Preservación y difusión del Fondo Histórico de la Universidad de la Sabana

Gladys Adriana Bello García, Héctor Martínez Torres, Sebastián Gómez Lozano, Emilio Lorenzo Gil

Palabras clave
Repositorios institucionales, Fondo Antiguo, preservación digital, DSpace, institutional repositories, digital preservation

Resumen

En el contexto de las instituciones de educación superior es necesario que los repositorios amplíen sus objetivos y capacidades de gestión y preservación más allá de las publicaciones científicas y académicas hacia la conservación de los Fondos Antiguos e Históricos.

Por lo anterior, se presentan los trabajos realizados en el repositorio *Intellectum* de la Universidad de la Sabana derivados de la incorporación de un Fondo Histórico de especial importancia recibido el año 2005 en donación constituido por dos colecciones con aproximadamente 40.000 folios de los siglos XVIII, XIX y XX. Esta donación corresponde a los fondos documentales Manuel María Mosquera y David Mejía Velilla conformando el Archivo Histórico Cipriano Rodríguez Santa María.

En *Intellectum* se requería mantener a la vez la máxima calidad y la máxima difusión (acceso) de las copias de preservación, copias de consulta y representaciones adicionales para diferentes usuarios y dispositivos de acceso. Para ello se ha realizado un enfoque en el que coexisten las soluciones específicas de preservación de los fondos documentales antedichos con las igualmente específicas de difusión y visualización, de forma amplia, para una variedad de usuarios.

Una parte del sistema de ingesta de *Intellectum* se ha orientado al depósito de las digitalizaciones en formato jpeg de alta calidad resultantes del proceso de digitalización de los fondos históricos. Estas digitalizaciones son los objetos principales de los sistemas de preservación que incorpora DSpace. Por su gran tamaño, derivado de su alta definición, estos ficheros no se ponen a disposición de los usuarios generales, almacenándose en un *bundle* específico.

Posteriormente, un conjunto de tareas automáticas de curación ensamblan a partir de esas copias de preservación diversas versiones o representaciones de cada objeto. En *Intellectum* se ensamblan los ficheros de preservación en un único fichero en formato pdf multipágina apto para su descarga por todo tipo de usuarios (almacenado en un *bundle* accesible), como representación en formato flash apta para un visor *pseudo-streaming* (específico de *Intellectum*) o como representación apta para dispositivos que soporten HTML4.

Abstract

Higher education institutions are asked to expand repositories and preservation management capabilities beyond scientific and academic publication management, such as History Collections.

This communication presents the work done in the repository Intellectum of La Sabana University, when in 2005 incorporated a History Collection of exceptional importance, receiving two collections of approximately 40,000 pages of the eighteenth, nineteenth and twentieth centuries, corresponding to the documentary fund of Manuel Maria Mosquera and David Mejia Velilla. These collections were integrated in the Historical Archive Cipriano Rodriguez Santa Maria.

Intellectum adaptation was focused on balancing the highest quality and maximum exposure (access) of preservation copies, consultation copies and additional representations for different users and access devices. For this, it has been made an approach in which specific preservation solutions coexist with, equally specific, visualization solutions, for a variety of users and access devices.

A part of the ingest system in Intellectum is oriented towards the deposit of high quality jpeg copies resulting from the process of digitization of historical collections. These scanned copies are the main objects of the preservation systems incorporated to DSpace. For its size, derived from its high definition, these files are not available to general users, and are stored in a specific DSpace bundle.

Subsequently a set of automated curation tasks assemble, from those copies, different versions and representations of each object. In Intellectum the preservation files are assembled into a single multipage pdf file, suitable for download by all users (stored in an accesible bundle), and also stored in flash format suitable for a pseudo-streaming viewer, or a copy apt for devices that support HTML4.

Contexto

El Repositorio Institucional *Intellectum* fue creado por la Universidad de la Sabana, Chía, Colombia, en el año 2012 ”para la publicación y visibilidad de la producción académica e investigativa de la comunidad universitaria.” Igualmente se consideró pertinente “acoger el sistema *Open Access* como lineamiento institucional para la publicación de la producción intelectual de la Comunidad universitaria”.

El Instituto de Humanidades de la Universidad de la Sabana recibió en el año 2005 la donación de los fondos documentales Manuel María Mosquera y David Mejía Velilla, dos colecciones con aproximadamente 40.000 documentos de los siglos XVIII, XIX y XX, muchos de ellos inéditos. En virtud del principio según el cual los archivos históricos conserven la memoria colectiva, esencial para la identidad de la Nación Colombiana, resultaba imperiosa su organización, clasificación y descripción, dentro de un cuerpo integrado en el archivo histórico Cipriano Rodríguez Santamaría como medio facilitador de consulta para investigadores.

La Universidad de La Sabana concedió un espacio en el quinto piso de la Biblioteca Octavio Arizmendi Posada que fue adecuado con algunas condiciones técnicas y ambientales mínimas para permitir el tratamiento archivístico correspondiente y el depósito de los documentos. Además, puesto que entre las políticas aprobadas para el proyecto del repositorio institucional estaba contemplado incorporar colecciones históricas al mismo, se decidió incluir estos fondos en el repositorio *Intellectum*.

El proyecto se dividió en varias fases: identificación, clasificación, descripción, elaboración de herramientas de consulta y control de calidad que se encuentran ya ejecutadas.

- Identificación: se realiza un diagnóstico general del estado físico y de conservación de los documentos, se realiza un inventario sobre los contenidos de manera general.
- Clasificación, ordenación, foliación y almacenamiento: se realizó un cuadro de clasificación de los fondos a partir del cual se realizó la organización física de los documentos.
- Descripción de los documentos: se realizó una base de datos en Excel con la descripción por unidad documental de acuerdo con la norma ISAD G (Comité de Normas de Descripción, 2000).
- Elaboración de herramientas de consulta: se realizó el catálogo de los fondos, fichas archivísticas e índices.
- Control de calidad: se hizo la revisión de catálogos, fichas archivísticas, índices y base de datos.
- Resultado de una investigación basada en documentos del Archivo Histórico, se editó y lanzó el libro “Rasgos poéticos que pueden servir de apuntamientos sobre la historia de nuestra revolución, escritos por Mariano del Campo Larraondo”; presentación, transcripción y notas de Marcela Revollo Rueda.
- En marzo de 2013 se realizó una exposición denominada “De audiencias, cantones y moradas: Estampas de la Independencia” y se editó el respectivo catálogo con el fin de hacer el lanzamiento oficial del Archivo Histórico Cipriano Rodríguez Santa María.
- Mediante acto administrativo se creó el Archivo Histórico Cipriano Rodríguez Santa María, con fecha retroactiva a la del evento anterior (4 de marzo de 2013), adhiriéndose el archivo como unidad a la Biblioteca Octavio Arizmendi Posada.
- Se definió el procedimiento a seguir para disponer los folios y su descripción archivística en una estructura de metadatación adaptada para el Repositorio Institucional *Intellectum* en formato Dublin Core. Adicionalmente se realizó la parametrización de un visor de acercamiento progresivo que le permite al usuario final tener una mejor experiencia de visualización y aprovechamiento sobre los documentos a consultar.

- Se ejecutó el proceso de digitalización del acervo documental con un tratamiento de los dos fondos documentales: Fondo David Mejía Velilla con 1679 registros y Fondo Manuel María Mosquera con 1818 registros.
- Se está llevando a cabo el proceso de catalogación del fondo consistente en la creación de los manifestos XML correspondientes a la metadatación de cada elemento, siguiendo indicaciones de la profesora Marcela Revollo, Historiadora investigadora de la Universidad.
- Creación de paquetes de ingesta específicamente diseñados para optimizar los flujos de trabajo de los agentes del proceso, con control estricto del ordenamiento numérico de los archivos, necesario para la creación correcta de los PDFs multipágina a partir de las imágenes individuales.
- Ingesta al repositorio *Intellectum* de forma progresiva y alineada con los trabajos de digitalización y metadatación.
- Microfilmación de la totalidad del fondo con el fin de salvaguardar una copia adicional de respaldo en este formato.
- Sellado de los documentos, así como el diseño del sello por parte de Comunicación Institucional y Gestión de la Información.
- Contextualización y apoyo en el apalancamiento del análisis, ordenación e investigación, realizando biografías y árboles genealógicos.
- Revisión general del documento Políticas y Lineamientos del Archivo Histórico Cipriano Rodríguez Santa María, con acercamientos con el Archivo General de la Nación para realizar el registro del archivo y generar convenios para la restauración de los documentos.
- Adecuación física del espacio destinado al Archivo Histórico con el fin de realizar ampliaciones para la disposición adecuada de estanterías, máquinas, espacio de consulta y oficinas.

Intellectum y el Fondo Histórico

La incorporación del Archivo Histórico Cipriano Rodríguez Santa María en *Intellectum* plantea una serie de problemas e interrogantes como consecuencia del tipo de material digital a incorporar: unas características digitales muy específicas derivadas principalmente de los propios fondos y el proceso de obtención de la copia digital.

La copia digital de preservación, producto de una primera fase de ese proceso de digitalización, se traduce en archivos de mucha definición, calidad y por lo tanto, de enorme tamaño para ser gestionados correctamente desde el punto de vista de su accesibilidad. Por este motivo las copias de consulta han sacrificado la calidad de los archivos en favor de su acceso.

Recientemente, otra derivada se ha añadido a esta problemática, a saber, la "movilización" de contenidos y los nuevos requerimientos a sus futuras representaciones

(e.g., la adaptación a su visualización por dispositivos móviles, la obsolescencia de determinados formatos, etc.).

Así, los interrogantes que surgieron en el planteamiento de *Intellectum* como depósito de las copias digitales de los fondos documentales Manuel María Mosquera y David Mejía Velilla fueron, ¿cómo almacenar y preservar las copias digitales?¿Qué nivel de definición requerían dichas copias? ¿Qué formatos son los más adecuados para la preservación dentro de un repositorio?¿Sirven esos formatos para la difusión al máximo de audiencias y dispositivos de consulta? ¿Existen experiencias previas? ¿Como difundir mejor, ahora y en el futuro, los objetos digitales?

En ese escenario el dilema que había que resolver era, ¿preservación y difusión son incompatibles? Creemos que no necesariamente.

Almacenamiento DSpace

La arquitectura funcional de DSpace sigue un modelo de tres capas: Aplicación, Lógica de negocio y Almacenamiento. Esta última incorpora el acceso a la Base de Datos y al denominado *Assetstore*. La Base de datos almacena las estructuras correspondientes a los metadatos, usuarios, workflows, permisos, estructura comunidades/colecciones, etc. mientras que el *Assetstore* almacena los ficheros primarios del repositorio (pdf, txt, docx,...), ficheros de licencia, extracciones de texto completo para la indexación, manifiestos xml adicionales, etc.

El enfoque seguido en *Intellectum* es almacenar todos los ficheros con algún valor de preservación en el *Assetstore*, con el fin de garantizar la coherencia y completitud de las funciones de Planificación de la Preservación, *Preservation Planning* en terminología OAIS (Consultative Committee for Space Data Systems, 2012) que proporciona los servicios de control del entorno del archivo y provee de capacidades de preservación, para asegurar que la información almacenada permanezca accesible a largo plazo, aún cuando el entorno original devenga obsoleto. De esta manera, la aplicación de las políticas de preservación, gestión de ciclos de vida, transformaciones, extracciones de objetos, e igualmente importante, migraciones de software, serán coherentes y comprensibles.

Por otra parte, las representaciones de los objetos de carácter eventual, sin valor de preservación o variantes transitorias, como pueden ser versiones para el visionado de un objeto en dispositivos móviles, ensamblajes de ficheros para visores específicos, etc. podrían y deberían ser mantenidas fuera del *Assetstore*. En *Intellectum*, en su versión actual, se dispone de dos representaciones adicionales a la correspondiente al PDF, una en formato flash y otra en formato png.

Estructura lógica de almacenamiento DSpace

Para el almacenamiento de los contenidos del Fondo Histórico planteamos un enfoque multinivel de conservación, preservación y difusión.

La capacidad de la estructura de almacenamiento en DSpace de almacenar ficheros en distintas agrupaciones lógicas, los denominados *bundles*, se ha usado en *Intellectum* para incluir en el repositorio las distintas versiones del mismo contenido digital, adaptando o personalizando cómo esos objetos son servidos a los distintos usuarios (distintas necesidades y requisitos).

A la vez, el sistema de gestión de objetos digitales de DSpace (que se puede extender con capacidades de análisis pormenorizado de formatos, transformación y migración) se usa como base del modelo de preservación del patrimonio digital del Archivo Histórico Cipriano Rodríguez Santa María.

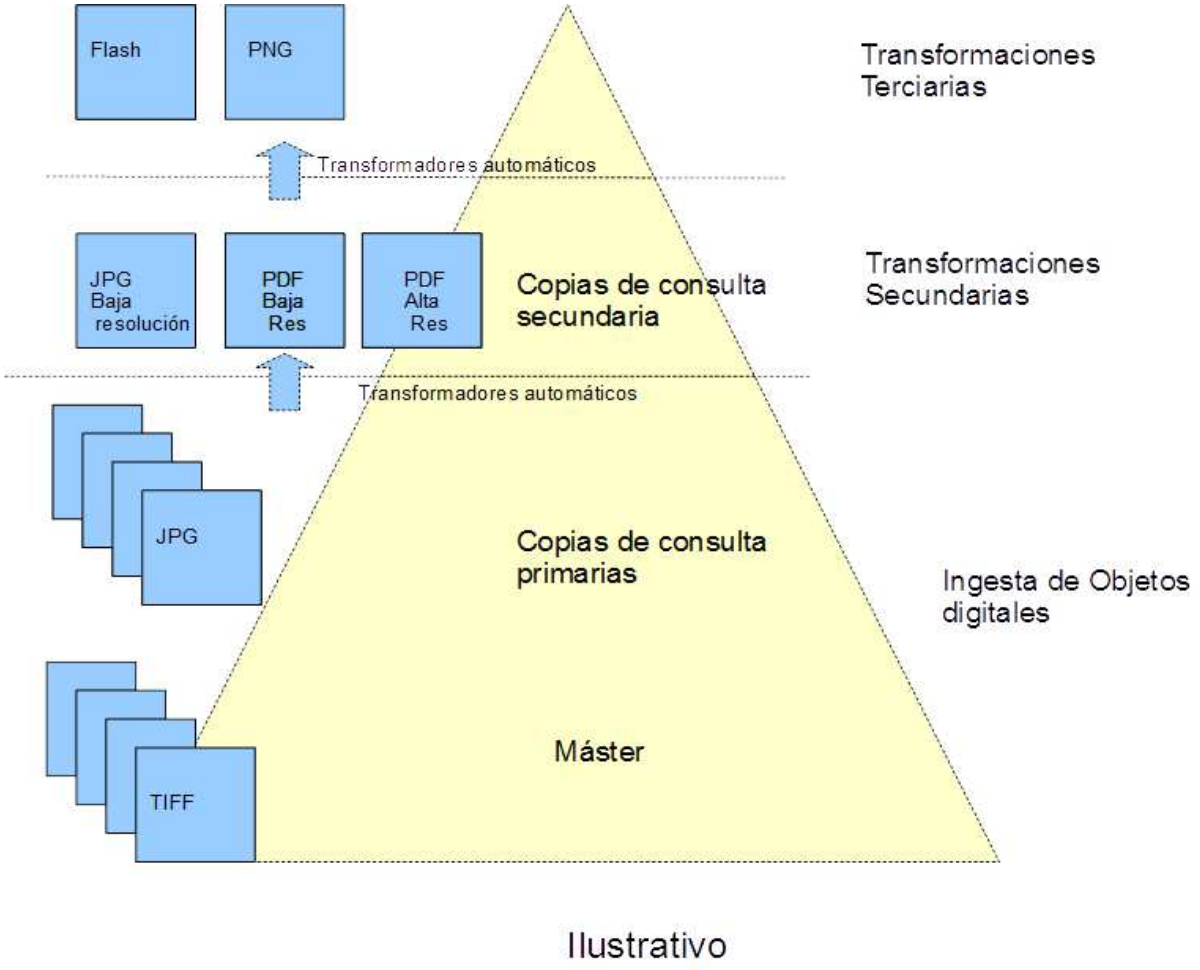
En *Intellectum* se ha utilizado la estructura de *bundles* que se presenta en la tabla adjunta.

Bundle	**Fichero**
Thumbnail	Miniaturas
Original	PDF multipágina, copias de consulta secundaria
TXT	Extracciones a texto completo de PDF u otro material de texto extraíble
License y cc_license	Ficheros de licencia de depósito y licencias de reutilización de contenidos creative-commons
Bundles Extendidos	
Preservacion	JPGs de consulta primaria
Master	TIFFs originales máster o TIFF editados

Adicionalmente al *Assetstore*, se ha creado un espacio de almacenamiento adicional para el depósito y gestión de las representaciones digitales que los visores específicos requieren para mostrar los contenidos del Fondo Histórico. Es un espacio de almacenamiento, desechable, donde no se requiere un enfoque de preservación y que almacenan los ficheros correspondientes a las transformaciones terciarias del contenido del Fondo Histórico.

Otra posibilidad que surge con el enfoque adoptado es desechar representaciones que devengan obsoletas, ya que las copias maestras, las realmente valiosas a efectos de preservación, seguirán existiendo en DSpace. Los tratamientos son múltiples y están imbricados precisamente en el enfoque multinivel de almacenamiento que se ha seguido.

De un modo gráfico, podríamos representar el modelo anterior de la siguiente forma:



Modificaciones al software DSpace

Los planteamientos anteriores se han traducido en las siguientes funcionalidades y modificaciones incorporadas a *Intellectum* (DSpace v4 con interfaz XMLUI):

1. Extensión del espacio lógico de Dspace mediante la definición de dos nuevos *bundles,* Máster y Preservación.
2. Adaptación de los sistemas de ingesta de ítems y de edición de ítems por interfaz de usuario para que los máster y copias primarias puedan almacenarse en los *bundles* anteriores. En la actualidad no se realiza ningún proceso automatizado que gestione la correspondencia entre máster TIFF y copias primarias, siendo la responsabilidad de los administradores del repositorio el adecuado ensamble de los paquetes SIP que se ingestarán en DSpace.
3. Tarea de curación para la creación de pdf multipágina. Una tarea específica de curación *JpgsToPdfCurationTaskCompresion* se programa para que comprima las

imágenes del *bundle* PRESERVACION a un nivel en el que resultan legibles y genera a partir de ellas un fichero PDF multipágina con la composición de esas imágenes. Para la correcta generación de los PDFs se requiere una estricta nomenclatura de los JPGs, con el fin de recrear la paginación del original. Este PDF multipágina se coloca en el *bundle* ORIGINAL y pasa a ser accesible por el público (si los permisos del ítem lo permiten).

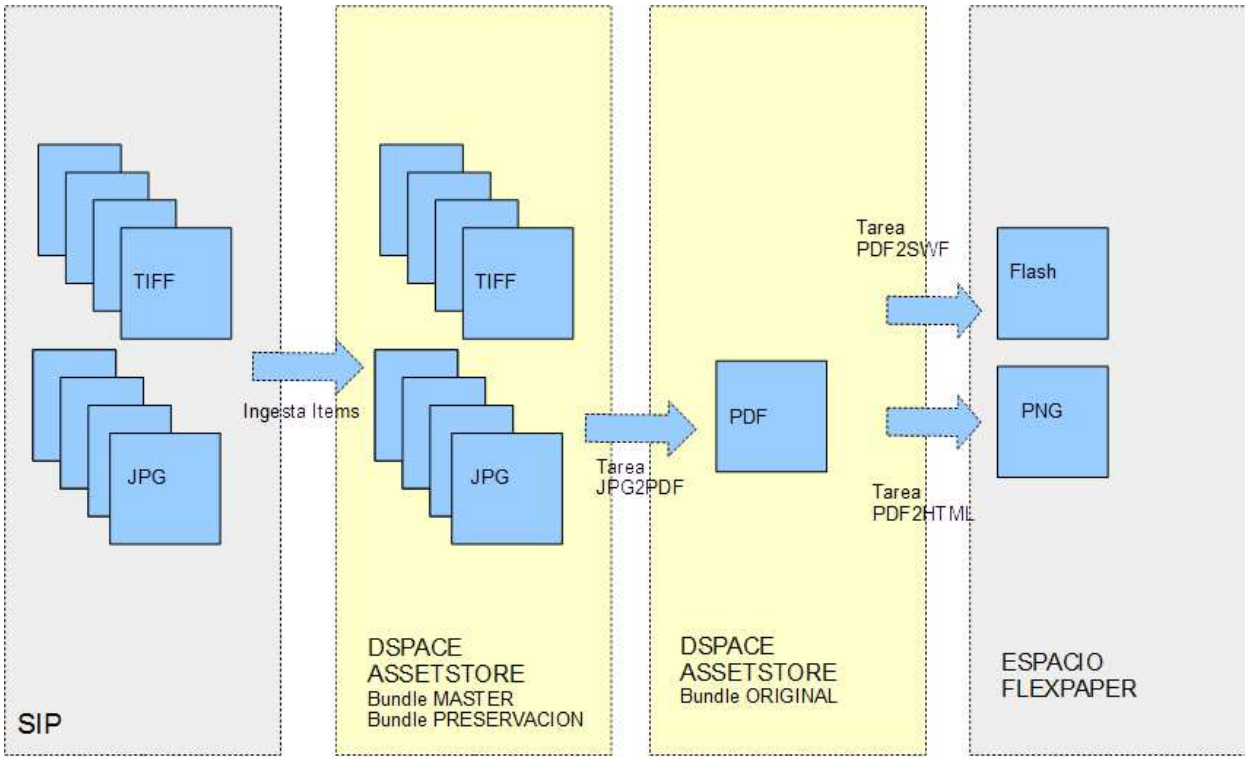
4. Tarea de curación para generación flash y png a partir de los pdf multipágina. La tarea de curación *FlexpaperGenerationCurationTask* procesa el pdf de un ítem y genera los ficheros necesarios para su visualización por el visor especializado. Utiliza los programas externos *pdf2swf* y *mudraw* para ello y deben estar instalados en la máquina. En ese proceso también se genera un fichero JSON, utilizando el programa *pdftojson*, que es utilizado para realizar búsquedas en el documento completo en el visor. Señalar que no se realiza un OCR de forma estricta, sino una extracción a partir del pdf.

5. Modificación de la vista simple de ítem para integrar el visor, capaz de ajustar su visualización a formato HTML si el dispositivo no soporta formato flash. Señalar que *Intellectum* está adaptado a la visualización por dispositivos móviles (incorpora la interfaz adaptativa del tema Mirage2) por lo que es importante la detección y correspondiente ajuste a las características de los nuevos dispositivos de navegación.

6. Incorporación de un visor específico, basado en el producto comercial Flexpaper (Flexpaper), solución que integra un visualizador rico de documentos, específicamente adaptado a material digital de gran calidad, proporcionando una experiencia vívida y realística de la lectura de los documentos digitales de *Intellectum.*[1]

Apuntaríamos que las funcionalidades número 3 y 4, al ser intensivas en consumo de recursos, están programadas para su ejecución nocturna, por lo que la aparición de las copias de consulta y resto de representaciones para la visualización puede presentar retrasos. No obstante, si en el período entre la ingesta del objeto al repositorio y la generación (nocturna) de estas representaciones adicionales algún usuario visitase el ítem, se ejecutan automáticamente los procesos de generación anteriormente apuntados y el usuario podría visualizar el contenido solicitado.

1 Ver, p.ej. http://intellectum.unisabana.edu.co/flexpaper/handle/10818/18424/REF%20566-71R.pdf?sequence=10&isAllowed=y



Conclusiones

El tratamiento integral de los aspectos de preservación y difusión de los ficheros correspondientes a un Fondo Histórico, requiere de enfoques novedosos y complementarios en los repositorios digitales tradicionales. En el repositorio *Intellectum* se ha logrado conjugar a la vez la máxima calidad y la máxima difusión (acceso) de las copias de preservación, copias de consulta y representaciones adicionales para diferentes usuarios y dispositivos de acceso. Para ello se ha diseñado y ejecutado un enfoque en el que coexisten las soluciones específicas de preservación de los fondos documentales con las igualmente específicas de difusión y visualización, de forma amplia, para una variedad de usuarios y dispositivos de acceso.

La ventaja del enfoque adoptado reside en generar *en tiempo* nuevas copias o representaciones de cada objeto digital sin necesidad de realizar una re-ingesta de los ficheros. De este modo, la adaptabilidad y flexibilidad del sistema son máximas al facilitar el acceso y representación de los objetos digitales con un mínimo esfuerzo, sea cual sea la evolución tecnológica o la opción de acceso de los usuarios.

Los elementos funcionales y tecnológicos que se han incorporado en el repositorio comprenden una variedad de soluciones: adaptación de los *bundles* lógicos de DSpace, generación automática de copias y representaciones por medio de tareas de curación, implantación de conversores de formato e implantación de visores específicos. El resultado es un repositorio que conjuga la especificidad de la preservación digital con la adaptación extremadamente flexible a la visualización multifomato y multidispositivo.

BIBLIOGRAFÍA

The Consultative Committee For Space Data Systems. (2012). Recommendation For Space Data System Practices: Reference Model For An Open Archival Information System (OAIS). CCSDS, Washington . Recuperado a partir de http://public.ccsds.org/publications/archive/650x0m2.pdf

Comité de Normas de Descripción. (2000). ISAD(G): Norma Internacional General de Descripción Archivística : Adoptada Por el Comité de Normas de Descripción. Estocolmo, Suecia, 19-22 Septiembre 1999. Consejo Internacional de Archivos, Madrid. Recuperado a partir de http://www.ica.org/download.php?id=1745

Devaldi Ltd. FlexPaper - The web pdf viewer solution. Recuperado 15 de septiembre de 2015, a partir de http://flexpaper.devaldi.com/

