Taller de estrategias de preservación digital
Informe Final

A modo de cierre de las actividades del Proyecto *Preservación del patrimonio digital en bibliotecas argentinas: estudio exploratorio y experiencia piloto*, se realizó un Taller de Estrategias de Preservación Digital con varias de las instituciones participantes. El objetivo era brindar un espacio de debate y puesta en común de las distintas alternativas que cada institución comienza aplicar en favor de la preservación de los activos digitales, a la luz de las nuevas experiencias y conocimientos desarrollados en el marco del Proyecto.

Para ordenar la presentación y debate de los temas más relevantes se definió una lista de cuestiones centrales, en parte motivada por intereses específicos de la investigación planteada en este Proyecto, y en parte por consenso con las instituciones participantes, según el siguiente detalle:

1) *Ingesta de archivos digitales*

2) *Esquemas de nombramiento de archivos digitales*

3) *Metadatos*

4) *Almacenamiento de las colecciones digitales*

5) *Estrategias específicas de preservación digital*

6) *Colaboración entre instituciones*

La elección de estos temas no fue azarosa, sino que refleja las preocupaciones dominantes en el campo emergente de la preservación digital en todo el mundo. En efecto, la producción e ingesta de archivos digitales, los esquemas de nombramiento de archivos, la producción y gestión de metadatos y las maneras de almacenar la información en los sistemas informáticos son cuestiones que deben preverse y organizarse en aras de garantizar el mantenimiento del valor y la funcionalidad de las colecciones digitales, y así se lo entiende en los crecientes consensos internacionales sobre estos puntos.

Diversas instituciones participantes, como se verá, presentaron sus estrategias en curso para cada uno de estos temas centrales. Sin embargo, todavía ninguna ha llegado al punto de emplear *estrategias específicas de preservación digital*, algo que requiere de cierta evolución, y de alcanzar un cierto grado de madurez en los programas de preservación de las bibliotecas y las instituciones de la memoria. Como decíamos en un informe anterior, el campo de la preservación digital es muy joven aún en la Argentina, y eso explica la ausencia de estrategias específicas de preservación digital.

Sin embargo, esta ausencia no debería oscurecer el hecho auspicioso de que las instituciones están comenzando a enfrentar de manera organizada las cuestiones básicas de la preservación digital. Y a medida que se va ganando experiencia para la resolución normalizada de estas cuestiones básicas, se van generando también las bases para programas más maduros y sofisticados de preservación digital, tal como ha ocurrido en las principales instituciones del mundo. Necesariamente, la emergencia de tales programas maduros sólo puede acontecer como un punto de llegada, como la consecuencia de unas prácticas institucionales –que empiezan por lo más simple, hasta poder abordar lo más complejo– sostenidas en el tiempo.

Asimismo, dado que las bibliotecas, en particular, presentan un terreno favorable para la aplicación de procesos de trabajo normalizados, es de esperar que las instituciones que hoy empiezan a ocuparse de la preservación digital se conviertan en multiplicadoras de esas buenas prácticas, en favor de otras instituciones de la memoria en nuestro país.

Por otro lado, el hecho de que las instituciones participantes hayan comenzado a definir y emplear prácticas y técnicas organizacionales para la preservación digital, nos está mostrando que los cursos, talleres y reuniones realizados en el curso este Proyecto han logrado un impacto favorable, cumpliendo así con uno de los objetivos principales que nos habíamos planteado. En este sentido, y considerando que esta es un área muy poco desarrollada en Argentina, ese impacto fue posibilitado en gran medida por el fuerte interés en esta problemática que todas las instituciones participantes pusieron de manifiesto.

Taller

¿Qué puede observarse a partir de los trabajos presentados en el Taller?
En primer lugar, todas las instituciones remarcaron que una de las cuestiones principales consiste en tomar conciencia de los riesgos que corre la información en formatos electrónicos –rápido ciclo de obsolescencia, pérdida del acceso, corrupción de los archivos o de los medios de almacenamiento– y la necesidad de desarrollar o adaptar procesos técnicos y estrategias institucionales de trabajo normalizadas para enfrentar esos mismos riesgos. Hay que decir que éste es un asunto realmente crítico, ya que una actitud bastante común frente a la

posibilidad de digitalizar originales o crear colecciones electrónicas, consiste en enfocar sólo las mejoras del acceso a la información que ella brinda, olvidándose de los riesgos que aquejan a toda información en formato digital.

Resumen de las presentaciones, debates y reflexiones del Taller

Biblioteca de la Facultad de Humanidades y Ciencias de la Educación de la Universidad Nacional de La Plata
Presentación del caso

En línea con su proyecto Memoria Académica, esta biblioteca presentó las estrategias diseñadas para lograr, desde la ingesta, el desarrollo de colecciones digitales sustentables en el largo plazo. Las colecciones abarcan una variedad de tipos documentales producidos por la propia actividad académica de esa casa de estudios, y el objetivo del proyecto consiste en dar acceso en línea a esos materiales al tiempo que se los preserva en sus formatos digitales. Los originales abarcan tanto materiales impresos, como otros "nacidos digitales", entre estos últimos las tesis y artículos académicos en formato electrónico.

Se discutieron diversas situaciones que afectan el ritmo de digitalización de los originales impresos, en buena medida producto de las limitaciones existentes tanto en recursos humanos como técnicos. Asimismo, se analizaron las posibilidades y oportunidades que se abrirían por el interés de algunos editores de publicaciones en favor de producir versiones digitales de las mismas. Teniendo en cuenta estas posibilidades más inmediatas, y para lograr productos digitales consistentes y sustentables, la Biblioteca elaboró los siguientes lineamientos para la producción, ingesta y gestión de objetos digitales:

Estrategias para la Ingesta

La Biblioteca elaboró un instructivo para la digitalización y gestión de documentos digitales, considerando estrategias de preservación digital. Este instructivo esta dirigido a todo editor que desee digitalizar sus publicaciones sobre papel, basado en la premisa de que todo proceso de digitalización debe procurar que el documento original sea escaneado por única vez, garantizando que la imagen obtenida reproduzca en la forma más fidedigna posible al material original.

Específicamente, los ejes de esta estrategia se basan en los siguientes elementos:

1) Uso de una planilla de cálculo o base de datos para la gestión y el control de los materiales digitalizados que se ingestan

2) Definición de un esquema de almacenamiento en árbol jerárquico de directorios en el disco rígido, con nombres normalizados para los propios directorios –incluyendo un prefijo que identifica cada clase de documento–, así como para las imágenes digitales de las páginas digitalizadas

3) Generación de tres subcarpetas:
    master: para almacenar los master de preservación

copias: para almacenar las copias derivadas
ocr: para guardar el archivo .DOC resultado del reconocimiento de caracteres

Por otro lado, el instructivo prevé la aplicación de ciertas normas y criterios para la digitalización, basadas en la evaluación de las características significativas de cada documento original, y en el intento de lograr una reproducción fidedigna. Esto incluye criterios para decidir la aplicación de una determinada resolución de captura, profundidad de bits, y formatos de archivo, distinguiendo entre imágenes master de preservación y copias derivadas para la consulta electrónica. A su vez, se define el proceso de reconocimiento por OCR, siguiendo ciertas premisas para obtener los productos deseados, y los formatos de salida.

Finalmente, en el instructivo se detalla la forma de grabación de los discos ópticos que serán usados para el almacenamiento de los diversos archivos electrónicos generados en los pasos previos.

El equipo de la biblioteca comenta que aunque utilizan PDF para la distribución electrónica de los textos, convierten los archivos recibidos en ese formato, en la ingesta, a otro de procesador de textos, para tener una versión más editable de esos materiales para el caso de futuras reediciones. Esto genera un intercambio de opiniones entre los distintos participantes, comentando experiencias similares, y se acuerda en la preferencia de mantener versiones editables en algún formato de procesador de textos (Word, OpenOffice, etc.), reservando los PDF para los efectos de la consulta.

Se produce otro intercambio de opiniones sobre esta presentación, analizando cuestiones más detalladas sobre el uso del estandar ISO 9660 para el nombramiento de archivos digitales, así como sobre las nuevas tendencias a digitalizar textos en color y guardarlos en el nuevo formato JPG2000, aprovechando las ventajas de compresión sin pérdida visible que éste presenta. Asimismo, se discuten los puntos débiles que aún existen en torno a este nuevo formato de archivo, tales como la relativa ausencia de herramientas de software que permitan su empleo universal.

La presentación finalizó con los esquemas de metadatos desarrollados por la Biblioteca para la gestión y preservación de sus colecciones digitales. Gracias a un trabajo realmente minucioso, que abarcó la investigación de los esquemas de metadatos existentes y una cuidadosa evaluación sobre su aplicabilidad a las clases de documentos digitales ya definidos, la Biblioteca terminó por crear sus propios esquemas, que fueron presentados en detalle a todos los asistentes. El proceso realizado para definir esos esquemas consistió de los siguientes pasos:

1) Revisión de la estructura y contenido de los documentos a incorporar a la colección
2) Búsqueda y análisis de esquemas de metadatos estándares para cada tipo documental
3) Uso de Dublin Core como esquema estándar de base
4) Generación de esquemas de metadatos propios
5) Necesidad de definir correctamente estos esquemas para que sean legibles por buscadores web *(tarea pendiente)*

Los esquemas de metadatos producidos son siete, uno para cada clase de documentos tal como fueron definidos por la Biblioteca:

1) Planes y Programas
2) Trabajos Presentados en Eventos
3) Tesis y Tesinas
4) Proyectos de Investigación
5) Artículos de Revistas
6) Normativa
7) Convenios

Adicionalmente, el equipo de la Biblioteca hizo un esfuerzo por definir metadatos específicos de preservación, para lo cual revisó los siguientes esquemas en uso:
1) METS (Metadata Encoding Transmission Standard) de la Library of Congress
2) Tech MD (metadatos técnicos para texto recomendados en METS).

Sobre la base de esta revisión, y considerando que no sería factible por ahora cumplir con la masiva producción de datos que estos esquemas requieren, se decidió adoptar algunos de los campos de información presentes en ellos, que fueron considerados como los elementos más necesarios para la preservación de largo plazo.

Por último, presentaron una serie de dudas y cuestiones pendientes en torno a los siguientes elementos:
1) Creación de metadatos para los documentos fuentes
2) Mapeo a Dublin Core - Cosecha de metadatos vía OAI – PMH
3) Esquemas (Schemas) en XML de los metadatos propios

Diapoteca de la FADU
Presentación del caso

Digitalización de una colección de alrededor de 32.000 diapositivas, que comenzó a crearse hacia 1960, por aporte de los profesores de la Facultad, quienes tomaban las imágenes como parte de sus proyectos y como recursos para su actividad docente.

Las actividades de digitalización se inician en el año 2002, para satisfacer varios objetivos:
1) Dar acceso a los profesores a la colección en formato digital.
2) Preservar el material original.
3) Dar acceso por Internet para los alumnos.

A medida que se gana en experiencia, también se piensa incluir la digitalización de una colección de libros antiguos.

En primer lugar, se presentan los parámetros técnicos para la producción de masters digitales de las diapositivas, siguiendo buenas prácticas internacionales:
Profundidad de 24 bits en color
Resolución de captura a 1200 dpi en el scanner de diapositivas.
Captura con cámara digital a 300 dpi.
Guardado del master en formato TIFF sin compresión.

Los masters se almacenan en el disco rígido y se produce una copia de respaldo en soporte óptico.

Producción de las imágenes para consulta:
1) Se derivan copias de uso y miniaturas, con un software por línea de comandos que permite hacer un proceso automático, por lotes, a partir de los masters.
2) Inserción de una marca de agua digital en las copias de uso.

Software utilizado:
Adobe Photoshop, para corrección de contraste y artefactos
Adobe Bridge, para normalizar el nombramiento de los archivos de las imágenes
Digital (DOS), para la generación de las copias de uso (550 pixeles en el lado mayor) y las miniaturas (150 pixeles en el lado mayor).

Muchos de los participantes se interesaron en el software Digital, para entorno DOS, por la posibilidad de derivar automáticamente las imágenes de consulta. Ese software permite reducir el tamaño en píxeles de las imágenes master para adaptarlas a las necesidades de la consulta electrónica. El equipo de esta Biblioteca ofrece compartir esta herramienta, ya que se trata de software libre.

La presentación continuó con las novedades introducidas por el equipo de esta Biblioteca en su flujo de trabajo, que consisten en la adopción del esquema XML de metadatos técnicos NISO/MIX, específico para la preservación de imágenes fotográficas digitales. En la experiencia de este equipo de trabajo, el estándar de preservación MIX se puede cumplir razonablemente bien, ya que:

1) los campos de información obligatorios no son excesivos, y
2) la mayoría de ellos se pueden extraer del header del archivo de imagen, sobretodo cuando fueron generados por cámaras digitales, ya que éstas producen un esquema de metadatos EXIF, utilizable para poblar los campos de MIX.

Se muestran las posibilidades del software Adobe Bridge para la creación de esas plantillas de metadatos, así como el sistema de acceso web desarrollado para esta colección digitalizada.

Ingesta de archivos
Discusión general

En función de las presentaciones anteriores, los asistentes discuten el problema de la ingesta de los archivos “nacidos digitales” en las nuevas colecciones. Particularmente, se ocupan del problema de las tesis electrónicas y artículos académicos. Dado que los formatos electrónicos que se reciben en las bibliotecas son el resultado de la elección de los autores, suele ocurrir que predominen los textos en formato PDF. Aunque este formato es un standard para la distribución de textos electrónicos, desde el punto de vista de la preservación digital de largo plazo, así como de la reutilización del material (reedición), no es el más adecuado y presenta varias limitaciones.

En el intercambio de experiencias sobre la ingesta se plantea un amplio consenso sobre los siguientes puntos:

1) Necesidad de verificar la autenticidad del archivo electrónico (comparar el original en papel con el contenido de la versión digital)

2) El formato PDF presenta muchas dificultades para la edición del contenido, así

como para la migración a formatos como HTML o XML.

3) Cuando el PDF no tiene incrustadas las tipografías utilizadas, suelen haber problemas para la correcta visualización de la información.

4) PDF ofrece muy poca información respecto de con qué tipo de software fue creado el documento. Si se recibe material en PDF es fundamental conocer qué software utilizó el productor para generarlo.

5) PDF es una excelente herramienta de visualización, pero no para almacenamiento o posteriores ediciones y migraciones.

6) La visualización estructurada y jerárquica en secciones, capítulos y hojas, es muy laboriosa y sólo se la obtiene a través de un intenso trabajo manual.

7) Los plugins para pasar de PDF a otros formatos son poco eficientes.

Por todas estas razones, los asistentes acuerdan en que sería preferible recibir los textos electrónicos en formatos más editables, tales como los de los procesadores de textos, antes que en PDF. Se discute también la alternativa de proceder a la migración de formato en la ingesta, como en el caso ejemplificado por el Archivo Nacional de Nueva Zelandia, para lo cual se podría incluso adaptar el software libre NZNL Metadata Extractor, desarrollado por ese Archivo. Una de las instituciones asistentes, que trabaja con XML para la marcación de los textos electrónicos que se pondrán en la web, ofrece su software libre –Pi– para realizar esta marcación. También se comenta que el software Greenstone utiliza XML para guardar los textos electrónicos. Otros recuerdan que la suite de software libre OpenOffice permite abrir cualquier formato de texto electrónico y guardarlo en un formato universal y abierto, basado en XML, manteniendo las características de formateo del texto.

En la discusión queda claro que todos los asistentes son concientes de que, idealmente, los formatos para la ingesta deberían adherir a estándares abiertos y no propietarios, con marcación XML. Sin embargo, se reconoce que no es razonable esperar que los autores presenten textos así generados. En este punto se discuten distintas opciones para normalizar la ingesta, con las siguientes propuestas presentadas:

1) Que el productor/autor entregue todos los archivos electrónicos con los que trabajó en la obra, antes que una presentación “empaquetada”. Por ejemplo, los textos en Word, con las imágenes separadas, en su formato original, y los cuadros de planilla de cálculo aparte, etc. Para algunos asistentes, el formato de Word (.doc) es mas fácil de convertir a HTML para su presentación en la Web.

2) Ofrecer una serie de pautas para los autores en la página web de la biblioteca, para mejorar la creación del PDF, indicando qué cosas se deben activar y cuáles no, para facilitar el trabajo posterior de la institución.

3) Entregar el archivo que produce Latex (software libre del mundo Linux para textos electrónicos, muy usado para crear textos para Internet).

4) Entregar un texto plano (.txt) con las imágenes por separado.

5) Es necesario el apoyo institucional para implementar pautas en la entrega del

material por parte de los productores.

Como resultado de esta discusión, y presentadas las alternativas ya descriptas, queda a cargo de cada institución decidir el empleo de alguna de ellas, bajo la inteligencia de promover así buenas prácticas de preservación digital desde la misma ingesta.

Metadatos
Discusión general

Durante el curso de capacitación se presentaron los nuevos esquemas de metadatos específicamente diseñados para sustentar la preservación de largo plazo de colecciones digitales –METS y PREMIS-. Estos esquemas, a su vez, están en sintonía con el también nuevo marco de referencia OAIS para la creación de repositorios digitales fiables.

Pero estos esquemas, basados en XML, requieren recolectar y registrar una gran cantidad de información, que incluye la de otros arreglos de metadatos en su habitual división tripartita -administrativos, estructurales y técnicos-, más otros elementos específicos para la gestión de largo plazo. Como se lo reconoce en otras instituciones del mundo, mientras no haya herramientas de software que automaticen buena parte del proceso de creación de estos metadatos, su adopción permanece como un desafío difícil de cumplir para cualquier organización promedio que carezca de grandes equipos de catalogadores capaces de realizar ese ingente trabajo en forma manual. De ese modo, la discusión sobre este punto logró consenso en cuanto a utilizar los distintos estándares de metadatos de preservación como una guía para enriquecer el esquema propio de cada institución. Se reconoce la imposibilidad actual de cumplir en profundidad con estos estándares, aunque las instituciones participantes se mantendrán a la espera de novedades tales como nuevas herramientas de software que vuelvan posible la adhesión y aplicación de estos esquemas. Debe destacarse que, en la comunidad internacional de preservación digital, en todas las instituciones medianas y pequeñas hay una activa discusión acerca de cuáles serían los elementos realmente necesarios para producir metadatos de preservación en una forma compatible con los recursos disponibles.

Por otro lado, las instituciones consideran más factible adherir al esquema de metadatos Dublin Core, un estándar internacional más sencillo que normaliza la identificación y descubrimiento de los recursos de información en la Web. Por ejemplo, en un caso se propuso realizar una tabla de equivalencia de los esquemas propios de metadatos con el esquema Dublin Core, para aumentar la visibilidad de los datos en la web y facilitar el descubrimiento de esos recursos. Una de las instituciones participantes trabaja actualmente con el objeto de “mapear” sus metadatos a las etiquetas de Dubin Core, para una mayor accesibilidad en la web. Aunque no alcanza el estatuto de metadatos de preservación, esta solución va por el camino de sumarse a las tendencias internacionales en las bibliotecas digitales desde el punto de vista de la descripción de los recursos de información disponibles. Es interesante remarcar aquí que, debido a su sencillez, el esquema Dublin Core ha sido adoptado ampliamente en el mundo, algo que todavía no ocurre con los exigentes esquemas de metadatos para la preservación digital de largo plazo.

También se presentó el caso de una institución que digitaliza colecciones fotográficas, y ha decidido utilizar el esquema de metadatos NISO-MIX, específico para mantener información técnica crucial sobre las imágenes master fotográficas. Pese a que este esquema también es relativamente exigente en la cantidad de elementos de información que incluye, para cada imagen en particular, el hecho de exista un mínimo de campos obligatorios que pueden capturarse casi automáticamente desde el header de las imágenes ha permitido su adopción y cumplimiento, aún sin contar con un nutrido personal en el proyecto. En general, esta captura casi automática es más sencilla cuando se trabaja con cámaras fotográficas digitales, porque ellas utilizan un estándar de descripción técnica de cada imagen –EXIF, incluido en el header- compartido por toda la industria, que sin mayores esfuerzos puede mapearse al esquema NISO-MIX, con algún agregado manual por software de imágenes –en este caso usaban Photoshop para esta tarea-. Esto se debe a que la industria de la fotografía digital tuvo que encontrar una manera de incluir información técnica sobre las imágenes tomadas para permitir resultados consistentes cuando los usuarios llevan a imprimir las fotos digitales en las casas comerciales de fotografía. En el caso de los escáneres, como su uso difiere del de las cámaras fotográficas –el producto digital no necesariamente tiene por destino su impresión como una foto-, todavía no hay un estándar de descripción técnica similar, aunque mediante el uso de los softwares para edición de imágenes puede accederse a cierto grado de información útil para poblar un esquema del tipo NISO-MIX.

Almacenamiento
Clacso

Más allá de la información recolectada en la encuesta sobre este punto, que mostraba la intención mayoritaria de recurrir a copias redundantes, así como a definir una agenda de copias de refresco en nuevos soportes, entre las instituciones representadas en el Taller se presentó el caso de CLACSO.

La institución utiliza el software libre Greenstone para gestionar y publicar en la Web sus recursos digitales. La estrategia de almacenamiento consiste en exportar a soporte DVD una copia de los nuevos recursos ingestados, más un backup completo en una unidad de disco externo, la que se repite con una frecuencia mensual. En este caso, y por la naturaleza de la misma institución, existe colaboración entre distintos nodos para el almacenamiento.

Fernando Boro
30 de mayo de 2008