Conferencia internacional
BIREDIAL-ISTEC
17-18-19 OCTUBRE 2016

Esta obra está bajo una Licencia Creative Commons Atribución-CompartirIgual 4.0 Internacional.

Las dificultades de la preservación digital: Problemas, desafíos y propuestas para los repositorios

Dra. Marisa R. De Giusti
Directora PREBI-SEDICI
Universidad Nacional de La Plata
Directora CESGI
Comisión de Investigaciones Científicas de la Provincia de Buenos Aires
Argentina

Objetivos de esta presentación

Esta presentación tiene la pretensión de discutir un conjunto de acciones concretas necesarias al objetivo de la preservación digital, dejando de lado los aspectos institucionales políticos, estratégicos e incluso económicos que pudieran regir esta actividad. El alcance y ámbito de esta propuesta, es el propio repositorio institucional con sus tecnologías y contenidos en diversos formatos.

- Motiva esta presentación el reconocimiento de que hay mucho por hacer en preservación digital y que el conocimiento de herramientas, normas, modelos, recomendaciones, acciones e incluso desarrollos para la implementación de repositorios confiables, si bien puede llevarnos a grandes confusiones, abre un camino de acciones a pensar en conjunto en América Latina y en el contexto particular de este congreso.
- Esta presentación no es un ejercicio de saberes sino un llamado a acordar tareas en conjunto en AL para la preservación de los objetos digitales y el logro de repositorios confiables.
- Toca pensar en términos de repositorios confiables: ISO 16363.

Preservación digital en el enfoque de esta presentación, no es digitalización de documentos, ni es guardado de copias múltiples, todos los cuales pueden ser parte del proceso de preservación pero no lo único, es decir el proceso de preservación no queda restringido a esos aspectos, digitalizar puede ser un proceso para incluir contenidos en un repositorio institucional y se sumarán a los archivos digitales a preservar que son el verdadero problema. Guardar múltiples copias de un documento, no va a asegurar el acceso y la legibilidad de los mismos, es una política de backups y sobre las tantas copias que haya, se deberán llevar adelante los procesos de preservación.

¿Qué hacer?

Estándares, normas, recomendaciones, metadatos, formatos herramientas, software, procesos automáticos, procesos manuales… muchas cosas,

Cada vez más tipologías documentales en el repositorio,

Todo el ciclo de vida

Figura: Pérez-Hernández, Dayni, & Delgado-Dapena, Martha Dunia. (2013). Modelo de gestión de objetos digitales para la gestión de soluciones tecnológicas. *Ingeniería Industrial*, *34*(1), 40-49. Recuperado en 04 de octubre de 2016, de: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1815-59362013000100005&lng=es&tlng=es

El desafío de la preservación digital

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. Es necesario asegurar que estén disponibles y sean accesibles a largo plazo.

El desafío de la preservación digital

El problema actual de la preservación digital abarca también a los documentos en papel que han sido sometidos a un proceso de digitalización, debe estar claro que **tras realizar la digitalización, comienza el problema de la preservación** y el acceso del mismo modo, de manera idéntica que para cualquier documento nacido digital.

Preservación digital: Si o No

- Criterios tradicionales para documentos en papel: en los documentos tradicionales en papel se habla de “negligencia benigna”: el olvido de un manuscrito en un arcón...
- En los documentos digitales:
  - 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.
  - No necesariamente se conserva la integridad externa del documento sino las propiedades significativas.
  - Se debe asegurar la autenticidad del recurso.

Problemas en la preservación digital

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.

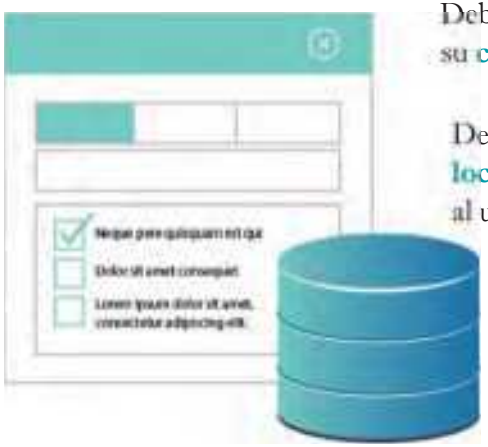
OD y metadatos de preservación

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

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

Deben guardarse las relaciones que lo vinculen con otros objetos

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



Debe conocerse su **creador**

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

Su soporte debe 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

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.

"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)

Problemas en la 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?,
- organizativo ¿de quién es la responsabilidad de cada acción? ¿cómo se asegura la continuidad de las decisiones?

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:

1. Los datos se mantendrán en el repositorio sin sufrir daños, sin perderse o sin ser alterados de forma malintencionada/o no.

2. Los datos podrán ser localizados y entregados al usuario.

3. Los datos podrán ser interpretados y comprendidos por el usuario.

4. Las metas 1, 2 y 3 serán realizables a largo plazo.

Niveles en la preservación digital

Nivel físico: (*Bitstream preservation*) incluye básicamente al objeto digital y su entorno: discos, dispositivos de lectura, insertos a su vez en otros tantos dispositivos, no se trata simplemente de copiar de un espacio físico a otro, en realidad eso involucra cuestiones lógicas como sistemas operativos, convenciones de nombres y hasta referencias para la localización de los archivos.

Andreas Rauber. “Digital Preservation Introduction” Department of Software Technology and Interactive Systems Vienna University of Technology http://www.ifs.tuwien.ac.at/~andi

Niveles en la preservación digital

Nivel lógico: los archivos digitales precisan abrirse con determinados programas, los cuales a su vez precisan determinados sistemas operativos que funcionan a su vez bajo un conjunto determinado de componentes de hardware (y no otros). Cualquier cambio en lo precedente puede conducir a problemas tales como la imposibilidad de abrir un archivo, de no poder acceder a algún objeto embebido en el archivo y otras muchas posibilidades que podrían resultar en pérdida de información digital.

Andreas Rauber. “Digital Preservation Introduction” Department of Software Technology and Interactive Systems Vienna University of Technology http://www.ifs.tuwien.ac.at/~andi

Niveles en la preservación digital

Nivel semántico: este nivel tiene que ver específicamente con la información que permite interpretar los datos, un caso sencillo para comprender esto puede ser la visualización de un mapa, si los límites entre países han cambiado respecto de la información del archivo, la comprensión del mismo va a ser errónea, este punto es bastante complejo y ayuda a poner en claro la necesidad de la actualización de la información de contexto o de la necesidad de post procesamiento o de transformaciones de algún tipo, empleando aquí el término transformación en un sentido muy general.

Andreas Rauber. “Digital Preservation Introduction” Department of Software Technology and Interactive Systems Vienna University of Technology http://www.ifs.tuwien.ac.at/~andi

Preservación a nivel físico

- Múltiples copias de cada objeto digital.
- Distintos tipos de medios de almacenamiento (al menos dos tipos)
- Diferentes copias almacenadas en lugares geográficamente diferentes.
- Mantenimiento de hardware y software.

Es crucial para implementar la conservación el control de calidad y seguimiento regular con el fin de detectar los errores que suceden, ya sea debido al hardware en sí o al mover o cambiar archivos.

Preservación a nivel físico

Almacenamiento de recursos de **SEDICI** (a modo de ejemplo)

- Almacenamiento local en múltiples discos espejados (RAID)
- Sincronización incremental diaria automática con servidor de backups local
- Sincronización incremental diaria automática con servidor de backups remoto
- Sincronización completa semanal automática con servidor de backups remoto

Para documentos de trabajo (archivos digitalizados, copias locales de documentos)

- Almacenamiento local (workstations) y en servidor de archivos centralizado
- Sincronización incremental diaria automática con servidor de archivos remoto

Preservación a nivel lógico

Pensar en los formatos de los objetos digitales: cómo mantenerlos accesibles y comprensibles, qué programa correr para leer un dado archivo, dónde, con qué dispositivos.

El proceso de conservación de la accesibilidad de archivos y el asegurar que siguen siendo comprensibles y legibles, independientemente de las tecnologías en evolución, es justo lo que se denomina como conservación a nivel lógico.

Preservación a nivel lógico: necesidades

- Una variedad de formatos y codecs elegidos por el repositorio, y una visión general del software compatible con ellos;

- La producción de metadatos para cada archivo: dependiendo del tamaño del repositorio y el tipo de datos, los metadatos pueden ser mantenidos en diferentes maneras. Con una gran colección de archivos, se recomienda que se almacenen metadatos en la situación ideal, es decir, la información básica incrustado en el archivo de metadatos y más complejo en una base de datos específica vinculada a los archivos. Con colecciones más pequeñas a veces es mejor mantener los metadatos de cada archivo incrustado dentro del propio archivo.

Preservación a nivel lógico

- Evaluación periódica de los formatos y codecs de software y archivos para evitar la obsolescencia. Esto puede implicar una amplia gama de enfoques. Cada enfoque varía dependiendo del contenido de los repositorios.

Preservación a nivel lógico: formatos

En relación a los formatos de archivo lo ideal sería definirlos cuando se establece un repositorio dentro de las políticas de datos (por supuesto esta tarea es iterativa porque la tecnología cambia) sin embargo en reportes, guías, recomendaciones y muy especialmente en la práctica de archivos y proyectos exitosos y de grandes instituciones lo que más existe es una gran diversidad de formatos aceptados o preferidos.

Diccionario de Datos PREMIS de Metadatos de Preservación

"El concepto de formato parece casi intuitivo, pero debido a la importancia que tiene la información del formato para la preservación digital, el grupo decidió ser muy concreto respecto a su significado. Debatiendo acerca de las características que definen un formato se llegó a la conclusión de que todo formato tiene que corresponderse con alguna especificación formal o informal, no puede tratarse de un diseño de bits al azar o sin previa documentación. La definición de Wikipedia, «una manera particular de codificar información para almacenarla en un archivo informático», no parece enfatizar lo suficiente esta característica. El grupo esbozó su propia definición: *una estructura específica y preestablecida para la organización de un fichero digital o cadena de bits*".

Diccionario de Datos PREMIS de Metadatos de Preservación

“Esta estructura preestablecida incluye la forma en que están codificados los datos y la forma en la cual los bits son interpretados para producir texto, imágenes y sonido”.

Formatos: problemas

El problema deviene entre otras cosas de la codificación elegida porque: *sólo en algunos casos la codificación es sinónimo de un formato específico de archivo; por ejemplo, la codificación mp3 es utilizada para codificar un formato de archivo .mp3,* pero los archivos de texto plano pueden tener formatos con codificaciones diferentes, ya que pueden ser codificados p.e. como ASCII o Unicode, entre un gran número posible de variantes, esto se complejiza con los archivos de imagen, música o video, más aún porque formatos como como *TIFF, WAV y AVI son formatos contenedores que están diseñados para combinar cadenas de bits de naturaleza distinta en un sólo archivo.*

Diccionario de Datos PREMIS de Metadatos de Preservación

"El formato es, evidentemente, una propiedad de los ficheros, pero también puede aplicarse a las cadenas de bits. Por ejemplo, una cadena de bits de una imagen dentro de un fichero TIFF podría tener un formato acorde a la especificación del formato del fichero TIFF. Por este motivo, PREMIS evita utilizar el término formato de fichero y emplea en su lugar formato, más genérico".

Diccionario de Datos PREMIS de Metadatos de Preservación

"Un repositorio debe registrar la información sobre el formato de la manera más específica posible. Lo ideal sería identificar los formatos con un enlace directo hacia la especificación completa del formato. En la práctica, es más cómodo un enlace indirecto como un código o una cadena que pueda a su vez asociarse con las especificaciones completas del formato".

Selección de formatos: generalidades

Para asegurar la preservación de la información, el formato elegido debe ser legible por una aplicación durante el mayor tiempo posible. Esto implica evitar formatos propietarios cerrados, como documentos en Word (en cualquiera de las versiones). Lo recomendable es utilizar formatos propietarios pero abiertos, como el formato TIFF, o mejor aún, formatos no propietarios como el ASCII.

Criterios a considerar cuando se eligen formatos

- Ubicuidad
- Asesoría o soporte
- Confidencialidad
- Calidad de la documentación
- Estabilidad
- Fácil identificación
- Derechos de propiedad intelectual
- Soporte de metadatos
- Complejidad
- Interoperabilidad
- Viabilidad
- Reusabilidad

Adrian Brown, “Selecting File Formats.” Disponible en: http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

¿Dónde buscar definiciones claras de los formatos?

El Registro Unificado de Formatos está disponible en: http://www.udfr.org/ establecido a partir de 2009 aúna los esfuerzos del registro PRONOM y el Registro Global de Formatos Digitales ambos dedicados a dar recomendaciones y detallar las características de los distintos formatos de archivo.

UDFR

The technical registry
PRONOM

Details: File format summary

Details for: Acrobat PDF/A - Portable Document Format 1a

Summary

Name	Acrobat PDF/A - Portable Document Format
Version	1a
Other names	PDF/A (1)
Identifiers	MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PUID: fmt/95
Family	
Classification	Page Description
Disclosure	Full
Description	The Portable Document Format/ Archive is a format designed for long term preservation by Adobe Systems. PDF/A is a simplified version of PDF 1.4, with all of the features from PDF 1.4 that would impede long term preservation removed. Removed features include Javascript, Audio/Visual content, LZW compression and encryption. A major principle of PDF/A is that it is self contained and not reliant on externalities thus all font and colour information is encoded into the file. PDF/A files are larger than other types of PDF files due to the need for embedded information. PDF/A supports two levels of compliance PDF/A1-a (Accessible) and PDF/A1-b (Basic). PDF/A1-a is fully ISO 19005-1:2005 (PDF/A-1) compliant whereas PDF/A1-b is less stringent and not compliant. PDF/A1-a requires tagged PDF and Unicode whereas PDF/A1-b does not.
Orientation	Binary
Byte order	Big-endian (Motorola)
Related file formats	Has priority over Acrobat PDF 1.0 - Portable Document Format (1.0) Has priority over Acrobat PDF 1.1 - Portable Document Format (1.1) Has priority over Acrobat PDF 1.2 - Portable Document Format (1.2) Has priority over Acrobat PDF 1.3 - Portable Document Format (1.3)

sedici.unlp.edu.ar

Formatos

Los objetivos perseguidos por un formato específico pueden ser diversos:

Almacenar un solo tipo de contenido plano sin ninguna codificación adicional. Ejemplo: .txt

Incorporar especificaciones para codificar la información (principalmente para su compresión, transmisión o cifrado) Ejemplo: .pdf

Combinar y sincronizar varios tipos de contenido en un solo archivo. Ejemplo: los archivos .mpeg o .AVI que incluyen pistas de audio, vídeo, subtítulos, metadatos, etc.

Formatos: cuestiones a tener en cuenta

¿Su uso está generalizado? ¿Existen varios programas para leer este formato? ¿Está utilizado por otras instituciones como formato de preservación?

¿Está abierto? El uso del formato no debe ser regido por patentes.

¿Está documentado? ¿Fue publicada su documentación (un formato puede estar documentado sin estar necesariamente abierto)? ¿Fue normalizado por instituciones como W3C o ISO? Esta documentación permitirá construir nuevos programas para leer el formato si sus vendedores ya no se encargan de su mantenimiento.

¿Existen programas para validarlo y caracterizarlo?

Cuando se elige el formato más adecuado al contenido, debe valorar los riesgos de

Formatos de archivos para la preservación digital

- Problemática
- Política de la BnF
- Formatos de archivos aceptados en SPAR
- Analizar los formatos

Problemática

El formato de datos digitales es uno de los aspectos más importantes para su preservación. Cuando uno elige el formato más adecuado al contenido, debe valorar los riesgos de obsolescencia que corre el formato. También se debe tomar en cuenta la cuestión jurídica: por ejemplo, el depósito legal impone que la institución acepte todos los datos, sea cual sea su formato. Los mayores criterios para determinar la sostenibilidad de un formato son los siguientes:

- **¿Su uso está generalizado?** ¿Existen varios programas para leer este formato? ¿Está utilizado por otras instituciones como formato de preservación?
- **¿Está abierto?** El uso del formato no debe ser regido por patentes.
- **¿Está documentado?** ¿Fue publicada su documentación (un formato puede estar documentado sin estar necesariamente abierto)? ¿Fue normalizado por instituciones como W3C o ISO? Esta documentación permitirá construir nuevos programas para leer el formato si sus vendedores ya no se encargan de su mantenimiento.
- **¿Existen programas para validarlo y caracterizarlo?**

sedici.unlp.edu.ar

Formatos de texto

El formato más difundido para la publicación de textos es el formato de documento portátil o **PDF**, pero no es raro encontrar autoarchivos de materiales hechos en formatos **.doc** o **.docx** u otro tipo de formatos de texto editable como **.odt.** En estos casos siempre es recomendable la trasformación del material al formato PDF. El formato PDF fue creado por Adobe, y es ahora un estándar abierto y oficial reconocido por la Organización Internacional para la Estandarización (ISO). A los fines de la preservación digital el formato recomendado es el PDF/A. El **PDF/A** es el estándar más común para los documentos de texto con formato, pero muchas entidades que ofrecen contenidos en formatos de texto electrónico en formato **EPUB**. Ambos formatos están basados en XML.

Formatos para texto recomendados

El **PDF/A** se presenta como el estándar aceptado para la creación de documentos digitales accesibles online y susceptibles de ser impresos, tanto aquellos basados en texto como los que incluyen imágenes, gráficos, etc. que requieren de un diseño preciso.

El **EPUB** es el estándar de facto recomendado para el texto electrónico. Aunque puede soportar imágenes está más orientado a la publicación de texto, por ello no es el formato más adecuado para documentos que requieren un diseño preciso o están basados en imágenes.

Ventajas de PDF/A

PDF/A es, de hecho, un subconjunto de PDF obtenido excluyendo aquellas características superfluas para el archivado a largo plazo de forma similar a como se ha definido el subconjunto PDF/X para la impresión y artes gráficas. Además, el estándar impone una serie de requisitos a los programas para la visualización de archivos PDF/A.

Un programa de visualización que se ajuste a los requisitos debe seguir ciertas reglas incluyendo la conformidad con las directrices en cuanto a la gestión de color, el uso de fuentes integradas a la hora de la visualización, o la posibilidad de realizar anotaciones por parte del usuario.

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 (por ejemplo, programas fuente o *streams* de datos), aunque se permite que tengan anotaciones (como hipertextos) que enlacen a documentos externos.

Niveles de cumplimiento PDF/A

PDF/A1 posee dos niveles de cumplimiento:

PDF/A-1a aplica corrección semántica y estructura. Cada carácter debe tener su equivalente Unicode. La estructura se expresa por medio de etiquetas.

PDF/A-1b aplica integridad visual.

Otros elementos de compatibilidad PDF/A

- El contenido de audio y video está prohibido (excepto en PDF/A3 –ISO estándar 3200
- 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.

En síntesis

Para la preservación se utilizan en mayor medida los formatos no propietarios, reconocidos como estándares. Cuando se trata de difundir, y aunque existe una mayor flexibilidad.

Los formatos presentados como válidos para la preservación, también lo son para la difusión de los contenidos aunque se suele aplicar algún tipo de compresión o se reduce de calidad en pos de su funcionalidad.

La principal divergencia encontrada en los formatos utilizados en la industria y en el ámbito de la preservación, reside en la utilización de medios de protección técnica.

Preservación a nivel semántico

Autenticidad, Interpretabilidad – preservación semántica

Capa Semántica: ¿cómo asegurar de entender/interpretar correctamente los datos?

¿Qué se puede hacer?

Preservación a nivel semántico

Amenazas a nivel semántico

- Cambio de significado de los términos: nombres de ciudades, …
- Escalas de medición, sensibilidad de sensores, …cambio
- Cambios de interpretación de los hechos: niveles de alcohol, …
- Datos dependientes del contexto: recalibración/sensor drift

Más bien a largo plazo, difícil de notar ¡y por ello peligroso!

Considerar el contexto de los objetos

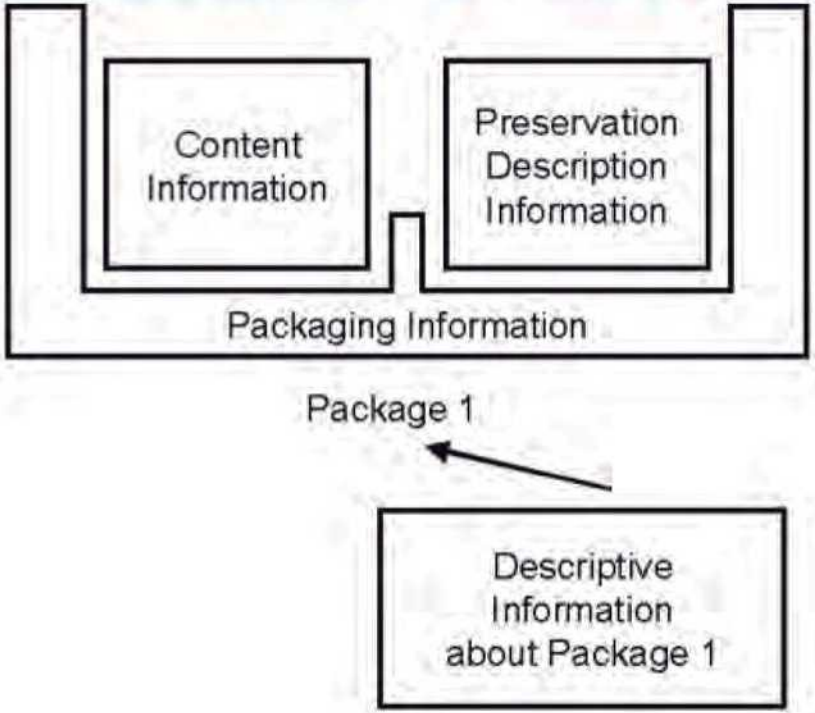
- Propósito, configuración, limitaciones, contexto cultural, objetos relacionados…

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).

Paquete de Información: IP: un elemento central



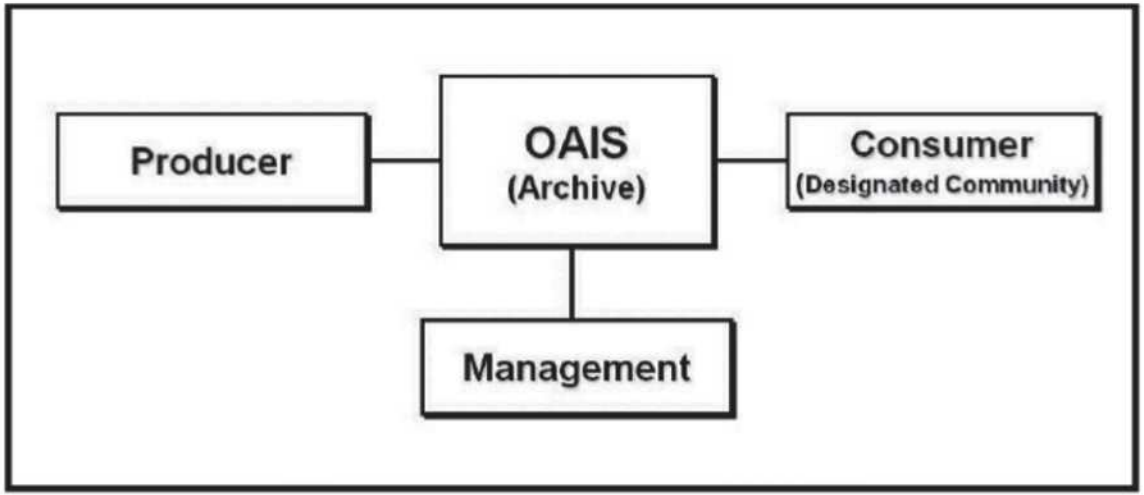
ISO 14721 (2012): Fig 2-3: Paquete de información: conceptos y relaciones

sedici.unlp.edu.ar

Definición del paquete de información

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**.

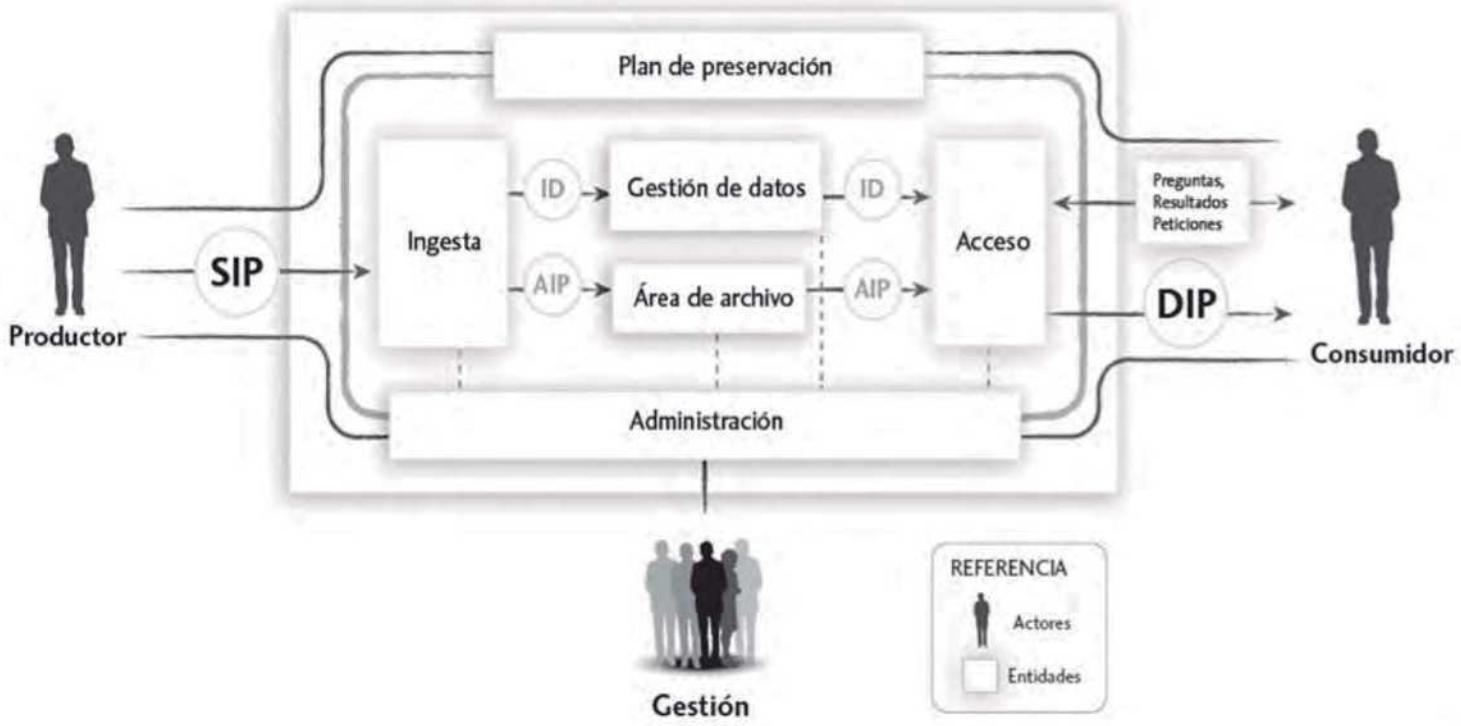
Actores en el OAIS



ISO 14721 (2012): Fig 2-1: Enviroment Model of an OAIS

sedici.unlp.edu.ar

Entidades funcionales



Basado en la Fig 4-1: OAIS Functional entities. ISO 14721 (2012)

El OAIS

- Es mucho más que una norma cuyo único eje es el paquete de información es decir el objeto digital, su representación y metadatos de la información descriptiva de la preservación o la información descriptiva que sirve para localizar el IP
- Las entidades funcionales del modelo abstracto describen las funciones esperables de un archivo para lograr el objetivo de la preservación a largo plazo y el acceso
- El estudio detallado del modelo abstracto nos pone en claro las funciones deseadas y las acciones que deben ejecutarse

Descripción simplificada de las funciones de las entidades

Ingesta/Ingreso (*Ingest*)

Proporciona los servicios y las funciones para aceptar un SIP de los productores, y preparar los contenidos para el almacenamiento y gestión en el archivo, a saber:

1. recepción,
2. aseguramiento de la calidad,
3. generación de AIP,
4. extracción de información descriptiva para su incorporación a la base de datos del archivo y contenido al archivo.

➢Ver p.e. ISO 16363: 4.2 Ingest: Creation of the AIP

Descripción simplificada de las funciones de la entidad Archivo (*Archival Storage)*

Proporciona los servicios y funciones para el almacenamiento, mantenimiento y recuperación de los AIP:

1. Recepción de los AIP desde Ingest e incorporación al almacenamiento permanente,
2. gestión de la jerarquía de almacenamiento,
3. renovación (*refreshing*) de medios de almacenamiento,
4. realización comprobaciones de error rutinarias y especiales,
5. recuperación ante desastres
6. Entrega del AIPs a la entidad funcional de Acceso para satisfacer los pedidos.

Descripción simplificada de las funciones de la entidad Archivo (*Archival Storage)*

4. proporciona informes periódicos de análisis de riesgos, supervisa cambios en el entorno tecnológico, así como en los requisitos de servicio para la comunidad designada.
6. Diseña plantillas de empaquetado de información (I*nformation Package)* y proporciona ayuda al diseño y revisión para adecuar estas plantillas a los SIP y los AIP.
7. Desarrolla planes detallados de migración, prototipos de software y planes de prueba para permitir la ejecución de las metas de migración de la Administración.

Descripción simplificada de las funciones de las entidade

Gestión de datos (*Data management)*

Proporciona los servicios y funciones para generar, mantener y acceder a la información descriptiva (*Descriptive Information*) que identifica y documenta los fondos de archivo, así como los datos administrativos usados para administrar el archivo:

1. administra y actualiza las funciones de la base de datos del archivo,
2. realiza consultas sobre los datos de la gestión de datos para generar respuestas,
3. y elabora informes con esas respuestas.

➢Ver p.e. ISO 16363: 4.5 Information Management

Descripción simplificada de las funciones de las entidades

Administración *Administration*

Proporciona los servicios y funciones para la operación global del sistema de archivo:

1. solicitud y negociación de acuerdos de envío (*submission agreements*) con los productores,
2. auditoría de las remisiones para asegurar que cumplen con los estándares del archivo,
3. mantenimiento de la gestión de la configuración del sistema de hardware y software.

También proporciona funciones de ingeniería de sistemas para supervisar y mejorar las operaciones del archivo, y para inventariar, informar sobre y migrar/actualizar los contenidos del archivo. Es responsable de establecer y mantener las normas y políticas del archivo.

Descripción simplificada de las funciones de las entidades

Planificación de la Preservación *Preservation Planning*

Proporciona los servicios y funciones para supervisar el entorno del OAIS, así como recomendaciones y planes de preservación para asegurar que la información almacenada en el OAIS permanece accesible y comprensible a largo plazo para la comunidad designada,:

1. evalúa los contenidos del archivo y recomienda periódicamente actualizaciones,
2. recomienda las migraciones de archivo,
3. desarrolla recomendaciones de normas y políticas de archivo,

➢Ver p.e. ISO 16363: 3.1.2 (Strategic Plan) y 4.3 Preservation Planning

Descripción simplificada de las funciones de las entidades

Acceso *Access*

Proporciona los servicios y funciones que ayudan a los consumidores a determinar la existencia, descripción, ubicación, accesibilidad y recepción de la información almacenada en el OAIS:

1. Se encarga de la comunicación con los consumidores para recibir solicitudes,
2. aplica controles para limitar el acceso a información especialmente protegida,
3. coordina la ejecución de las solicitudes para su finalización con éxito,
4. generar y entrega respuestas a los usuarios (DIP)

➢Ver p.e. ISO 16363: 4.6 Access Management

En resumen: qué se precisa para la preservación

Una profunda comprensión de los objetos digitales con todo su entorno

A lo largo de todo el ciclo de vida

Realizado de manera sistemática

Una estructura confiable de flujos de trabajo

Requiere vigilancia para asegurar el acceso

Cuanto más complejos se vuelven los objetos más complicado es

Se necesita definir un plan de preservación

Acciones

Perfil de los objetos digitales presentes en el repositorio

Acciones de estandarización, migración, curación en general

Vigilancia para evitar obsolescencia

Todo el ciclo de vida del objeto digital

Muchos metadatos

Plan de preservación que a su vez dispara nuevas acciones

Decisión

Ó elegir alternativas...

La implementación no es única, existen diferentes propuestas a analizar y elegir:

Muchas herramientas: modelos más eclécticos

Un sinnúmero de posibilidades

Uso de una implementación dedicada a la preservación exclusivamente

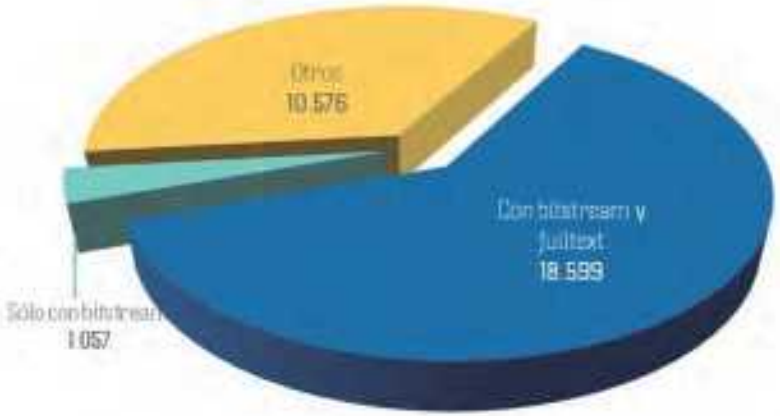
Acciones realizadas 2014-2015

Revisar los objetos digitales y comprobar la existencia de muchos metadatos

De Giusti Marisa. Tesis doctoral: "UNA METODOLOGÍA DE EVALUACIÓN DE REPOSITORIOS DIGITALES PARA ASEGURAR LA PRESERVACIÓN EN EL TIEMPO Y EL ACCESO A LOS CONTENIDOS".
http://sedici.unlp.edu.ar/handle/10915/43157

Acciones

- Evaluar los objetos del repositorio como IP
- Generar un reporte sobre su estado.
- Informar si cuentan con los elementos que la norma define para el IP.
- 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 IP está bien formado el repositorio es funcional en sus procesos (ingesta, preservación y entrega).



ITEMS EN TOTAL: 30.232

Herramientas usadas 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.

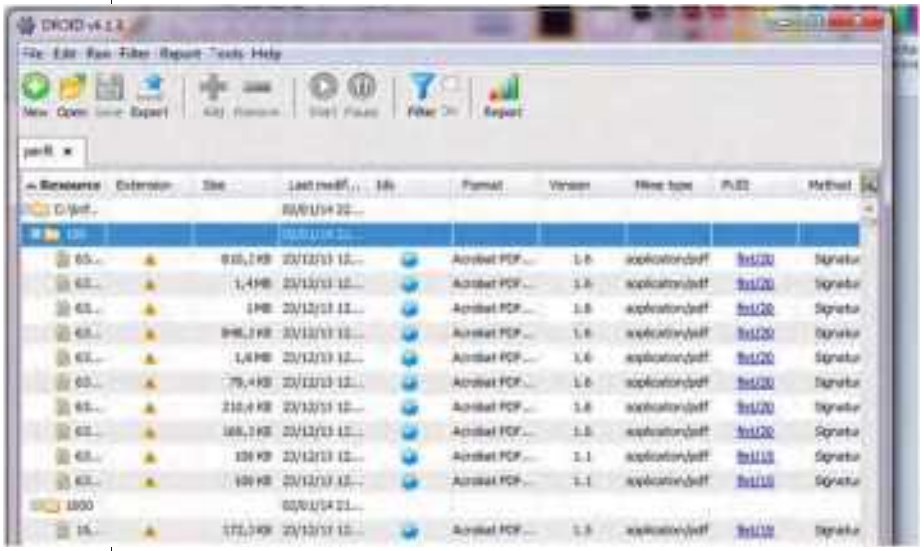
El paquete de información en el OAIS

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.

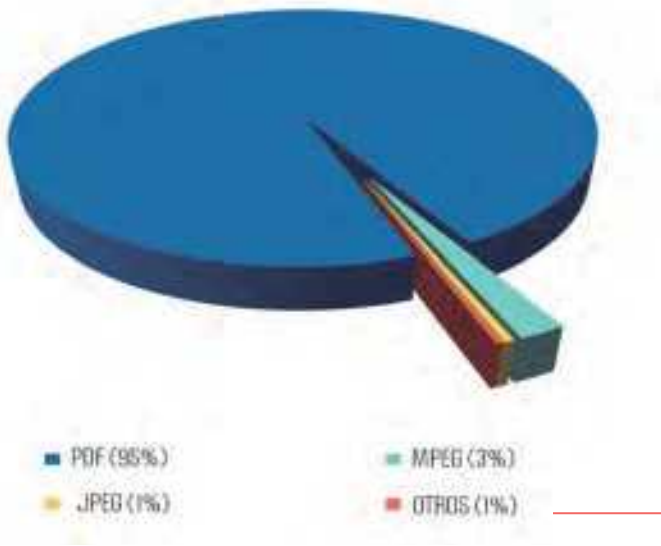
Perfil en DROID

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



Formatos-Mime Types

El gráfico representa de manera simplificada el perfil elaborado por DROID



SHOCKWAVE – FLASH
POWERPOINT
BMP
QUICKTIME
HTML
EPUB-ZIP
BINARIO
MS EXCEL
ODS
EPUB
OGG

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

- Migrar a PDF/A1-a en su defecto PDF/A1-b.
- En el caso de los libros, utilizar la opción de optimización porque incide en numerosos aspectos (correcta recuperación del texto en búsquedas a texto completo, óptima visualización en web. Hay que cumplimentar ambos objetivos: que el archivo cumpla con el estándar y que quede correctamente optimizado.
- 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 formatos abiertos.
- Extraer automáticamente los metadatos técnicos (FITS) y guardarlos en la base de datos acompañando al bitstream correspondiente.

El paquete de información en el OAIS

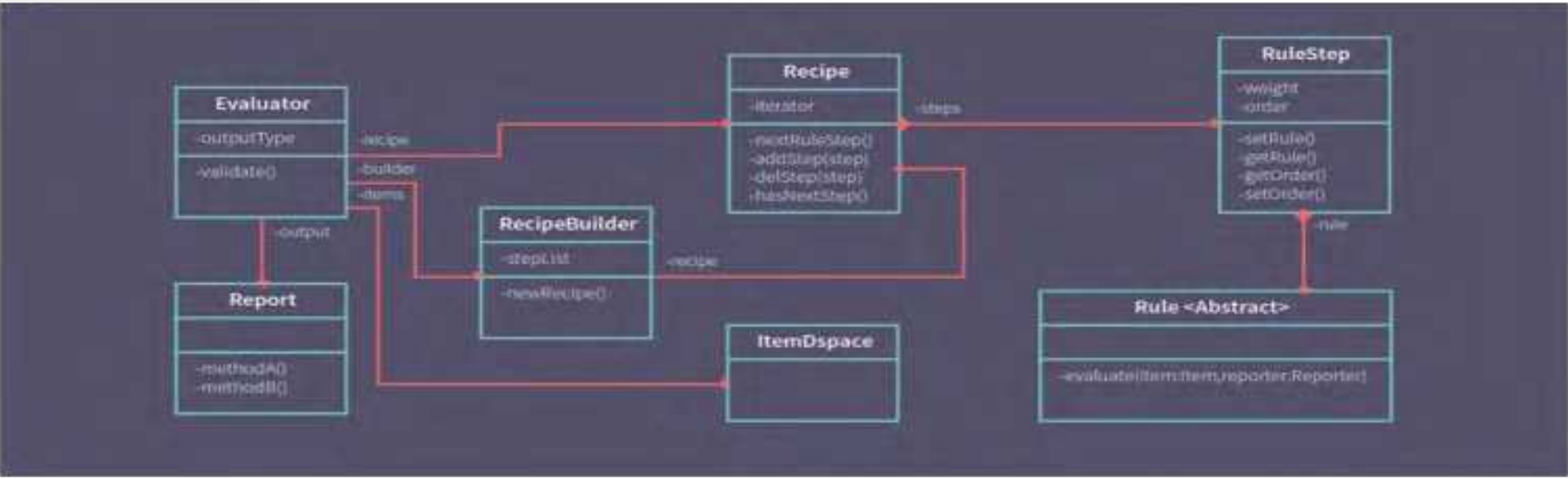
La Información descriptiva de la preservación

- 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.
- Propuesta: desarrollo de un validador a modo de una tarea de curación aplicable a nivel de ítem, colección o comunidad con el propósito de efectuar el el mantenimiento de ítems a lo largo de todo su ciclo de vida de los ODs

Metodología propuesta

- Desarrollo de un validador a modo 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 fue efectuar el mantenimiento de ítems en el tiempo y a lo largo de todo su ciclo de vida de los ODs

Validador



Enfoque: DSpace

Ejemplo de tarea: "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:

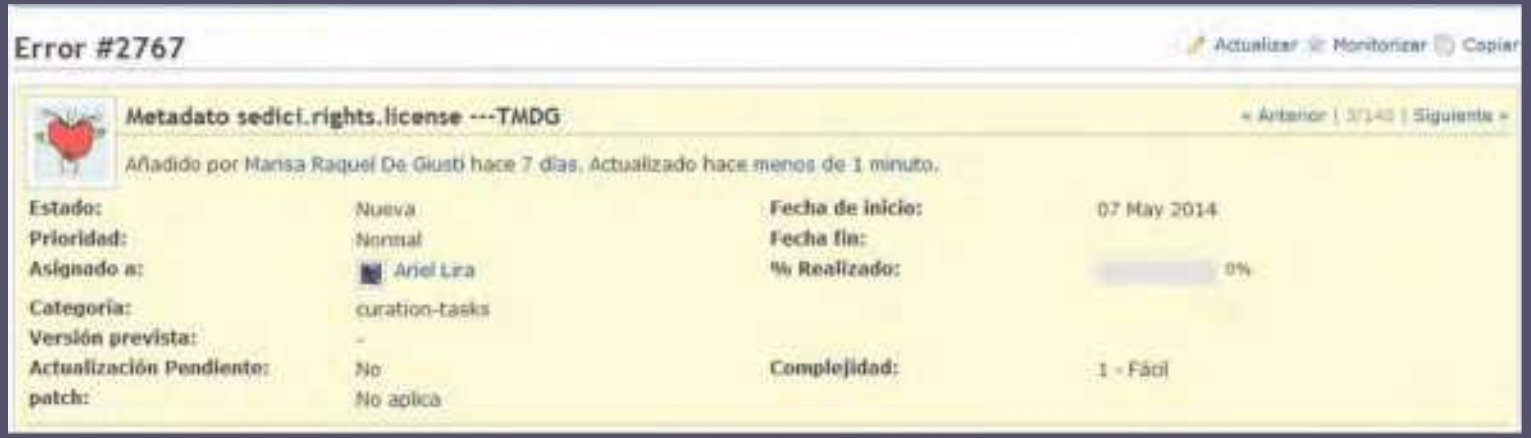
✓ True → Válido

× False → Inválido

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 resultados y acciones



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.

Análisis de la información descriptiva

Verificaciones siguiendo las directrices DRIVER

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 items 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: conclusiones

- Estuvo centrada en el contenido, la representación del contenido (formato), la PDI y la información descriptiva.
- Se verificaron muchos metadatos.
- La visión integral de los contenidos permitió 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 generaron 104 tickets en el gestor de incidencias de SEDICI (algunos involucran cientos y hasta miles de ítems).

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.

Qué trabajos futuros se plantearon

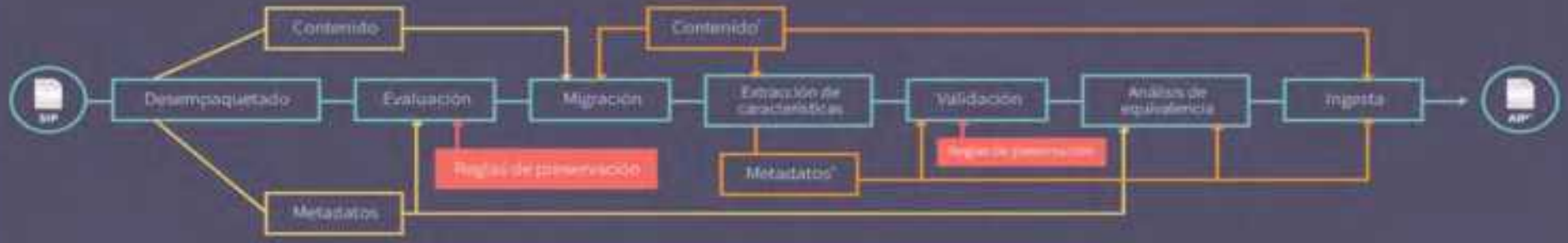
- Análisis de las migraciones masivas y automatizadas. Problemas.
- Selección de herramientas de migración.
- Reporte de los eventos de migración y sus responsables de modo de asegurar la trazabilidad en el ciclo de vida (evento y agente en PREMIS).
- En relación a las posibles "capas" de las reglas, generar reglas más abstractas, pensadas más allá de que el repositorio se encuentre implementado en DSpace o en cualquier otro software para repositorios. Nivel repositorio.
- Implementación de un lenguaje de sintaxis sencilla que permita a la administración del repositorio, realizar controles sobre los items de manera individual, por colección, etcétera y obtener un reporte.

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

Plato 4.4

SCAPE

marisadg

Welcome to Plato 4.4! [M. Kraxner06.03.2014 13:14]

A copy has been created: *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)*
It is marked as playground. If you want to use it for serious planning, please change this in Plan Settings. [PLATO11.03.2014 12:59]

My Plans

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	

sedici.unlp.edu.ar

Digital Preservation

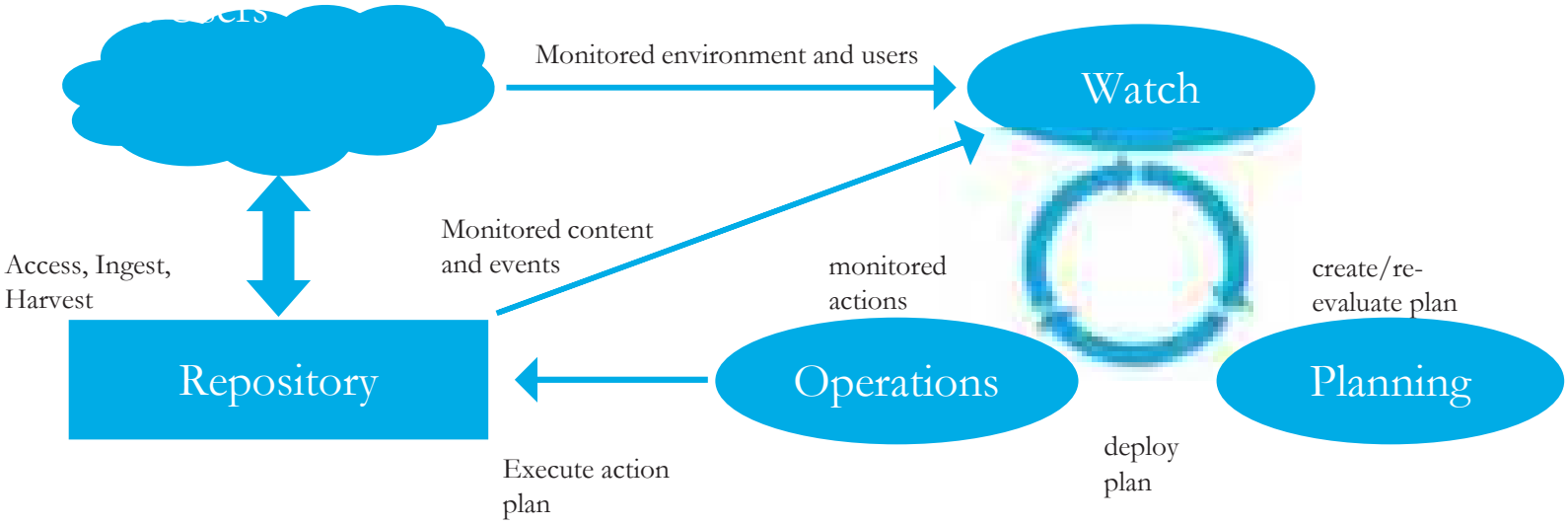
Information Management and Preservation (IMP)
Information and Software Engineering Group (IFS)
Institute of Software Technology and Interactive Systems (ISIS)
Vienna University of Technology

IMP Digital Preservation Publications Software

- Digital Preservation
- Software Tools for Processes
- Benchmarking
- Planning
- Content Profiling
- Monitoring
- Process Management Plans
- Data Citation
- Escrow
- more

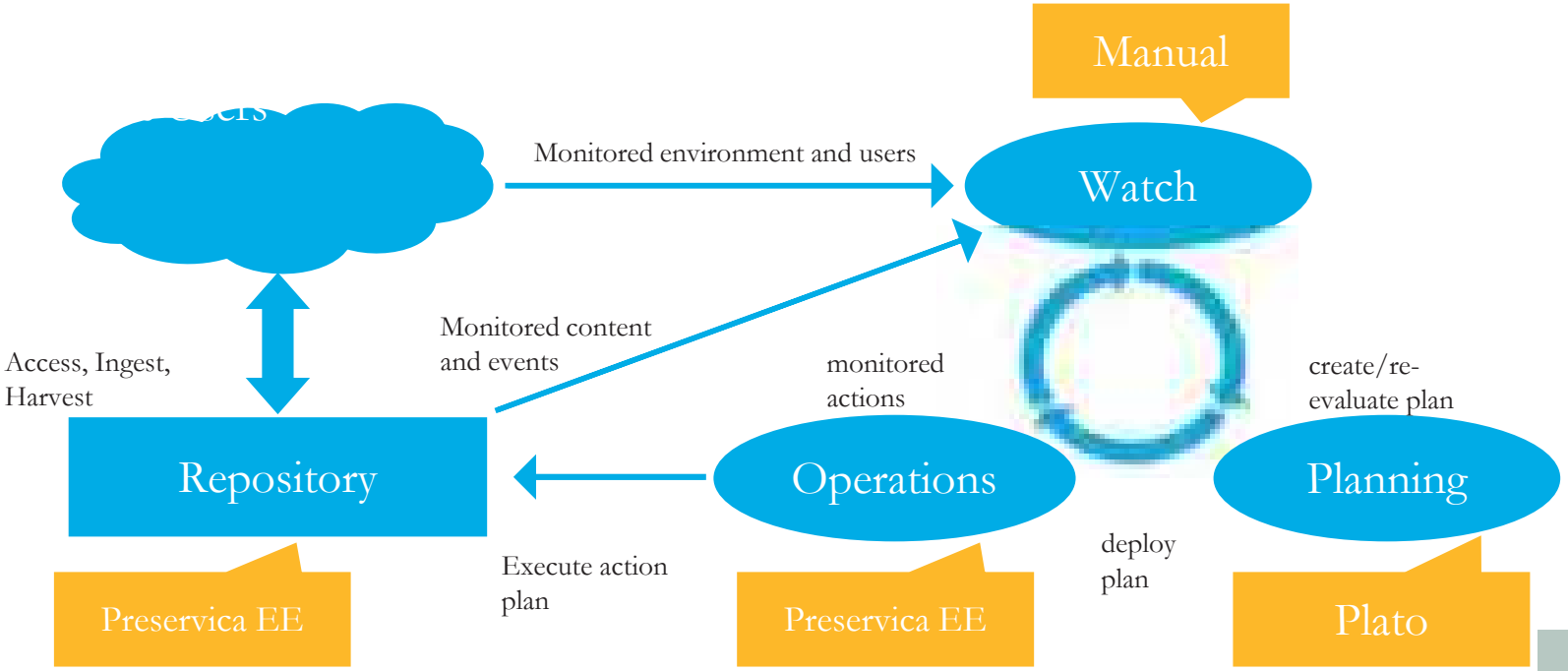
sedici.unlp.edu.ar

Digital preservation lifecycle del SPE



Miguel Ferreira. Report on compliance validation. Scape. Mayo de 2014

Digital preservation lifecycle



Andreas Rauber, Hannes Kulovits. *Digital Preservation*. Vienna University of Technology & Austrian State Archives Vienna, Austria. Biredial ISTEC 2015, Universidad del Norte, Colombia. Octubre 2015.

Preservica

- Preservica permite una preservación activa
- Gestiona la migración de los objetos digitales en el repositorio
- Tiene un registro interno de formatos basado en PRONOM
- Realiza de manera separada el proceso de planeamiento de la preservación
- Ejecuta el plan en Preservica
- Selecciona los objetos en el repositorio:

– Formatos de archivo

– Riesgos

Preservica

- Análisis (“caracterización”)
    - Identificación
    - Extracción de características
    - Validación
- Acciones
    - Migración, emulación

Aseguramiento de la calidad

- Autenticidad y propiedades significativas

Metadatos

Reportes

Preservica y DSPACE

Preservica
Digital Preservation

Webinars | Upcoming Events | User G

Why Preservica? | How it Works | Preservica Editions | News | Resc

Knowledge Center / DSpace Ingest Workflow

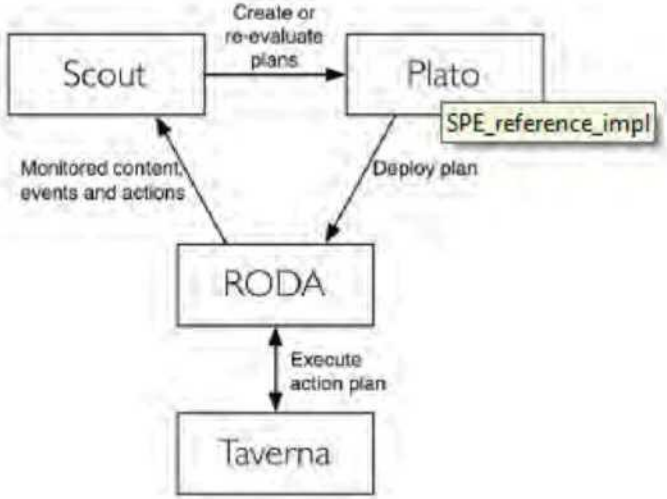
The DSpace Ingest Workflow automates the bulk import of DSpace packages (records, files and metadata) into Preservica for long term active preservation in the cloud, ensuring valuable DSpace content and collections are safeguarded against loss, degradation and file format obsolescence and remain accessible and readable for decades to come.

Retomando SCAPE: Modelo SPE

El mediombiente de preservación SCAPE es un sistema acoplado ligeramente que permite extender un sistema de repositorio existente (p.e RODA) con varios componentes que cubren: el perfilamiento de la colección (p.e C3PO) monitoreo de preservación (p.e SCOUT) y planeamiento de la preservación (p.e Plato) estos componentes abordan las funcionalidades clave definidas por el OAIS. La mayoría de los sistemas de repositorios existentes carecen de estas funcionalidades o las ejecutan de manera superficial.

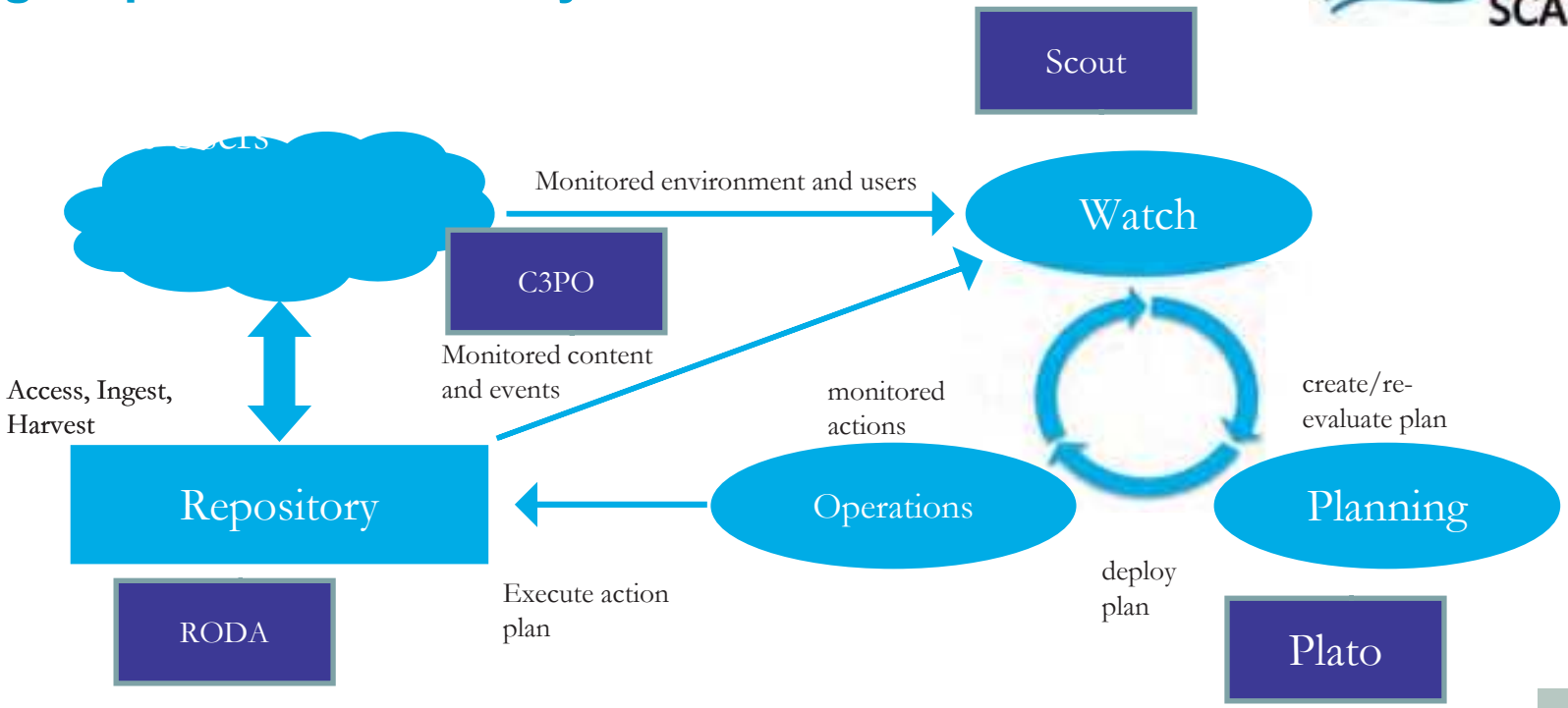
Miguel Ferreira. Report on compliance validation. Scape. Mayo de 2014

SPE



Miguel Ferreira. Report on compliance validation. Scape. Mayo de 2014

Digital preservation lifecycle



Andreas Rauber, Hannes Kulovits. *Digital Preservation*. Vienna University of Technology & Austrian State Archives Vienna, Austria. Biredial ISTEC 2015, Universidad del Norte, Colombia. Octubre 2015.
Kresimir Duretec, Artur Kulmukhametov, Michael Kraxner, Markus Plangg, Christoph Becker, and Luis Faria. 2014. The SCAPE preservation lifecycle. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '14). IEEE Press, Piscataway, NJ, USA, 425-426.

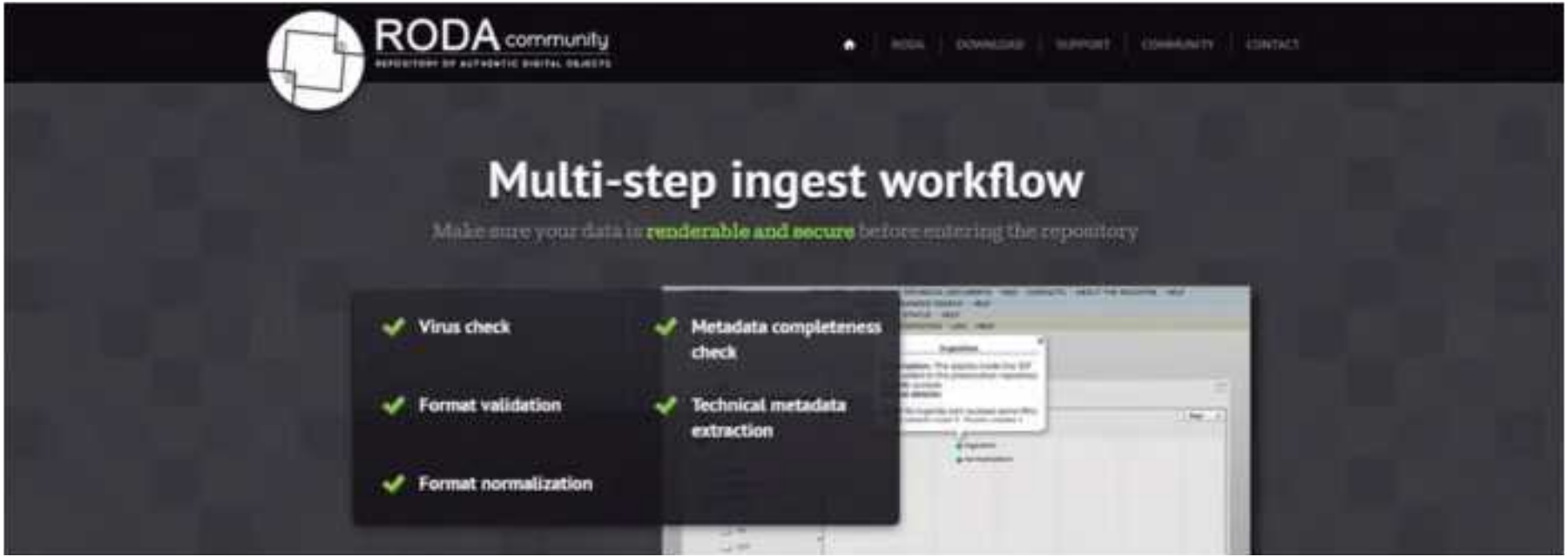
RODA: El contenido digital en la SPE es administrado por RODA, el sistema de depósito elegido para integrar los diversos componentes funcionales desarrollados en SCAPE.

Scout: El proceso de supervisión / watch es implementado por Scout. Una base de conocimientos que centraliza toda la información necesaria para detectar los riesgos de preservación.

Plato: El proceso de planificación es realizado con Plato, una herramienta que sistematiza la planificación de la conservación. Permite la definición de los objetivos de conservación, criterios y restricciones necesarias para la toma de decisiones.

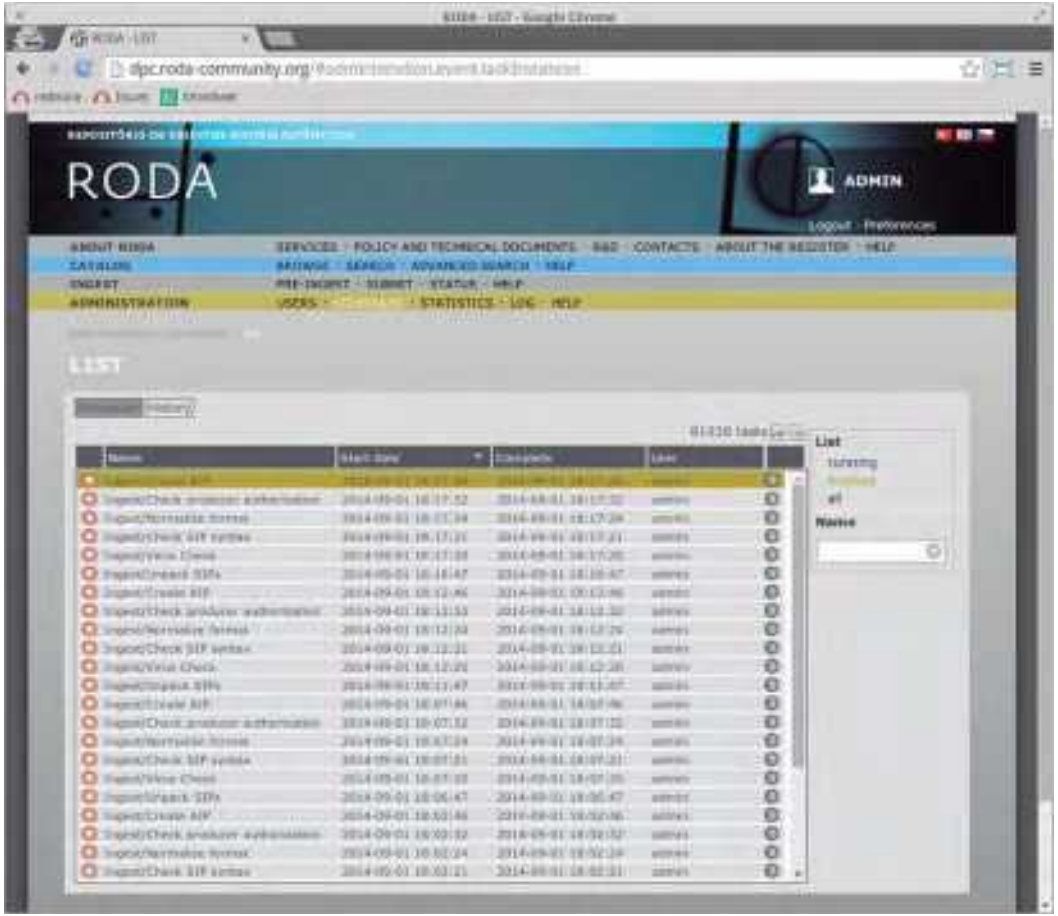
Taverna: el proceso de las operaciones lo soporta Taverna, un sistema de gestión que permite la ejecución de flujos de trabajo complejos de preservación: herramientas de caracterización, de migración y de garantía de calidad.

RODA: an Open source Digital repository designed for preservation



http://www.roda-community.org/

RODA



http://www.roda-community.org/

RODA

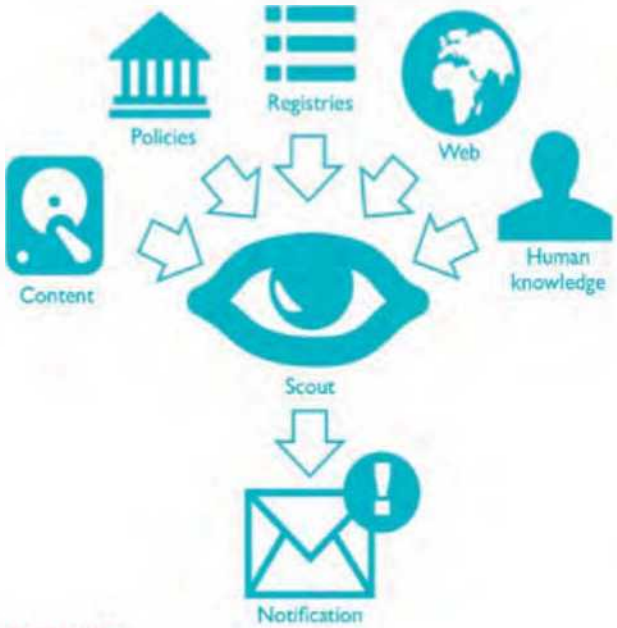
http://www.roda-community.org/

RODA

http://www.roda-community.org/

Scout

Scout – a preservation watch system



SCOUT

Preservation Watch System

What is SCOUT?
SCOUT is a preservation watch system that monitors the whole world to warn you of preservation risks or opportunities.

Why did I received this email?
A trigger was created to warn you when something important happened. Your email was defined as a notification that this important event has occured.

There is a notification for you!

Request: Check collection policy conformance

Question

Question assessed on this trigger.
SPARQL:?s watch:entity ?colle[illegible]:property ? compressionSchemeDist ; watch:stringDictionaryValue ?value . ?compressionSchemeDist watch:id "cl-KDNE_rjmuRKxjIhnqeHpYgnw"^^xsd:string . ?value ?l ? dictionaryItem . {?dictionaryItem watch:key ? compressionType1 . ?policy1 a cp:FormatObjective ; cp:measure measure:117 ; cp:value "none"^^xsd:string . FILTER regex(?compressionType1, "^(Unknown|Uncompressed)") } UNION {?dictionaryItem watch:key ?compressionType2 . ?policy2 a cp:FormatObjective ; cp:measure measure:117 ; cp:value "lossless"^^xsd:string . FILTER regex(?compressionType2, "^(Conflicted|JPEG)") }
Target: PROPERTY_VALUE

http://openplanets.github.io/scout/

Scout

Scout es capaz de monitorear el perfil de un dado contenido, es decir brindar un resumen rápido que permite la caracterización de un archivo.

SCOUT no hace los análisis o la extracción de información de los objetos digitales, sino que se basa en diversas herramientas que realizan estas tareas puntuales, cuyos resultados son luego almacenados, analizados y expuestos. Para hacer esto, SCOUT utiliza un sistema de plugins propio que simplifica la integración con nuevas fuentes de información, registros de formatos de archivos, herramientas para caracterización de archivos, sistemas de migración y aseguramiento de la calidad, políticas, conocimiento humano, entre otros.

Scout

La base de conocimiento generada puede ser explorada fácilmente, y es posible instalar disparadores (triggers) que se ejecutan ante determinados eventos y permiten notificar automáticamente a los usuarios sobre nuevos riesgos u oportunidades. Por ejemplo, una notificación podría ser que el contenido no se ajusta a las políticas definidas, que un formato se volvió obsoleto, o que nuevas herramientas para mostrar el contenido están disponibles.

Scout

Scout no hace los análisis o la extracción de información de los objetos digitales, sino que se basa en diversas herramientas que realizan estas tareas puntuales, cuyos resultados son luego almacenados, analizados y expuestos. Para hacer esto, SCOUT utiliza un sistema de plugins propio que simplifica la integración con nuevas fuentes de información, registros de formatos de archivos, herramientas para caracterización de archivos, sistemas de migración y aseguramiento de la calidad, políticas, conocimiento humano, entre otros.

Scout

El sistema de plugins de SCOUT utiliza unos pequeños módulos, llamados adaptadores (adaptor), a partir de los cuales le es posible interactar con otras aplicaciones. Un adaptador se encarga de recibir órdenes desde SCOUT y traducirlas al lenguaje, protocolo o formato que espera la aplicación, y a su vez recibe la salida generada por la aplicación y la transforma en un formato que SCOUT es capaz de comprender y procesar. Existe un adaptador para C3PO, otro para conectarse con el registro PRONOM, uno para consultar APIs de reportes de repositorios y parsear los eventos registrados.

Open Preservation Foundation

Sign in | Create Account

Search

Demo de C3PO: https://vimeo.com/53069664

Technology ▼

Adopt our digital preservation tools and collaborate in open source development.

Knowledge ▼

Discover best practice through our blogs, training, events, and interest groups.

Our Organisation ▼

Meet the team, members, and partners behind our strategy and projects.

Join the OPF

Knowledge

- Events
- News
- Interest Groups
- Blogs
- Surveys
- Training

C3PO: a content profiling tool for preservation analysis

In the last months, I have been researching the problem of large-scale content profiling for preservation analysis. I do this for a number of reasons. For one, I support the opinion that formats are just another property. Undoubtedly, a very important one, but knowing which formats you have is not sufficient for good preservation planning and actions. I believe a good content profile sets the foundation of a preservation plan and helps reduce bias during the experiments phase. And lastly, it is a great source for preservation monitoring, but more on that later.

In this blog post I present a content profiling prototype tool called Clever, Crafty Content Profiling of Objects – c3po.

What is c3po?

c3po is a tool that deals with meta data of digital objects and helps you to get an idea of what you have to deal with. It consists of two parts, a CLI (Command Line Interface) application and a Web Application. The CLI app reads in and processes FITS meta data files and stores them in a document store. The Web Application offers visualisation, filtering, export of the data and much more.

Share this page

Latest news

nestor and the Open Preservation Foundation sign Memorandum of Understanding

The Open Preservation Foundation and nestor have signed a Memorandum of Understanding with the aim to facilitate discussion between the two organisations and to co-operate on activities to promote digital preservation. Sabina Kistner Hidalgo, Head of the nestor Office said: "nestor and the OPF share central goals

sedici.unlp.edu.ar

What is Plato?

Plato is a decision-making support tool, which:

- guides you through the preservation planning workflow
- uses external information sources to increase efficiency
- utilises controlled experiments
- documents the process.

Andreas Rauber, Hannes Kulovits. *Digital Preservation*. Vienna University of Technology & Austrian State Archives Vienna, Austria. Biredial ISTEC 2015, Universidad del Norte, Colombia. Octubre 2015.

Documentación de Plato: http://www.ifs.tuwien.ac.at/dp/plato/documentation/

Taverna

myGrid

Google™ Custom Search

Introduction | Documentation | Download | Developers | Cite | News | Collaborations | About

Taverna has now moved to the Apache Software Foundation. For updated information, see Apache Taverna (incubating).

Taverna Workflow Management System

Powerful, scalable, open source & domain independent tools for designing and executing workflows. Access to 3500+ resources.

RECENT NEWS

- The Apache Taverna (incubating) team is pleased to announce the release of:
- BioVeL – SEEK and Taverna addressing climate change
- Google Summer of Code Taverna Projects
- Apache officially given control of Taverna

Workbench | Server | Player | Command Line | Taverna Online

COMMUNITY

- Taverna for astronomy, bioinformatics, biodiversity, digital preservation
- Workflow components
- Taverna 3 OSGi
- Taverna Online
- Next generation sequencing on Amazon cloud
- Taverna-Galaxy

Taverna

- Es un sistema de gestión de flujos de trabajo, compuesto por un conjunto de herramientas que permiten diseñar y ejecutar estos flujos de trabajo.
  - Herramientas:
    - Aplicación de escritorio (workbench): ofrece un entorno gráfico para crear, editar y ejecutar flujos de trabajo en la computadora donde se está ejecutando. Permite acceder a servicios via REST o SOAP, y ofrece herramientas desde la línea de comandos.
    - Línea de comandos: esta herramienta permite la ejecución de workflows desde un intérprete de comandos (terminal). Es una versión del workbench, a la cual se le eliminó la interfaz gráfica (GUI), lo cual permite la ejecución desde sistemas sin interfaz gráfica o desde una conexión remota (por ej. via ssh).

Taverna

- Servidor Taverna: el objetivo del servidor es contar con un espacio dedicado para la ejecución remota de flujos de trabajo. Permite la carga (upload) y ejecució de workflows , provee acceso a interfaces REST y SOAP, brinda una comunicación segura (encriptada) con los usuarios así como también aislación entre usaurios y entre flujos de trabajo en ejecución, permite monitorear el uso del sistema y soporta eficiencia y robustez para la ejecución de workflows, soportando incluso caídas temporales (ej. reinicios) del servidor.

En la dirección http://galaxy.nbic.nl/ hay un servidor Taverna, llamado Taverna Galaxy.
Wiki de Taverna Galaxy: https://trac.nbic.nl/elabfactory/wiki/eGalaxy
Experiencias del uso de Taverna Galaxy: http://bergmanlab.smith.man.ac.uk/?p=943-
Taverna Player: es una interfaz web para ejecutar workflows en un servidor Taverna desde un navegador. Los workflows pueden ejecutarse con datos que había cargado el creador al momento de diseño, o con datos cargados por el usuario que ejecuta el player. Los workflows se acceden en modo "solo lectura", ya que Taverna Player no permite su edición (sólo los datos de la ejecución pueden alterarse).

Referencia: Villarreal, Gonzalo L. “Dichos sobre Taverna”. Reporte técnico PREBI-SEDICI- UNLP.

SPE: conformidad con ISO 16363

SCAPE

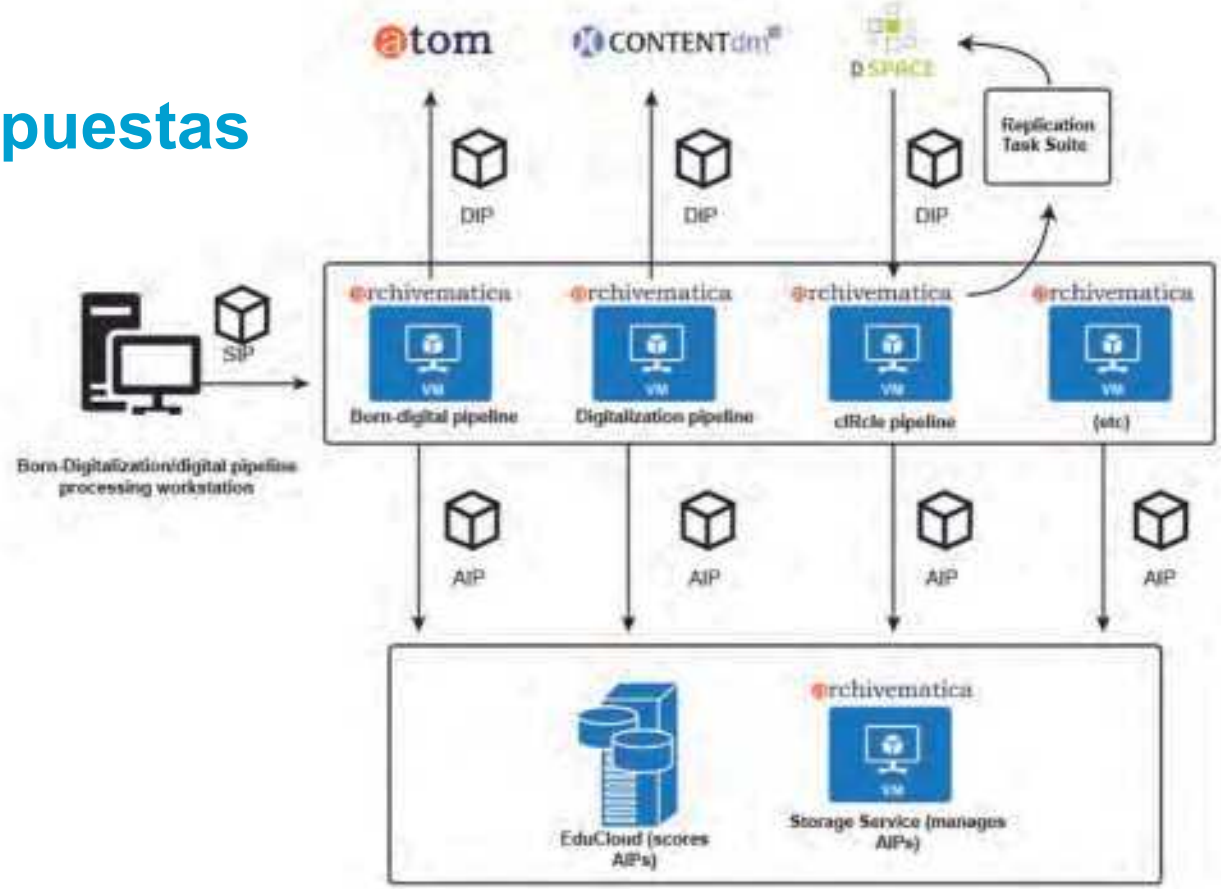
Table 1 - Assessment results.

Metric/Requirement	FULLY SUPPORTED	PARTIALLY SUPPORTED	NOT SUPPORTED	OUT OF SCOPE
Organizational Infrastructure	3	1	5	16
Governance & organizational viability	-	-	-	5
Organizational structure & staffing	-	-	-	4
Procedural accountability & preservation policy framework	3	-	-	4
Financial sustainability	-	-	-	3
Contracts, licenses, & liabilities	-	1	5	-
Digital Object Management	56	0	1	1
Ingest: acquisition of content	10	-	-	-
Ingest: creation of the AIP	27	-	1	-
Preservation planning	6	-	-	-
AIP preservation	5	-	-	1
Information management	4	-	-	-
Access management	4	-	-	-
Infrastructure and Security Risk Management	10	1	0	13
Technical infrastructure risk management	10	1	-	9
Security risk management	-	-	-	4
Totals	69	2	6	31

sedici.unlp.edu.ar

Miguel Ferreira. Report on compliance validation. Scape. Mayo de 2014

Otras propuestas



Philip C. Bantin. *Building Trustworthy Digital Repositories. Theory and Implementation*. Capítulo 9. IMPLEMENTATION: Building a Preservation Strategy around Archivematica. A cargo de Bronwen Sprout y Sarah Romkey.

Archivemática

Archivemática

Basado en estándares

Archivematica es un conjunto de herramientas de software libre que permiten al usuario procesar objetos digitales desde que son introducidos en el sistema hasta su publicación acorde al modelo funcional ISO-OAIS. El usuario puede monitorear y controlar los micro-servicios de ingestión y preservación a través del panel de control. Archivematica utiliza estándares como METS, PREMIS, Dublin Core, la especificación BagIt (Library of Congress) entre otros estándares reconocidos internacionalmente con el objetivo de generar fiablemente paquetes AIPs (Archival Information Package) para ser grabado en su sistema de almacenamiento preferido.

Ejemplo

cIRcle is the University of British Columbia's digital repository for research and teaching materials created by the UBC community and its partners. Materials in cIRcle are openly accessible to anyone on the web, and will be preserved for future

Digital POWRR Tool Evaluation Grid	Ingest					Processing						Access		Storage					Maintenance			Other		
	Copy	Fixity Check	Virus Scan	File Dedupe	Auto Unique ID	Auto Metadata Creation	Auto Metadata Harvest	Manual Metadata	Rights Management	Package Metadata	Auto SIP Creation	Public Interface	Auto DIP Creation	Auto AIP Creation	Reliable, Long-Term Bit Preservation	Redundancy	Geographically Dispersed Data Storage Model	Exit Strategy	Migration	Monitoring	Auto Recovery	Open Source	Clear Documentation	Cost
**Archivematica	x	x	x		x	x	x	x	x	x	x		x	x					x			x	x	Free
Preservica(Tessella)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	Varies
Roda	x	x	x		x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	Free

Fuente http://digitalpowrr.niu.edu/tool-grid/

Hay demasiado más por ver y elegir juntos!

RODA community
REPOSITORY OF AUTHENTIC DIGITAL OBJECTS

Xena
Digital Preservation Software

JHOVE2

archivematica

Plato

planets

Preservica
Digital Preservation

fits

premis
solutions

SCAPE

Scout

DROID

sedici.unlp.edu.ar

Esta presentación estará disponible en la colección de SEDICI: http://sedici.unlp.edu.ar/handle/10915/25293

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

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