VI Conferencia Internacional BIREDIAL/ISTEC - 2016
Benemérita Universidad Autónoma de San Luis Potosí (UASLP), México
http://congresos.unlp.edu.ar/biredial-istec

Un modelo de trabajo para agilizar la generación de documentos de texto para su preservación

Salamone Lacunza, Paula [1]; Villarreal, Gonzalo L. [2]; De Giusti, Marisa R.[3]; Lira, Ariel J[4].

Resumen

Los repositorios institucionales (RI) tienen la responsabilidad de gestionar, preservar y ofrecer acceso libre a la producción científica de una institución particular. Para ello, el repositorio debe establecer políticas que aseguren la autenticidad de los objetos digitales, que prevengan la pérdida parcial o total de los mismos, y que permitan acceder a su contenido por una comunidad de usuarios designada. Para la correcta ejecución de estas políticas se deben realizar un conjunto de actividades de preservación que, idealmente, deberán integrarse al conjunto de tareas de administración del repositorio cotidianas, y así estandarizar y asegurar la realización de las actividades de preservación que se hacen sobre los objetos digitales. Desde luego, también será necesario realizar revisiones periódicas sobre los métodos y circuitos implementados, estudiar la efectividad de las herramientas y formatos en uso, y realizar perfilamientos y análisis de los objetos digitales del repositorio a fin de controlar la eficacia de las tareas de preservación.

Como es de suponerse, las actividades de preservación pueden requerir una importante carga adicional para los administradores del repositorio. Por ejemplo, como se mencionó, será necesaria la incorporación manual o la verificación y corrección, si son incorporados por el software, de un nuevo conjunto de metadatos de preservación que pueden ser descriptivos (soporte, identificadores), estructurales (capítulos, índices, relaciones) y administrativos (formato, versión del software, resolución, compresión).

Además de la inclusión de los metadatos, las preservación digital implica el análisis de los formatos de los archivos digitales que se ingestan al repositorio, la selección del mejor formato de transformación o migración, la transformación o migración en sí desde el archivo original hacia su correspondiente formato preservable, la verificación de la correcta transformación a fin de comprobar que no se han generado efectos indeseados que impidan la reproducción apropiada del contenido, la validación según las reglas del estándar utilizado y el cumplimiento de las normativas requeridas para el archivo resultante según el formato al cual ha sido migrado, y finalmente su almacenamiento en un medio adecuado. Sin embargo, las actividades de preservación no finalizan aquí: los archivos almacenados deben almacenarse en distintos medios mediante un sistema de copias de seguridad desatendidas (sin intervención humana), en lo posible geográficamente distribuidas, a fin de asegurar que no se perderán en caso de

1 Estudiante avanzado de Licenciatura en Sistemas. PREBI-SEDICI, Universidad Nacional de La Plata. paula@sedici.unlp.edu.ar

2 Doctor en Ciencias Informáticas. PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. gonzalo@prebi.unlp.edu.ar

3 Doctor en Ciencias Informáticas. Investigador independiente de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires; PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. marisa.degiusti@sedici.unlp.edu.ar

4 Licenciado en Sistemas. PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. alira@sedici.unlp.edu.ar

catástrofes; también deberán realizarse controles periódicos a fin de asegurar su integridad, y es aconsejable una revisión periódica de los formatos para asegurar la mejor selección de los mismos atendiendo los supuestos de la preservación digital. En fin, no basta con establecer un plan de preservación, sino que es necesario ejecutarlo y revisarlo periódicamente.

Materiales y metodología

En este trabajo se hará énfasis en el conjunto de tareas que realiza la administración del repositorio relativas al análisis, transformación y validación de los objetos digitales a los fines de la preservación. También se expondrán algunos casos especiales detectados, en donde la conversión del objeto digital no sigue un camino estándar establecido y que por lo tanto requieren una evaluación individual para determinar su tratamiento. Se mencionarán las pruebas y análisis realizados con un conjunto de herramientas informáticas y utilizando objetos reales del repositorio SEDICI de la UNLP.

Resultados parciales y conclusiones

Se presentará aquí un modelo de trabajo semi-automático, mediante el cual los administradores delegan el análisis y transformación de estos objetos a un conjunto de herramientas informáticas, las que a su vez brindan un reporte de las tareas realizadas y los resultados obtenidos. Del conjunto de herramientas analizadas, se destacarán aquellas que fueron finalmente seleccionadas para la realización de las actividades de preservación y los motivos de su selección. También se explicará en detalle la metodología de trabajo implementada a fin de agilizar las tareas de la administración del repositorio a la hora de transformar los objetos digitales y disminuir la carga de procesamiento de los equipos informáticos.
El modelo aquí propuesto se ha implementado por el momento en los documentos de texto, para los que se utiliza el formato PDF/A (en alguna de sus variantes) descrito en las normas ISO 19005-1, ISO 19005-2 e ISO 19005-3, pero como se verá, su diseño e implementación permite fácilmente la incorporación de otros tipos de documentos no textuales, como por ejemplo imágenes o archivos de audio.

Palabras claves: repositorios institucionales, preservación digital, metadatos de preservación, actividades de preservación.

Abstract
Institutional repositories have the responsibility to manage, preserve and give access to the scientific production of the institution. To this end, they must establish policies that ensure digital objects authenticity, that avoid partial or total lost of files, and that make contents available to the designated users' community. These policies must be carried out through a set of preservation activities that, preferably, should be integrated into the everyday administration tasks. Clearly, periodical reviews about the methods, tools and workflows under use must be performed, as well as profilings over random samples from the repository, in order to make sure that the preservation plan is being kept and that results are correct according to the current formats and access platforms.

Preservation activities, such as verification and correction of metadata, impose an additional workload to the repository administration. Besides these efforts, correct documents preservation require to understand file formats, and to transform files info preservable versions if possible, which in turn require also to check for correctness of transformed objects to make sure they can still be opened and accessed. In addition to this work, data preservation requires every chunk of information to be mirrored in more than one device, preferably in geographically distributed environments. Ideally, this must be done transparently for users and on an automatic basis.
This work is focused on all the administration tasks required to analyze, transform and validate digital objects to preserve them. A semi-automatic work model is presented here, in which administrators delegate the analysis and transformation of these objects to a set of computing tools, which besides the actual execution of tasks, give back to the administrators a report of all performed events and its results. It will also explained in detail the workflow performed by the staff of the repository to speed up all tasks and to minimize the processing load of computing hardware.
The model proposed here has been implemented for text documents so far. PDF/A format and its variants (ISO 19005-1, ISO 19005-2 and ISO 19005-3) has been used here. However, this implementation allow the addition of other file types, such as audio or images.

Keywords: institutional repositories, digital preservation, metadata preservation, preservation activity.

Introducción

La preservación de la producción científica y académica de las instituciones cobra mayor relevancia a medida que las mismas generan recursos, resultantes de los conocimientos de las personas y de sus expresiones, que nacen cada vez más en formas digitales (De Giusti et al. 2014; De Giusti et al. 2012). Los productos de origen digital son valiosos, constituyen un verdadero patrimonio a conservar a futuro para la sociedad, pero pueden no contar con un respaldo físico (por ejemplo en papel). Cabe destacar que el simple hecho de almacenar copias digitales de estos recursos no es suficiente para preservarlos a lo largo del tiempo: los objetos digitales (OD) deben ser descriptos correctamente para, por ejemplo, conocer su trayectoria y modificaciones, deben utilizarse formatos apropiados que permitan continuar transformándolos a medida que surgen nuevos formatos o plataformas de acceso, deben almacenarse de forma tal de evitar las pérdidas por fallas de los dispositivos de almacenamiento, etcétera (PREMIS 2016; CCSDS 2002). Estos son sólo algunos de los desafíos que las instituciones deben enfrentar a la hora de asegurar la perpetuidad de su patrimonio digital, y es aquí donde los repositorios digitales cobran un rol central gracias a su posicionamiento en la organización como principal medio para almacenar, dar visibilidad y difundir la producción institucional.

Contexto

Este proyecto se realiza en el marco del repositorio SEDICI, perteneciente a la Universidad Nacional de La Plata (De Giusti et al., 2008). El repositorio alberga recursos de todas las unidades académicas de la UNLP, así como también de sus programas, direcciones y proyectos. Esto constituye una de las características principales de SEDICI: un flujo constante de recursos ingresan constantemente al repositorio, de orígenes y tipologías muy variadas, que son verificados, catalogados, adaptados, preservados y difundidos por un conjunto de personas (administradores del repositorio). Estos administradores pueden cumplir distintos horarios laborales, poseen distinto grado de conocimientos técnicos, y sus entornos de trabajo pueden alterarse a lo largo del tiempo: compra de nuevas computadoras, instalación de nuevos sistemas operativos, cambios en las aplicaciones que utilizan (navegadores, herramientas de oficina, sistemas de seguridad), etcétera. Si bien esta variabilidad otorga una gran flexibilidad desde el punto de vista organizacional, puede presentar un desafío a la hora de disponer de las herramientas que aseguren el correcto y equitativo procesamiento de los OD que se alojan en el repositorio. Una de las ventajas principales del modelo que se propone en este trabajo es la separación entre los equipos de trabajo de los administradores y el software de procesamiento de documentos, lo cual ayuda a mantener la flexibilidad de este tipo de entornos. Por otro lado, como se verá más adelante, el modelo es fácilmente escalable y replicable, lo que constituye una enorme ventaja tanto si la carga de trabajo continúa creciendo como también para otras organizaciones que deseen implementar un modelo similar.

Al día de hoy SEDICI cuenta con cerca de 50 mil recursos, y el 95% de ellos son documentos de texto en formato PDF (De Giusti, 2014). Si bien el modelo aquí propuesto no está sujeto al formato PDF exclusivamente, debido al elevado porcentaje

de este formato en su primera implementación para este repositorio, se trabajó sobre documentos de texto en este formato.

Formatos de preservación

La preservación de objetos digitales requiere, como se comentó arriba, el uso de formatos apropiados que aseguren el acceso a los datos alojados dentro de dichos objetos y que permitan realizar transformaciones y adaptaciones para asegurar su acceso a medida que las herramientas de acceso avanzan y que surgen nuevos formatos de archivos. Esto se aplica a cualquier tipo de recurso digital, pero como es de esperarse, la selección del formato de preservación deberá adecuarse al tipo concreto de OD: documentos de texto, imágenes, audios, videos, datos crudos, objetos espaciales, etcétera (Kresse et al. 2015). A la hora de seleccionar un formato de preservación, deben tomarse en consideración diversos aspectos, como por ejemplo el grado de estandarización del formato, el nivel de apertura de la licencia de dicho formato, la posibilidad de pérdida de información al transformar OD hacia ese formato (por ejemplo, al utilizar algoritmos de compresión con pérdida), la facilidad a la hora de transformar OD hacia estos formatos, el tamaño final de los archivos, entre otros (Brown, 2008; Giménez Chornet, 2014).

En el caso particular de los documentos de texto, el estándar PDF/A, descrito en las normas ISO 19005-1, ISO 19005-2 e ISO 19005-3, se impone como el formato más apropiado para su preservación. Este formato está basado en el estándar PDF 1.4, al que le incorpora algunos requerimientos adicionales, como ser:

1. Especificaciones sobre los metadatos y la estructura del archivo.
2. La paleta de colores (incluyendo escala de grises y blanco/negro) no deben ser representados en un espacio de color de dispositivo (DeviceRGB, DeviceCMYK, DeviceGray).
3. Las fuentes usadas en texto visibles deben estar embebidas (incluídas dentro del archivo).
4. Para un PDF/A-1a, la estructura original del documento se mantendrá igual al documento PDF/A original. Es decir, no se crearán nuevas etiquetas y la estructura no deberá cambiar, por lo que para crear un archivo PDF/A-1a, el OD debió haber sido creado estructurado y etiquetado. En cualquier otro caso, se optara por PDF/A-1b.

La principal diferencia entre PDF/A-1b y PDF/A-1a, es que PDF/A-1a posee especificaciones adicionales sobre PDF/A-1b, como por ejemplo:

1. Las fuentes embebidas tienen otros requerimientos, como ser, su representación en Unicode (ISO 19005-1, capítulo 6.3.8).
2. El documento debe contener una estructura lógica (ISO 19005-1, capítulo 6.8).

Uno de los propósitos de los requerimientos del estándar PDF/A-1a es de proveer soporte para personas con capacidades diferentes, por ejemplo, incorporando la información requerida y necesaria para aplicaciones que hagan el pasaje de texto a voz. La estructura lógica del documento es una descripción del contenido de las páginas, que debe ser dada por el documento original y consiste en un correcto etiquetado jerárquico que distingue el verdadero contenido de los artefactos incluidos en el documento (números de páginas, pies de página, artefactos de maquetación,

etc.). El etiquetado provee una breve descripción, y esta debe ser sencilla y fácilmente comprensible para las personas, por eso no puede ser generada en una conversión (no es posible derivarla automáticamente por el software de conversión) sino que debe ser generada al momento de la creación del OD. Esta es una de las razones por la cual no cualquier PDF puede ser convertido a PDF/A-1a.
El PDF/A-2 descrito en la norma ISO 19005-2, está basado sobre la norma ISO 32000-1, correspondiente al estándar PDF 1.7, y este a su vez es una extensión complementaria del estándar PDF/A-1. Las principales diferencias entre PDF/A-1 y PDF/A-2 son:

1. Se agrega la compresión de JPEG2000 para imágenes.
2. Las imágenes con transparencias son permitidas.
3. Contenidos adicionales (conocidos como *layers* o capas) pueden ser visibles o no.
4. Muchos PDF/A pueden combinarse para crear un solo archivo.
5. Se crea un nuevo nivel de conformidad, el nivel U (unicode), que permite crear un PDF de búsqueda sin tener que cumplir con el estricto nivel de conformidad A (A por accesibilidad).

Los documentos que contengan las características descritas anteriormente, en particular las transparencias o layers, deberán ser convertidos a PDF/A-2, rara vez a PDF/A-1.
El estándar PDF/A-3, descrito en la norma ISO 19005-3, está basado sobre ISO 32000-1 (PDF 1.7). Es a su vez una extensión complementaria de PDF/A-2. Algunas de las diferencias principales entre PDF/A-2 y PDF/A-3 son:

1. Los archivos, de cualquier formato y conformancia, deben ser embebidos en el PDF/A creado.
2. Los archivos embebidos pueden ser asociados a cualquier parte del PDF/A-3 creado.

A partir de las descripciones previas, es evidente que sería ideal alcanzar el estándar PDF/A-1a. Sin embargo, esto no siempre es posible debido a las características particulares de cada archivo. En caso de no poder llegar a ese formato, se aceptan también otros formatos de PDF/A.

Propuesta de servidor de validación y transformación

La realización de actividades de preservación puede requerir una importante carga adicional para los administradores del repositorio, como por ejemplo al convertir archivos entre formatos o en el momento de verificar y corregir estos archivos. Al realizar estas actividades, es necesario minimizar tanto la dependencia con los administradores (o sea, qué nivel poseen tanto de conocimientos técnicos como del uso de las herramientas de transformación y verificación) así como también cualquier decisión sujeta a consideraciones subjetivas, como por ejemplo cuál es el nivel de calidad mínimo requerido para determinar que una transformación sea o no aceptable. De este modo, no solo se asegura un nivel de calidad mínimo general para todos los documentos, sino que también se facilita la incorporación o el intercambio de personas en el staff de administración del repositorio.
La política de preservación digital establecida en SEDICI requiere que la carga de objetos digitales al repositorio debe realizarse siempre en un formato preservable. Por

otro lado y como es de esperarse, las actividades de catalogación y carga no se concentran en una única persona, sino que participan varias personas que realizan por lo general cualquiera de las tareas necesarias para obtener archivos aptos para su carga: división de documentos, transformación de formatos, corrección de metadatos de los documentos, incorporación de OCR, entre otras. Por este motivo, al diseñar el modelo de trabajo aquí descrito, se buscó implementar un sistema independiente de las personas y que promueva el trabajo colaborativo en equipo. Como ganancia adicional, se logró una implementación que puede accederse desde cualquier computadora de la red de trabajo, independiente de su sistema operativo, arquitectura o incluso capacidad de cálculo, y que evita la instalación de herramientas particulares en las computadoras de trabajo de los administradores, lo que agiliza la incorporación de equipos y brinda flexibilidad a la hora de cambiar equipos o realizar actualizaciones o cambios en los sistemas operativos subyacentes.

La propuesta de este trabajo se basa en un modelo de trabajo centralizado, que concentra el procesamiento de los documentos en un equipo dedicado, y desatendido, lo que significa que los administradores no necesitan intervenir en el procesamiento de dichos documentos, aunque se requiere su intervención a la hora de analizar y validar los documentos procesados, y eventualmente realizar operaciones adicionales para completar este procesamiento. La implementación de este modelo implica un sistema de red estilo cliente-servidor muy sencillo y fácil de replicar, en el que los clientes (los administradores) acceden a un directorio compartido en la red y depositan los documentos que deben ser procesados, que luego son tomados por un proceso (servidor) que los analiza, transforma y deposita en otro directorio de la red el resultado de esta transformación. Para la compartición de los directorios se utilizó la tecnología de directorios sobre redes TCP/IP de MS Windows conocida como *Server Message Block* (SMB, Microsoft 2016), que también está disponible en equipos con el sistema operativo MAC OS X y en sistemas GNU/Linux mediante el cliente Samba (Samba 2016); de este modo, cualquier computadora de la red capaz de acceder a una *carpeta compartida en la red* podrá utilizar este servidor de transformaciones. Asimismo, dado que la propuesta es independiente del mecanismo de compartición de directorios en red, este mismo modelo puede también replicarse sobre el protocolo Common Internet File System (CIFS, 2016), lo que permite superar la red de área local y ampliar el espacio de trabajo hacia cualquier equipo conectado a Internet. Incluso es posible implementarse mediante otros sistemas de directorios compartidos, como Network File System (NFS, 2016) o utilizando aplicaciones en la nube como por ejemplo Owncloud (Owncloud, 2016) o Dropbox (Dropbox, 2016).

El espacio de trabajo (el directorio compartido) se organizó en distintas áreas o secciones, que internamente se corresponden con directorios del sistema de archivos. Esto permite organizar el trabajo ya realizado del trabajo que aún falta realizar, y ayuda a los administradores a comprender y utilizar este modelo. En una primera implementación, el modelo contaba con tres secciones:

1) Sección In: espacio en donde los administradores depositan los archivos a convertir.
2) Sección Out: espacio donde son depositados los archivos convertidos satisfactoriamente, o sea que fueron transformados y verificados.
3) Sección Advertencia: espacio donde se colocan los archivos que no han pasado el validador post conversión.

Si bien este esquema resultaba muy simple de utilizar, al revisar el modelo fue necesario incorporar dos secciones adicionales para manejar ciertos casos especiales:

4) Sección Válido: directorio donde se copian los archivos que ya se encuentran en un formato de preservación (o sea, que no requieren ningún procesamiento) y que han pasado su correspondiente validación.
5) Sección OCR: directorio donde se depositan los documentos que directamente no se pueden convertir. Por lo general a estos documentos no se les ha realizado un reconocimiento de caracteres y posterior extracción de texto, requisito fundamental para los documentos de texto del repositorio.

Para los clientes o usuarios finales (los administradores que requieren las transformaciones), la metodología de trabajo es muy simple (imagen 1). Cuando se recibe un documento en formato PDF (por ejemplo una tesis, un artículo o un libro), el administrador a cargo del mismo accede al directorio compartido en la red, y copia el archivo original en la *Sección In*. Luego, espera a que el servidor realice las transformaciones necesarias, y a continuación ingresa a la Sección Out, donde encontrará tanto el documento transformado a un formato preservable junto a un archivo de registro de la transformación. En caso que el archivo obtenido no haya cumplido con la validación, se deberá buscar en la Sección Advertencia, y revisar el registro de transformación correspondiente para analizar qué ha sucedido y evaluar su reparación. Si el archivo esta corrupto y no hay forma de convertirlo o repararlo, pasa a la sección OCR donde se realiza el reconocimiento óptico de caracteres. Este procesamiento se realiza al día de hoy con el software ABBYY FineReader (ABBYY, 2016), el cual una vez finalizado permite exportar el documento en formato PDF/A. Por otro lado, en caso que el archivo ya se encontraba en un estándar de PDF/A válido, el administrador lo encontrará en la Sección Válido, junto a un breve reporte donde se confirma la validez del mismo. En todos los casos, además del reporte particular de transformación y validación que se obtiene por cada archivo, se registran los eventos en una bitácora general, que luego permite obtener estadísticas de uso de la aplicación (archivos procesados, transformaciones incorrectas, tipos de errores encontrados, etc). Este registro general sirve como retroalimentación para continuar mejorando y ampliando este modelo.

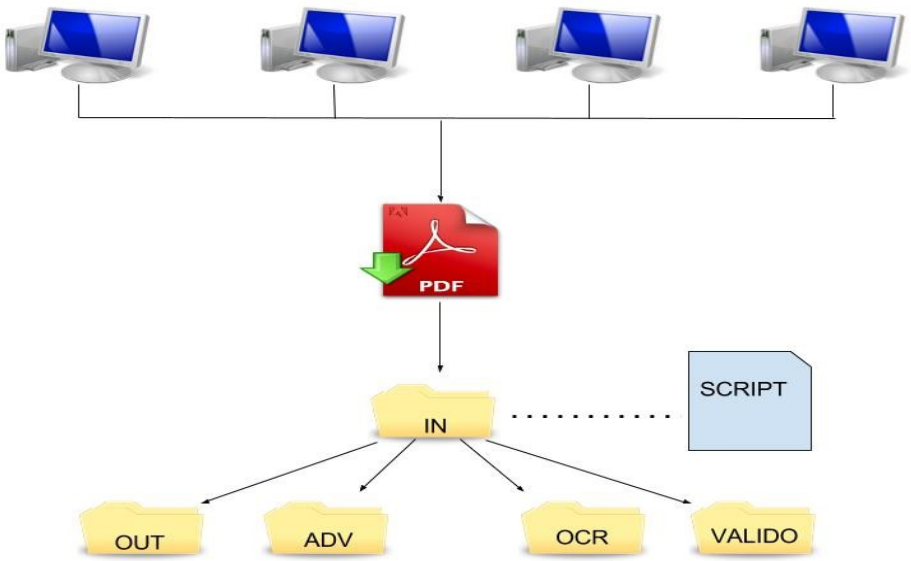

Imagen 1: Esquema de trabajo general. Los administradores acceden a un directorio de red, donde depositan los documentos PDF, que luego son tomados por un *script*, procesados y almacenados en otro directorio según el resultado del procesamiento (OUT, ADV, OCR, VÁLIDO)

Algunos OD no aprueban el post análisis porque no cumplen el estándar en el que se intentó convertir. Esto suele pasar cuando el algoritmo falla en la detección del nivel o de la conformidad si, por ejemplo, se intentó convertir un OD a PDF/A-1b, pero el archivo contiene imágenes con transparencias. En este ejemplo, un posible camino sería intentar convertir el OD a PDF/A-2b; otra posible opción sería eliminar las transparencias. Al día de hoy, este análisis y reparación se debe realizar de manera manual por los administradores, que encontrará estos "casos especiales" en la sección Advertencia. Sin embargo, en caso de que no haya un camino posible para una conversión exitosa, el archivo PDF deberá pasar a la sección OCR para procesarse con el software ABBYY.

Respecto a la conversión de PDF a PDF/A, se realizó un relevamiento sobre distintas herramientas en el que se consideraron diferentes parámetros como ser la licencia (software libre vs sistema cerrado), la interfaz de uso (línea de comandos, aplicación de escritorio y proceso servidor) y plataforma (MS Windows, GNU/Linux). La herramienta pdfaPilot resultó muy apropiada, pero su versión para servidores tenía un costo de adquisición mayor al presupuesto disponible, y por ellos se optó por adquirir la versión CLI para GNU/Linux, cuyo uso se debe realizar por medio de un intérprete de comandos de dicho sistema operativo. Adicionalmente, por medio de un programa (script) escrito en lenguaje Bash, se encapsularon y se brindó acceso a los módulos encargados de realizar las distintas tareas sobre los documentos, como ser la detección de archivos pendientes de transformación, la conversión entre formatos, la validación, el registro de las actividades realizadas en la bitácora, la salida hacia el usuario, etcétera. Este script se desarrolló con un área dedicada a las variables de configuración del sistema (por ejemplo, niveles de registro o rutas hacia los programas ejecutables), lo que permite adaptar su funcionamiento en diferentes entornos de ejecución. El script en cuestión se ejecuta de manera contínua en segundo plano, y cada vez que detecta que se han copiado archivos o directorios dentro de la Sección IN, ejecuta las aplicaciones

que realizan los análisis y transformaciones correspondientes, y almacena los resultados en los directorios apropiados según el esquema previamente descrito.
El procesamiento recursivo de directorios es un arma de doble filo. Si bien puede ahorrar tiempo a los administradores, ya que podrían copiar un directorio con muchos subdirectorios dentro con archivos a procesar o con más subdirectorios, esto puede generar una sobrecarga de trabajo en el servidor, manteniendo procesos en ejecución por mucho tiempo y ocupando demasiados recursos de procesamiento. Por este motivo, en el módulo de procesamiento central de la herramienta se incorporó la posibilidad de copiar directorios enteros con archivos dentro de la sección IN, pero se limitó la recursión a un sólo nivel de acceso. Esto también acorta los tiempos de copiado de archivos en la red, lo cual podría generar otro punto de falla dado que los archivos sólo deben procesarse una vez que se han copiado completamente. Desde el punto de vista de la seguridad, se incorporaron dos módulos adicionales. El primero de ellos se encarga de brindar exclusión mutua sobre cada archivo, verificando que el mismo no esté siendo accedido por algún proceso de conversión previo. El segundo módulo está encargado de verificar que el archivo ingresado es un documento PDF correcto, o lo que es lo mismo, que no se ha corrompido al descargarse o copiarse a través de la red.
El programa pdfaPilot permite que se le especifique el nivel (1, 2 o 3) y el nivel de conformidad buscado (a,b,u). Para hacer uso de estas opciones, se encapsuló en un módulo la configuración de la ejecución de las transformaciones, en el cual se determina cuál es estándar más indicado para cada PDF. Para ello, previamente se realiza un análisis de los archivos PDF, mediante otras herramientas como exiftool o jhove que, entre otras opciones, permiten extraer los metadatos contenidos dentro de los archivos PDF. Un simple análisis de los metadatos extraídos permite determinar el nivel de conformidad alcanzable; por ejemplo, si el documento ingresado no es “taggeado” (etiquetado), no será posible alcanzar la conformidad “a” de PDF/A, y por lo tanto se optará por la conformidad “b”.
Todas las decisiones tomadas, los resultados alcanzados, el motivo y los errores encontrados durante el procesamiento deben ser informados de alguna manera a las personas que solicitan el procesamiento de los archivos. Por este motivo, se incorporó un módulo encargado de registrar los eventos sucedidos y generar un reporte final para los usuarios. Este módulo, además de incorporar el reporte generado por el software convertidor, genera tres tipos de reportes:

1. Reporte estadístico general: Archivo de texto donde, para cada OD detectado en la Sección In, se agrega el nombre del archivo, fecha y resultado del procesamiento, es decir, si el archivo terminó en la Sección Out, Válido, Ocr o Advertencia
2. Reporte del documento: por cada archivo procesado se crea reporte con una breve descripción de lo que la herramienta realizó, su resultado post-conversión y un código informativo (imágenes 2, 3 y 4).
3. Reporte diario: Archivo de texto generado por día con una síntesis de los OD procesados.

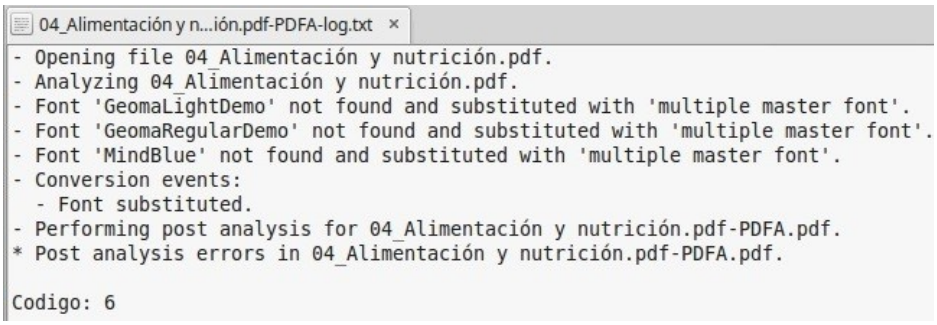

Imagen 2: La imagen muestra el reporte de conversión del archivo “04_Alimentación y nutrición.pdf”. Falla el post-análisis por lo que el archivo original junto a su reporte son enviados a la Sección Advertencia (código de error 6).

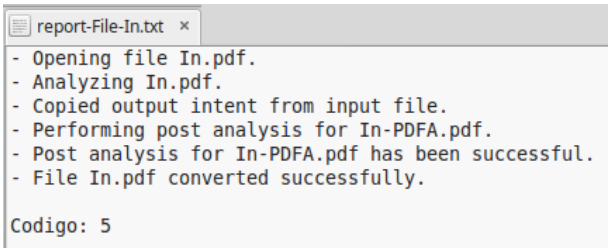

Imagen 3: Reporte de conversión de un archivo satisfactorio (código de aceptación 5). El OD original, el OD convertido a PDF/A y el reporte son enviados hacia la Sección Out.

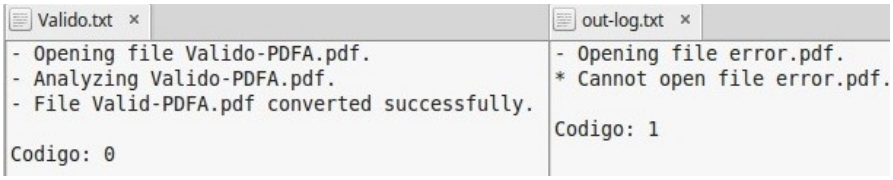

Imagen 4: El reporte de la izquierda corresponde a un documento que ya se encontraba en formato PDF/A valido, por lo que es llevado a la Sección Valido, con código de aceptación 0. El reporte de la derecha corresponde a un archivo que no se pudo procesar por un algún error, generalmente, el archivo está malformado o corrupto, por lo que es enviado a la Sección OCR con código de error 1.

Mejoras y actualizaciones

Durante la primera etapa de uso de este desarrollo, se observó que la detección del nivel y conformidad de PDF/A no resultaba apropiada en todos los casos (estaba muy sujeta a las características internas de cada archivo a procesar). Sumado a esto, el tiempo de transformación de estos archivos se incrementó de manera considerable, pues se repetían numerosas veces los procesos de análisis y conversión, hasta obtener finalmente un resultado fallido. Por estos motivos se buscó una segunda herramienta capaz de realizar estas tareas, o que permitiera dar un nivel y conformidad por defecto, y que sea capaz de alterar el nivel o la conformidad buscados en caso de ser necesario. Se optó por la herramienta *3Height converter shell* (3H, imagen 5) para GNU/Linux, que además de contar con estas opciones, la empresa desarrolladora (PDF-tools) proveyó un buen servicio de servicio técnico y otorgó un importante descuento de adquisición por tratarse de una institución educacional. El cambio de aplicación no requirió mayores modificaciones en el script de ejecución, ya que sólo fue necesario actualizar la ruta del programa ejecutable y modificar los parámetros según son requeridos por la nueva herramienta.

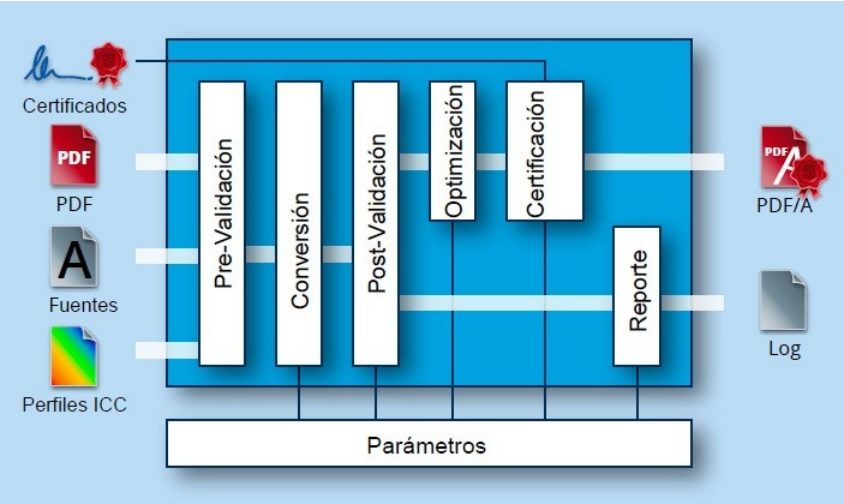

Imagen 5: Esquema de procesamiento por etapas de 3H (Fuente: manual de usuario de 3H, traducido al castellano).

Una de las especificaciones del estándar PDF/A establece que las fuentes deben incluirse (embeberse) en los documentos; dicho de otro modo, no se aceptan fuentes externas. La elección de la fuente a utilizar presenta algunos inconvenientes, principalmente cuando se trata de fuentes privativas o fuentes no disponibles en todas las plataformas y sistemas operativos. Para resolver esta situación, se sumó un conjunto de paquetes de fuentes (por ejemplo fuentes *TrueType*) al conjunto de fuentes por defecto, con lo cual en caso de no encontrarse una fuente específica, podrá optarse por una versión alternativa del paquete adicionado. Esto requirió unas modificaciones en el software 3H por parte de la empresa desarrolladora, lo que generó una nueva versión de esta aplicación.

Otras situaciones también requirieron intercambios con el soporte técnico. Como se mencionó previamente, si no se especificaba un nivel y conformidad, el software seleccionaba aquel que parecía correcto mediante su algoritmo de decisión, pero en algunos casos el OD convertido era nivel “2” conformidad “u”, cuando el mismo cumplía los requerimientos para ser “1a”. Al observar este comportamiento, se notificó al soporte técnico, quienes generaron una nueva versión incorporando un provisorio “upgrade detection” más exacto. En la versión actual estos asuntos fueron resueltos mediante la incorporación de dos parámetros adicionales, uno que sube el nivel (de 1 a 2) y otro que baja la conformidad (de “a” a “b” en caso de nivel 1, y de “a”,”u”,”b” en nivel 2).

Desde el punto de vista de la validación de los documentos convertidos, se realizaron pruebas tanto con 3H (en su última versión) como con el validador de Acrobat Pro DC. Si bien en líneas generales ambos funcionaban correctamente, se encontraron dos metadatos que eran tratados de manera diferente. Por un lado, la forma de definir los metadatos de fecha no era detectada apropiadamente por Acrobat DC. Se notificó al servicio técnico de Adobe con respecto al *XMP metadata Create Date*, que debe especificarse en el estándar ASN.1 como se indica en la norma ISO/IEC 8824; el soporte técnico confirmó que el mismo será corregido en una futura versión del software. Finalmente, se encuentra abierto el debate acerca del contenido del metadato History, ya que es interpretado de manera diferente por 3H y por Acrobat DC, y el estándar no aporta claridad al respecto.

Conclusiones

Una de las ventajas del modelo aquí presentado yace en la flexibilidad que ofrece en cuanto al trabajo de los administradores y la independencia de sus estaciones de trabajo respecto al as herramientas de procesamiento de documentos. En contextos donde surgen constantemente nuevas plataformas y formas de acceso y compartición de recursos, este tipo de flexibilidad sirve de apoyo para continuar expandiendo los equipos de trabajo o adaptando los entornos según las necesidades de cada usuario.

Por otro lado, el modelo propuesto se destaca por su facilidad de replicación en otras organizaciones. Las tecnologías que utiliza son compatibles con la mayoría de las plataformas, y en las estaciones de trabajo sólo es necesario acceder a un directorio compartido en red. Desde el lado del servidor, el script desarrollado ofrece un conjunto de parámetros de configuración, lo que demostró ser muy útil al cambiar de herramientas y por lo tanto los parámetros de ejecución de las mismas (de pdfaPilot a 3H), y que servirá también para la instalación en otros contextos. En lo relativo a los costos, si bien se adquirió una licencia de software para el software 3H, se obtuvo un importante descuento que permitió mantener el proyecto dentro del presupuesto. Asimismo, muchas de las herramientas utilizadas son de código abierto (servidor Linux, Samba, exiftool, jhove), lo que también asegura que los costos se mantendrán al mínimo a lo largo del tiempo.

Particularmente en este trabajo con respecto al software de 3H, solamente se mencionó de su módulo de conversión y validación, pero el mismo cuenta con más opciones de configuración. Por ejemplo, una de ellas le aporta al PDF/A creado la optimización (linearizado o con vista rápida para la web). Con Vista rápida en Web, el servidor Web sólo carga la página solicitada, en lugar del PDF completo. Esta opción es deseada en documentos gran tamaño, que pueden tardar mucho tiempo en descargarse desde un servidor.

En cuanto a los tiempos de procesamiento, la conversión y validación de un documento PDF tradicional es casi instantáneo (aproximadamente 1 segundo). La revisión de la Sección IN para detectar OD a procesar se realiza cada 10 segundos (este valor puede configurarse también desde el script). Con estos valores temporales, en la mayoría de los casos los administradores copian archivos y directorios en la sección IN, y obtienen los documentos transformados en menos de 20 segundos, lo cual agiliza enormemente la tarea de carga. Si bien se requieren conocimientos más avanzados en los distintos estándares de PDF/A para la reparación de "casos especiales", en la mayoría de los casos esto no sucede y, por lo tanto no se ha generado un cuello de botella por el tratamiento de estos documentos. Esta necesidad de acción manual en ciertos casos presenta uno de los principales puntos a mejorar, pues sería ideal lograr un procesamiento totalmente automático o al menos minimizar tanto como sea posible los casos particulares que requieren intervención humana. Esta automatización puede alcanzarse mejorando los algoritmos de elección de estándares o incorporando nuevas herramientas de detección y reparación, y será uno de los trabajos a futuro del proyecto.

La incorporación de fuentes presenta también algunos desafíos particulares. Día a día se crean nuevas fuentes, algunas de ellas abiertas pero otras privativas, y en muchos casos no disponibles en todas las plataformas. Se hace muy difícil contar siempre con todas las fuentes requeridas, por lo que en ocasiones se opta por una fuente alternativa, que puede no ser igual a la fuente original del documento. Si bien esto

asegura que el contenido del documento (el texto por ejemplo) será accesible a lo largo del tiempo, que es el fin último de estos esfuerzos por asegurar la preservación digital, sería deseable mantener los documentos transformados tan parecidos a los originales como sea posible. Quizás este punto requiera un trabajo de concientización de los usuarios, para que minimicen o eviten el uso de fuentes incompatibles o privativas, pero esto ya excede los objetivos del proyecto.

Es posible también considerar la escalabilidad de esta propuesta. Actualmente el procesamiento de los OD se realiza de manera individual y secuencial (uno detrás del otro). Sería deseable paralelizar estas tareas, ya sea para aprovechar las ventajas de los equipos modernos con múltiples unidades de procesamiento, o incluso para distribuir la carga de trabajo en equipos de la red. Si bien al día de hoy esto no es un problema en el contexto de SEDICI, el crecimiento sostenido del repositorio o incluso la realidad de otras organizaciones y contextos amerita considerar seriamente una solución que permita multiplicar la carga de trabajo por varias veces sin que esto repercuta en la eficiencia del servicio.

Una observación no menor respecto a esta propuesta es que no se encuentra integrada con el software Dspace que utiliza el repositorio. Los administradores deben descargar los documentos desde Dspace, procesarlos por fuera, y luego cargar las versiones ya procesadas. Sería ideal que esto se realice de manera transparente, y que los administradores dispongan directamente de los documentos originales y procesados, y se concentren en la descripción de los mismos. Esto requerirá probablemente modificaciones en el software Dspace, así como también evaluar nuevamente el flujo de trabajo de los administradores del repositorio, pero sin dudas ofrecerá grandes ventajas para el trabajo diario del repositorio.

Para finalizar, cabe destacar que el proyecto completo se encuentra disponible en el repositorio Github, para que cualquier organización pueda descargar los scripts, utilizarlos o personalizarlos según sus necesidades particulares. La URL del proyecto en Github es https://github.com/sedici/scripts-digitalizacion.

Bibliografía

Brown, A. (2008) Digital Preservation Guidance Note 1: Selecting file formats for long-term preservation. The National Archives. Recuperado en https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

CCSDS, Reference Model for an Open Archival Information System (OAIS):ISO 14721 . 2002.

Common Internet file system https://technet.microsoft.com/en-us/library/cc939973.aspx

De Giusti M. R., Sobrado A.,Lira A. J., Vila M. M., and Villarreal G. L. (Sep. 2008) “SeDiCI (Servicio de Difusión de la Creación Intelectual),” D-Lib Magazine, vol(14).

De Giusti M.R, Lira A. J., Texier J., and Villarreal G. L. (2012) “Las actividades y el planeamiento de la preservación en un repositorio institucional”. Conferencia Internacional BIREDIAL-ISTEC, Universidad del Norte, Barranquilla - COLOMBIA. Recuperado en http://sedici.unlp.edu.ar/handle/10915/26045

De Giusti, M. R., Lira, A. J., Villarreal, G. L., Terruzzi, F. A.; Adorno, F. G. (2014) Preservación digital: un experimento con SEDICI-DSpace. XX Asamblea General de ISTEC (Puebla, México). Recuperado en http://sedici.unlp.edu.ar/handle/10915/34889

De Giusti, M. R.(2014) "Una metodología de evaluación de repositorios digitales para asegurar la preservación en el tiempo y el acceso a los contenidos." Tesis doctoral, Universidad Nacional de La Plata. Recuperado en http://sedici.unlp.edu.ar/handle/10915/43157
Descarga la versión de prueba de adobe acrobat | acrobat pro DC https://acrobat.adobe.com/la/es/free-trial-download.html
Dropbox (IE) https://www.dropbox.com/
ExifTool by Phil Harvey http://www.sno.phy.queensu.ca/~phil/exiftool/
Giménez Chornet, V. (2014). Criterios ISO para la preservación digital de los documentos de archivo. Códices, 10(2), 135-150. Recuperado de http://revistas.lasalle.edu.co/index.php/co/article/view/3267/2607
ISO 19005-1. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A)
ISO 19005-2. Document management - Electronic document file format for long-term preservation - Part 2: Use of ISO 32000-1 (PDF 1.7)
ISO 19005-3. Document management - Electronic document file format for long-term preservation - Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)
PDF/A-3: ISO 19005-3
ISO 32000-1:2008. Document management -- Portable document format -- Part 1: PDF 1.7
JSTOR/Harvard object validation environment http://jhove.sourceforge.net/
Kresse W., Pau J.M., “Development of an ISO-standard for the preservation of geospatial data and metadata: ISO 19165,” Photogrammetrie, Fernerkundung, Geoinformation, vol. (Dec. 2015), Pages 449-456.
Linux NFS faq http://nfs.sourceforge.net/
Microsoft SMB protocol and CIFS protocol overview https://msdn.microsoft.com/en-us/library/windows/desktop/aa365233(v=vs.85).aspx
Microsoft typography - OpenType specification https://www.microsoft.com/en-us/Typography/OpenTypeSpecification.aspx
OCR software from ABBYY. Best text recognition for windows and Mac https://www.abbyy.com/finereader/
Opening windows to a wider world https://www.samba.org/
OwnCloud.org https://owncloud.org/
PDF to PDF/A converter http://www.pdf-tools.com/pdf/pdf-to-pdfa-converter-signature.aspx
PREMIS, “PREMIS: Preservation Metadata Maintenance Activity (Library of Congress),” (2016). Recuperado en http://www.loc.gov/standards/premis/
Products | Callas Software. https://www.callassoftware.com/en/products/pdfapilot