se encuentre implementado en DSpace o en cualquier otro software para repositorios. Es decir, que se estarían considerando reglas a nivel repositorio.
- Extensión de la herramienta de validación: Implementación de un lenguaje que permita a la administración del repositorio, a través de una sintaxis sencilla, realizar controles sobre los items de manera individual, por colección, etcétera y se obtenga un reporte del estado de los items consultados.

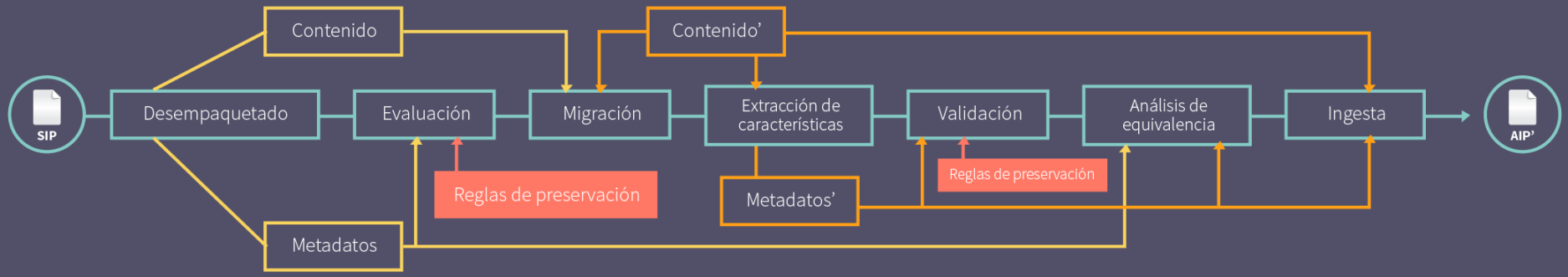
Trabajos futuros

Introducción en el flujo de trabajo de herramientas que permitan la verificación y validación de formatos:

En la ingesta: al ingresar el SIP realizar el desempaquetado y que la herramienta de extracción de características realiza la caracterización del objeto digital.

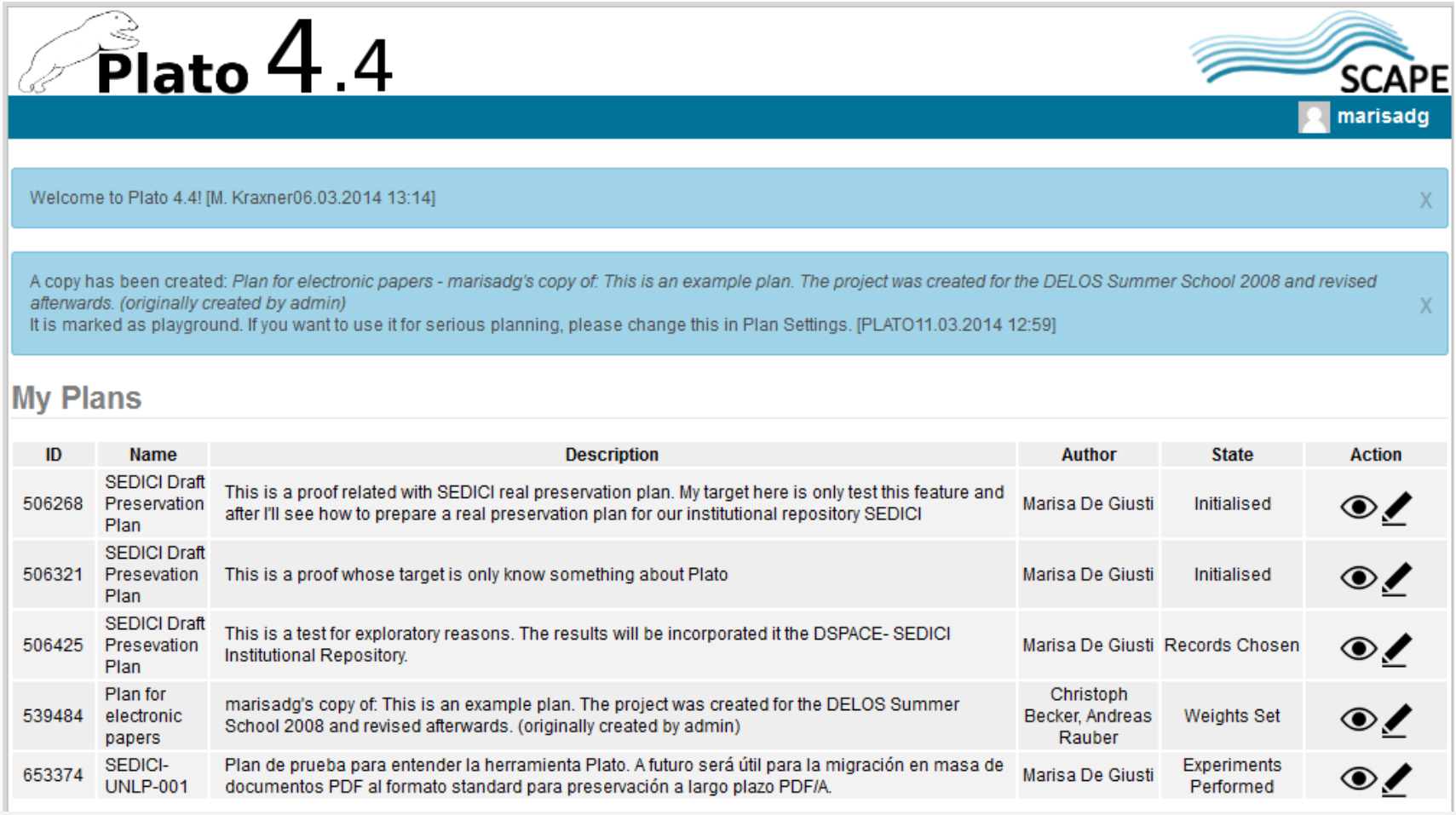


En el AIP, se evalúa si es necesaria la migración, en caso de serlo se reingresa un nuevo AIP y metadatos.



Trabajos futuros: plan de preservación

Planes de preservación en etapa de inicio del repositorio SEDICI



ID	Name	Description	Author	State	Action
506268	SEDICI Draft Preservation Plan	This is a proof related with SEDICI real preservation plan. My target here is only test this feature and after I'll see how to prepare a real preservation plan for our institutional repository SEDICI	Marisa De Giusti	Initialised	
506321	SEDICI Draft Presevation Plan	This is a proof whose target is only know something about Plato	Marisa De Giusti	Initialised	
506425	SEDICI Draft Presevation Plan	This is a test for exploratory reasons. The results will be incorporated it the DSPACE- SEDICI Institutional Repository.	Marisa De Giusti	Records Chosen	
539484	Plan for electronic papers	marisadg's copy of: This is an example plan. The project was created for the DELOS Summer School 2008 and revised afterwards. (originally created by admin)	Christoph Becker, Andreas Rauber	Weights Set	
653374	SEDICI-UNLP-001	Plan de prueba para entender la herramienta Plato. A futuro será útil para la migración en masa de documentos PDF al formato standard para preservación a largo plazo PDF/A.	Marisa De Giusti	Experiments Performed	

El futuro es  y grande!

GRACIAS

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