VII CONFERENCIA INTERNACIONAL SOBRE BIBLIOTECAS Y REPOSITORIOS DIGITALES DE AMÉRICA LATINA
BIREDIAL-ISTEC `17
2·3·4 de octubre 2017 LA PLATA · ARGENTINA

EDUCACIÓN PÚBLICA Y GRATUITA

UNIVERSIDAD NACIONAL DE LA PLATA

istec
Ibero American Science & Technology Education Consortium

Revisión de distintas implementaciones para preservación digital: hacia una propuesta metodológica para la preservación y la auditoría de confiabilidad de RI

De Giusti, Marisa R., Villarreal, Gonzalo L.
PREBI-SEDICI Universidad Nacional de La Plata
CESGI Comisión de Investigaciones Científicas

CIC COMISIÓN DE INVESTIGACIONES CIENTÍFICAS
Ministerio de Producción, Ciencia y Tecnología

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Universidad del Rosario

UNIVERSIDAD DE COSTA RICA

UNIVERSIDAD DEL NORTE

UASLP

Objetivos del trabajo

- Relatar una investigación sobre estructuras aptas para la preservación de documentos digitales en repositorio.
- Describir antecedentes similares y en particular tres experiencias exitosas de conexión de un repositorio con herramientas capaces de asegurar la preservación digital.
- Todos los modelos tienen subyacente el Modelo OAIS ISO 14721[1].

[1] International Organization for Standardization (ISO). (2012). ISO 14721. Space data and information transfer systems - Open archival information system (OAIS) - Reference model. Disponible en https://www.iso.org/standard/57284.html

Noción de preservación digital

La preservación digital consiste en los procesos destinados a garantizar la accesibilidad permanente de los objetos digitales. Para ello, es necesario encontrar las maneras de representar lo que se había presentado originalmente a los usuarios mediante un conjunto de equipos y programas informáticos que permiten procesar los datos.

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.

Estrategias en la preservación digital

- La preservación digital supone la selección y puesta en práctica de un conjunto evolutivo de estrategias con objeto de lograr el tipo de accesibilidad anteriormente mencionado, considerando las necesidades de preservación de las diferentes capas de los objetos digitales.
- También están los aspectos legales de permiso para poder preservar.

Estrategias en la preservación digital

- Colaborar con los productores (creadores y distribuidores) para aplicar normas que prolonguen la vida efectiva de los medios de acceso y reduzcan la variedad de problemas desconocidos que deben ser tratados
- Reconocer que no es realista tratar de preservar todo, hay que seleccionar
- Guardar el material en un lugar seguro
- Controlar el material utilizando metadatos estructurados y otros documentos que faciliten el acceso y ayuden durante todo el proceso de preservación
- Proteger la integridad y la identidad de los datos
- Elegir los medios apropiados para proporcionar acceso pese a los cambios tecnológicos

Pretensiones de este trabajo

Mostrar antecedentes y el inicio de un trabajo de investigación con el motivo de generar preguntas sobre un tema incipiente en AL
cuando todavía no hay un consenso sobre qué estructura utilizar para preservar en los repositorios digitales de AA

Adelantando las conclusiones

- Luego de un análisis previo de las estructuras disponibles para preservación digital, al tener las distintas implementaciones facilidades básicas comunes, tales como agregado de metadatos especiales para preservación y seguimiento del ciclo de vida del OD, la elección final atiende a dar una respuesta para la implementación más común de RI de la región latinoamericana que está basada en la herramienta de código abierto DSpace.

Antecedentes
estructuras e implementaciones exitosas

1. Estructura utilizada en el proyecto SCAPE

- **Sca**lable **P**reservation **E**nvironments (SCAPE)[1] fue un proyecto coordinado por el Austrian Institute of Technology, financiado por la Unión Europea. El proyecto comenzó en 2012 y finalizó, según lo previsto, a fines de 2015.

Los puntos centrales del proyecto fueron:

- ❏ El análisis de formatos de archivos de repositorios
- ❏ La descripción formal de planes y políticas de preservación
- ❏ La automatización y puesta a punto de herramientas y procesos escalables

El control de calidad de procesos de preservación

Cumple con la mayor parte de los requerimientos de la norma ISO 16363[2]

[1] Sitio web: http://scape-project.eu/

[2] International Organization for Standardization (ISO). (2012). ISO 16363. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. Disponible en https://www.iso.org/standard/56510.html

Estructura utilizada en el proyecto SCAPE para el ciclo de vida de la preservación

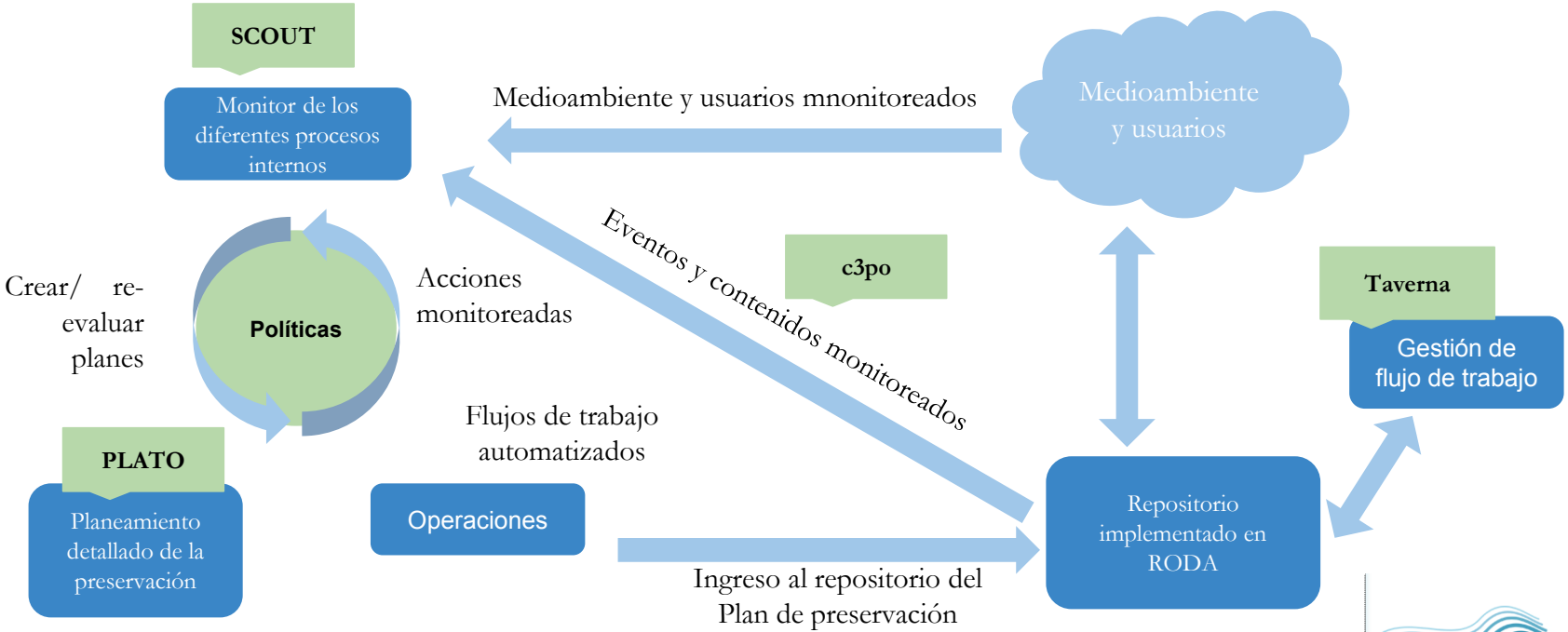


Gráfico basado en: 4. M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria, "The scape planning and watch suite," in Proc. 10th Int. Conf. Preservation Digital Objects, Lisbon, Portugal, 2013, pp. 262-265.

RODA fue desarrollado para ser un repositorio digital completo, y provee la funcionalidad necesaria para las principales unidades que componen el modelo de referencia OAIS.

RODA implementa todo el flujo de trabajo de ingesta, en el que no sólo valida los SIP sino también se encarga del proceso de negociación entre el archivo y el productor. Para el proceso de acceso (*access*) provee diferentes posibilidades de búsqueda y navegación a través de los metadatos, además de visualizaciones y descargas de los OD almacenados. Los componentes de la administración (*Administration*) también fueron desarrollados para permitir a los archivistas modificar los metadatos descriptivos y definir reglas para intervenciones de preservación, como el control de integridad sobre todos los OD almacenados, la iniciación de un proceso de migración o el control de usuarios.

RODA tiene un modelo de contenido atomístico y muy orientado a PREMIS. Cada entidad intelectual es descrita por un componente EAD (Encoded Archival Description) de registro de metadatos. Estos registros se organizan jerárquicamente a fin de constituir una descripción de archivo completa, pero manteniéndolos separados dentro del modelo de contenidos. Estos componentes EAD son creados a través del mecanismo de enlaces RDF de Fedora, y cada nodo, hoja del árbol jerárquico es enlazado a un objeto de representación (ejemplo: un objeto Fedora que incluye todos los archivos y *bitstreams* que componen la representación digital). Se mantienen relaciones lógicas entre todos estos objetos, por medio de un conjunto de entidades PREMIS (nodos PO), a fin de conocer la historia y origen (*provenance*) de cada objeto.

RODA Los eventos de preservación que se llevan a cabo se registran como nuevos nodos de eventos de preservación. Algunos eventos especiales, como migraciones de formatos, establecen relaciones adicionales entre dos nodos de representación de preservación (eventos de enlazado). Cada evento de preservación es ejecutado por un agente, que puede ser un usuario del sistema o un evento disparado automáticamente por el software.

Como es de esperarse ¡y muy deseable! la información del agente que disparó el evento también se registra dentro del nodo PO.

Estructura utilizada por British Columbia University

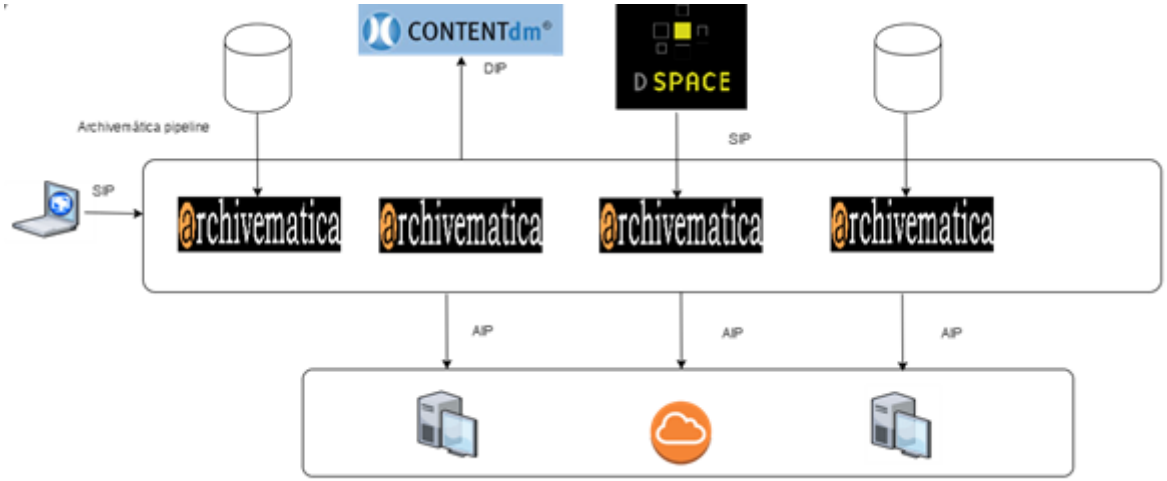


Diagrama general de un pipeline de Archivematica Fuente: Sprout, Bronwen and Romkey, Sarah. (2016). “Implementation: Building a preservation strategy around Archivematica”. En Bantin, Philip C. (ed.). *Building Trustworthy Digital Repositories: Theory and Implementation*. Rowman & Littlefield: London.

Experiencia UBC

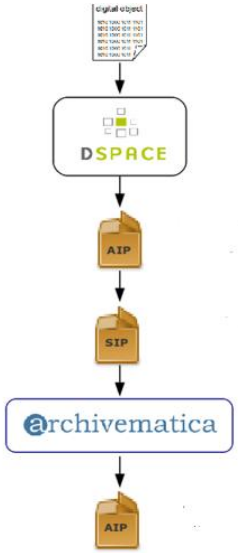
- Implementación del repositorio CIRCLE en DSPACE.
- Análisis de las prácticas de preservación de materiales digitales y digitalizados.
- Distintos mecanismos de ingreso de un SIP (ingreso al pipeline).
- Uso de Content dm para el almacenamiento y gestión de los archivos digitales.
- Cada instancia de Archivematica dentro del pipeline se encarga del procesamiento de contenido.

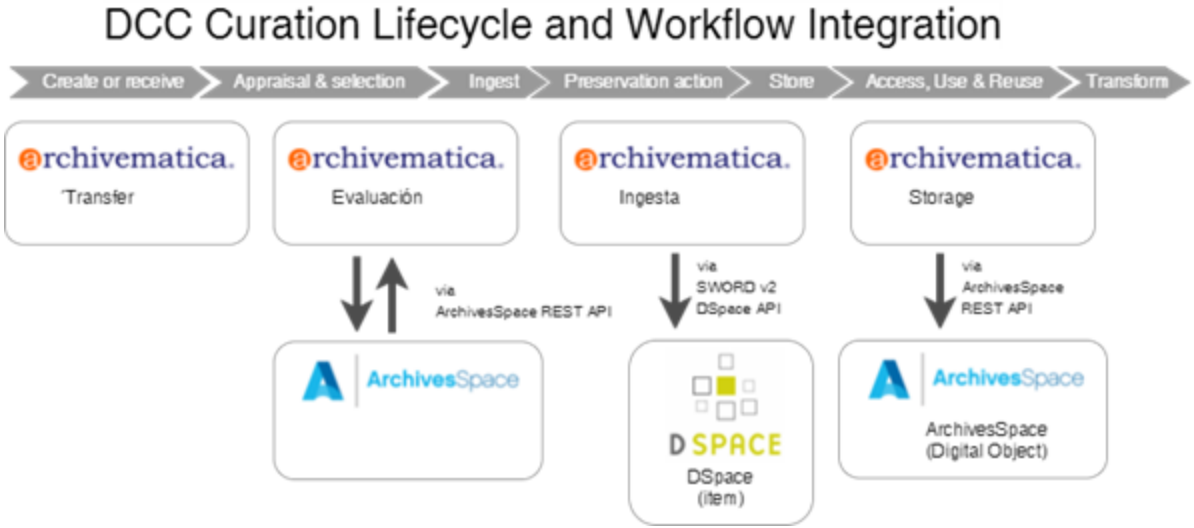
cIRcle http://circle.ubc.ca/

- DSPACE es la herramienta de depósito de los SIPs y de entrega de los DIPs.

Estructura utilizada por la la Biblioteca Histórica de Bentley (Universidad de Michigan)

Ítem-AIP-SIP-AIP





Estructura utilizada por la Biblioteca Histórica de Bentley. Fuente: Eckard, Max; Pillen, Dallas and Schallcross, Mike. (2017). "Bridging Technologies to Efficiently Arrange and Describe Digital Archives: the Bentley Historical Library's archivesSpace-Archivematica-DSpace Workflow Integration Project". *Code(4)Lib Journal*, 35, 1-30. Disponible en http://journal.code4lib.org/articles/12105

Experiencia Michigan

Tras la revisión de herramientas open-source para software de gestión de RI la biblioteca Bentley decidió integrar las funcionalidades de ArchivesSpace, Archivematica y DSpace en una estructura con funciones distribuidas.

- ❏ Facilitar la creación/reutilización de metadatos descriptivos y administrativos en los sistemas de conservación y gestión.
- ❏ Simplificar la ingesta y el depósito de contenido en un repositorio de preservación.
- ❏ Encontrar soluciones para Bentley, pero extrapolables a otras instituciones. Compartir todo el código y la documentación con los archivos y las comunidades de preservación digital.

Experiencia UNLP

- Se ha instalado Archivematica en un servidor GNU/Linux. La instalación incluye el storage service.
- Se han analizando sus funcionalidades básicas logradas a partir de las herramientas libres y de código abierto que incluye. Estas permiten procesar el OD desde la ingesta.
- Se ha probado la ingesta de directorios de archivos y disparado los distintos microservicios.
- Se ha instalado Archives Space.

Actividades que realiza la estructura propuesta

- El repositorio en DSpace está encargado del ingreso y la entrega de los contenidos digitales.
- La estructura de Archivematica realiza las actividades de preservación digital a través de la implementación de un conjunto de microservicios, que actúan sobre una estructura conceptual asimilable al paquete de información (IP) en sus distintas versiones.
- La estructura física resultante del paquete de información en sus diferentes versiones (SIP, AIP, DIP) incluye archivos, checksum, logs, documentación de la transferencia y metadatos en una estructura XML.

Archivematica

Su estructura se basa en los microservicios y las foss tools

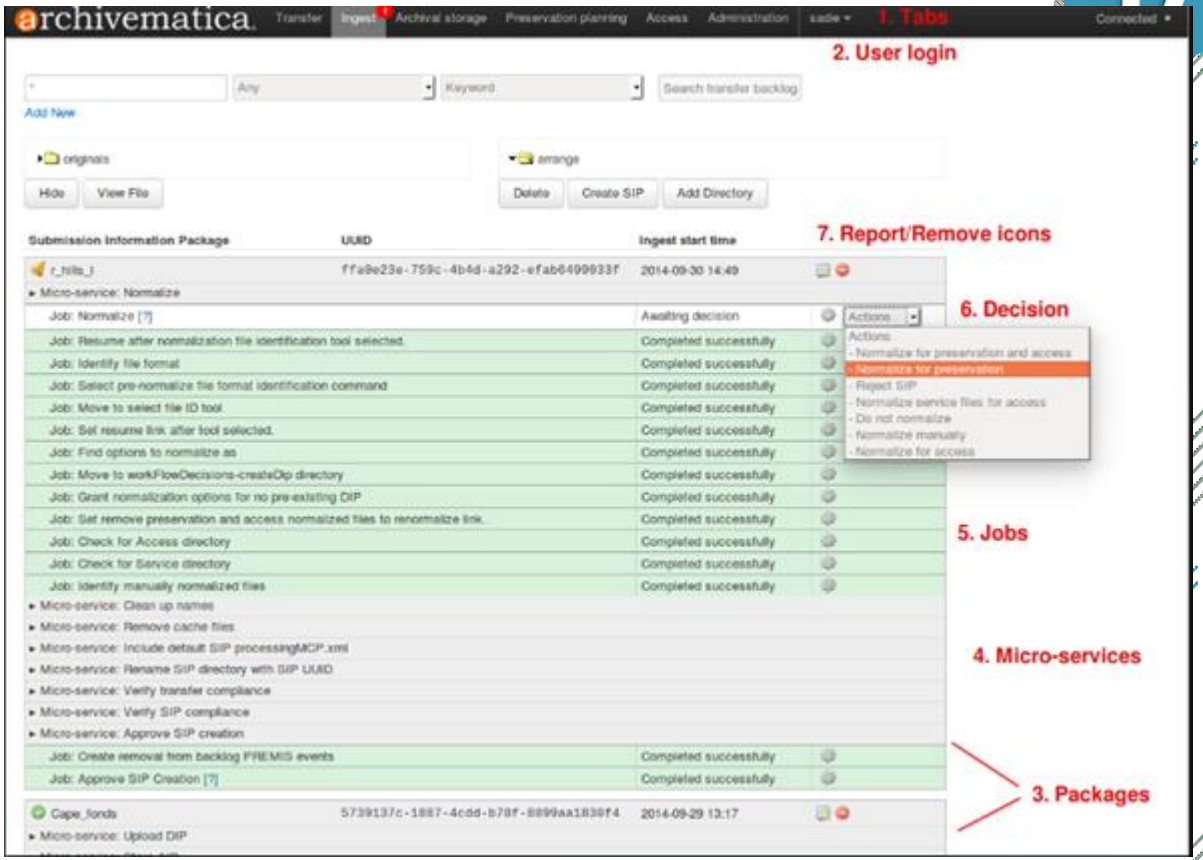
Archivematica

Usa los esquemas PREMIS, METS y DC.

El Escritorio de Archivematica permite seguir las acciones que suceden en los distintos procesos y microservicios.

El flujo de trabajo comienza tras una transferencia de archivos (pueden venir de un DSPACE).

Usa muchas herramientas: Bagit, Fido, Siegfrid, Jhove, Fits...



Escritorio de Archivematica. Fuente: https://www.archivematica.org/es/docs/archivematica-1.5/_images/Dashboard.png

Funciones de Archival Storage

- Archivematica utiliza una estructura de árbol de directorios para almacenar los AIP localmente.
- La estructura en árbol está basada en los identificadores persistentes (16 bits) del AIP; también permite múltiples sitios de almacenamiento locales o remotos e incluso localizaciones LOCKSS.
- El detalle del procedimiento de integración se encuentra en la wiki de Archivematica.

Archives Space: descripción y funciones

ArchivesSpace es un software abierto de gestión de archivos que permite dar seguimiento a las sesiones de acceso y la gestión de colecciones. Permite:

- La incorporación de nuevos registros
- La publicación de materiales
- La gestión de autoridades
- La gestión de lugares
- La gestión de derechos
- Brinda servicio de referencia
- Ofrece generación de informes y reportes
- Genera metadatos EAD, MARCXML, MODS, Dublin Core, y METS
- Brinda la posibilidad de exportaciones

Los procesos que se han ejecutado en esta fase de pruebas son:

- ❑ Aprobar la entrega
- ❑ Comprobar el cumplimiento de requisitos
- ❑ Renombrar los archivos añadiendo un identificador único
- ❑ Crear sumas de verificación y comprobarlas
- ❑ Crear archivos METS XML
- ❑ Colocar los archivos en cuarentena si es necesario
- ❑ Identificar formatos de los archivos
- ❑ Extraer archivos empaquetados
- ❑ Comprobar la existencia de virus
- ❑ Mover los archivos al directorio de entregas finalizadas
- ❑ Crear el SIP[1]

[1] Como alternativa pueden enviarse los archivos al *backlog* para su posterior procesamiento e incluso crear reportes, como puede verse en la documentación de Archivematica, disponible en: https://www.archivematica.org/es/docs/archivematica-1.5/user-manual/transfer/manage-backlog/#retrieve-from-backlog

Conclusiones y trabajos futuros

- Sólo se tienen pruebas aisladas sobre Archivematica y un gran trabajo pendiente.
- Es necesario realizar la conexión con ArchivesSpace
- Es necesario crear la conexión con un DSPACE de pruebas
- ¡Hay mucho por hacer!

¡Muchas gracias por su atención!

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