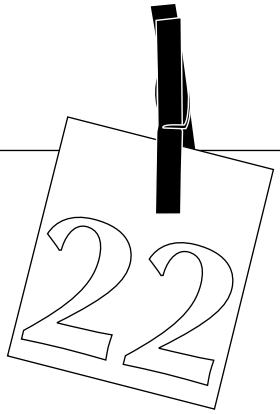


PRESERVACIÓN DIGITAL, EL RETO DEL FUTURO

Luis Torres Freixinet
lfreixinet@gmail.com

Unidad de Sistemas de Reproducción de Documentos del Archivo Municipal de Zaragoza

La realidad supera ampliamente a nuestra imaginación cuando tenemos que enfrentarnos al reto del almacenamiento y preservación de las imágenes.

Compartir archivos: Actas de las VIII Jornadas de Archivos Aragoneses. Huesca, 25-28 de noviembre de 2008. Huesca: Gobierno de Aragón; Diputación Provincial, 2008, t. II, pp. 75-94

RESUMEN

La realidad supera ampliamente a nuestra imaginación cuando tenemos que enfrentarnos al reto del almacenamiento y preservación de las imágenes. Es necesario que se garantice: que sean auténticas, fiables, que estén íntegros y disponibles.

- Dar una primera base que permita comprender la problemática de la preservación de objetos digitales desde el punto de vista técnico.
- Examinar los componentes de la preservación de objetos digitales.
- Conocer las principales estrategias técnicas de preservación digital.
- Examinar las buenas prácticas de diseño de repositorios y servicios digitales.

SUMMARY

The reality surpasses our imagination when we have to face the challenge of storage and preservation of images. It is necessary to ensure: what are authentic, reliable, that are integrated and available.

- Give an initial basis for understanding the problem of preserving digital objects from a technical point of view.
- Examine the components of the preservation of digital objects.
- Knowing the major technical strategies for digital preservation.
- Examine the best practices of design and digital repositories and services.

PALABRAS CLAVE

Preservación digital, OAIS, XENA, HASH, formatos.

KEY WORDS

Digital preservation, OAIS, XENA, HASH, formats.

1. INTRODUCCIÓN

La realidad supera ampliamente a nuestra imaginación cuando tenemos que enfrentarnos al reto del almacenamiento y preservación de las imágenes digitales a largo plazo. Y hablo de reto usando un término "suave" dada la problemática que se nos avecina. Son muchos los sectores afectados por el problema de la preservación de los documentos electrónicos, pero en todos los casos es necesario garantizar:

- Que sean *auténticos*.
- Que sean *fiables*.
- Que estén *íntegros*.
- Que estén *disponibles*.

La problemática y variedad existente hará que nos centremos en este artículo en lo que entendemos todos por imágenes digitales, dejando a un lado todo lo relacionado con bases de datos, hojas de cálculo y/o aplicaciones corporativas. Todo ello a pesar de presentar elementos coincidentes en cuanto a las políticas de seguridad y almacenamiento. Como información podríamos citar las normas que actualmente tratan este tema:

- Trabajos previos del TC 171
- ISO/TR 15801
- ISO/TR 18492
- ISO 19005/A
- Proyecto 26102

Las exigencias de digitalización del Patrimonio Documental difieren en gran medida de las de entornos de oficinas y similares. Además de poder garantizar la mayor fidelidad de la imagen reproducida debemos garantizar que en el futuro éstas sean accesibles. Existe una necesidad imperiosa de trabajar de manera multidisciplinar sabiendo que la tecnología digital nos abre muchas puertas, pero que también nos convierte en esclavos de su propia evolución tecnológica. Evolución que por el ritmo que lleva bien

podríamos denominar revolución. Anticiparse y planificar son las palabras claves, y naturalmente vale la pena, ya que de lo contrario se perdería todo lo realizado hasta ese momento. Estoy de acuerdo con vosotros en que es un problema, pero sabiendo la naturaleza del problema es más fácil intentar solucionarlo. Toda institución que crea ficheros digitales tiene que tener un Plan de Preservación Digital. En lugar de pensar en los ficheros de imágenes como un almacén de datos estacionario, podemos pensar en ellos como en una tribu nómada que puede sobrevivir al paso de los años sólo si sabe moverse en busca de mejores territorios de caza.

Podríamos incluso, como inicio, plantearnos una pregunta: ¿Por qué "*archivar*" los documentos digitales? Aquí las respuestas podrían ser variadas pero fundamentalmente para garantizar la permanencia por su valor informativo, administrativo, legal, cultural (patrimonio), etc.

La Ley 30/1992, de 26 noviembre, de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común, en su artículo 45.5 nos dice: *"Los documentos emitidos, cualquiera que sea su soporte, por medios electrónicos, informáticos o telemáticos por las Administraciones Públicas, o los que éstas emitan como copias de originales almacenados por estos mismos medios, gozarán de la validez y eficacia del documento original siempre que quede garantizada su autenticidad, integridad y conservación y, en su caso, la recepción por el interesado, así como el cumplimiento de las garantías y requisitos exigidos por esta u otras Leyes"*. Este artículo fue posteriormente desarrollado por el Real Decreto 263/1996, de 16 de febrero, por el que se regula la utilización de técnicas electrónicas informáticas y telemáticas por la Administración General del Estado. Y las iniciativas, tanto de la Administración General del Estado como de otras administraciones (CC.AA., entidades locales, etc.) no hacen más que reafirmar la clara tendencia hacia la e-administración.

Ahora bien, ¿cómo definiríamos el archivo de documentos electrónicos? Una buena definición podría ser: "*Conjunto de documentos producidos, recibidos o reunidos por una persona física o jurídica de modo involuntario, natural y espontáneo en el transcurso, y como apoyo, de su actividad, de la que es testimonio, haciendo uso de la electrónica; que se conservan y transmiten también mediante medios electrónicos en depósitos de conservación permanente, tras efectuar una selección a partir de la identificación y valoración de las series, con medidas de autentificación y de preservación adecuadas y con una organización respetuosa con su modo de producción, con el fin de garantizar su valor informativo, legal y cultural, así como de permitir su acceso y uso también mediante las tecnologías de la información*".

Nos encontramos con el problema de que todavía hay pocas soluciones para el tema de la Preservación Digital (en adelante P.D.). Además deberemos tener en cuenta que para garantizar la P.D. deberíamos tener resueltos cuatro problemas:

1. Los aspectos económicos: ¿quién pagará la P.D.? Es un gasto constante, fijo y en ascenso. De manera que tenemos el gasto eléctrico, el del personal informático, el del cambio de hardware obsoleto, etc. Y el riesgo (no cuantificable), puesto que si se produce una pérdida ésta suele ser total y no parcial como ocurría con el papel[1].

2. Aspectos de Responsabilidad o Institucionales: ¿quién asume la P.D.? Informáticos, documentalistas, archiveros, gestión de métodos, etc. La lógica impone que tiene que ser un trabajo multidisciplinar. La dificultad estriba en que este acuerdo se debe mantener a lo largo de los años ¿Cómo podemos atar el futuro?

3. Aspectos Legales: actualmente según la legislación española, europea y americana (USA) casi todas las técnicas de P.D. son ilegales. Los cambios de formatos sobre material protegido (revistas, libros, periódicos, etc.) según la SGAE, y similares, no pueden admitirse, por lo que es necesario una reforma legal. Dos ejemplos:

    - Compramos las fotografías digitales de un fotógrafo actual. Él entrega la obra en un formato determinado. El cambiar el formato de la misma entraría en colisión con lo establecido por la legislación actual.

    - Revista de suscripción on-line. Guardo en mis equipos los tres años que he estado suscrito. Ahora ya no estoy suscrito. ¿Puedo consultar los ficheros guardados? Tendríamos la paradoja de tener los ficheros pero no el derecho legal de usarlos o consultarlos.

4. Aspectos Técnicos: obsolescencia, soluciones o aplicaciones normalizadas…

Curiosamente existe un consenso internacional al indicar el punto 4 como el de más fácil de solución para la próxima década. Otra cosa es que en España se implante… Mientras los puntos 1, 2 y 3 son más fáciles de postular pero por el contrario más difíciles de solucionar.

1. SUDERMAN, Jim. "Soportes y formatos de conservación y transferencia: Hacer elecciones responsables". *Conferencia SARBICA (Hanoi, 5 de mayo de 2004)*. Traducido por Alejandro Delgado Gómez [en línea]. <http://www.interpares.org/>.

2. CUANDO LA TÉCNICA NOS CREA PROBLEMAS…

Lo más importante es el procedimiento, más que el hardware o similares. Por ahora no contamos con ningún sistema que podamos comprar, ya que todo el mundo está realizando pruebas. Dada la complejidad del tema ni siquiera las grandes instituciones han podido crear el sistema completo.

En los últimos tres o cuatro años empezamos a acceder a experiencias de "líderes" internacionales como la Universidad de Cornell[2], NARA, Univ. Oxford, Australia, etc.: observamos estrategias que convergen, pero aún no se ha podido cerrar el círculo.

El tiempo nos ha ido dando lecciones que no debemos olvidar:

1. Es un error muy habitual pensar que las copias "ya las realizará alguien" y así obviar el problema.

2. ¿Qué uso se le da a la información almacenada? Creo que se guarda demasiada información.

3. La obsolescencia tecnológica del hardware y el software no se deben perder de vista. Ejemplo: el proyecto Domesday Book.

4. Los mismos datos pueden ofrecer resultados diferentes al cabo de los años. El reestudio de los datos almacenados con el uso de tecnologías más modernas puede ofrecer nuevas perspectivas en determinados campos.

5. Queda totalmente descartado guardar hardware+software que en el futuro pueda dar soporte a esos datos: los "Museos Tecnológicos" o lo que yo denomino "Almacenes de Chatarra" no son el camino adecuado.

Si tuviéramos que enumerar las etapas de la P.D. indicaríamos lo siguiente:

1. Archivar los documentos: control de copias, etc.

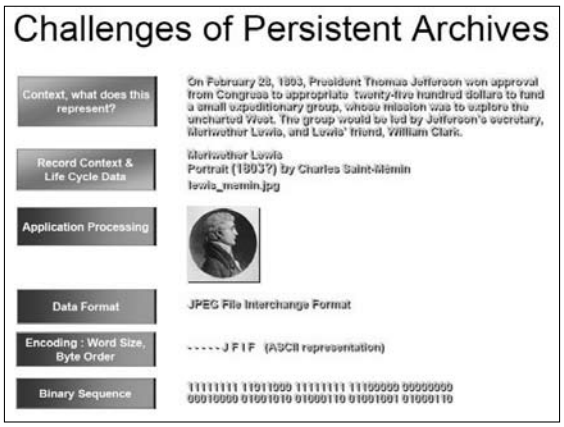
2. Preservar la cadena de bits (el *bit stream*).

3. Garantizar el acceso a largo plazo: nosotros y nuestros recursos pensamos en un plazo de + 10 años. Cualquier solución que no prevea garantizar el acceso debe

2. "10x10: Training Corneéis Digital Preservation Efforts". *The Promise and Perils of Digital Prservation. North Carolina Preservation Consortium Annual Conference (November 18, 2005)* [en línea]. <www.library.cornell.edu/iris/dpo/docs/NC-DP%20at%20Cornell-1105.ppt.

ser descartada. Pero P.D. implica como objetivo la preservación permanente de cualquier fichero o documento electrónico[3].

Pero, ¿qué debemos preservar? Tenemos diferentes niveles, comenzando de abajo a arriba[4]:



- La secuencia binaria.
- La codificación interna (JFIF).
- La codificación externa (formato: jpeg, tiff, etc.). Mantener el conocimiento del formato.
- El resultado una vez procesado: imagen, texto, vídeo, etc.
- Catalogación, registro, metadato… ¿qué es esto? Descripción.
- Contexto histórico de la representación: importante no perderlo de vista.

Deberíamos poder preservar las partes técnicas del documento electrónico. De manera que el *bit stream* (la ristra de ceros y unos) pueda ser rescatada, independientemente del soporte. Pero este *bit stream* debe conservar su formato lógico y la capacidad de poder decodificarlo, por ejemplo, ¿es un texto o una imagen o un vídeo? Si no somos capaces de poder usarlo no hemos solucionado el problema. Pero no hay que olvidar la información intelectual, la información jurídica, la información de gestión, etc. El problema, como uno puede imaginarse, tiende a complicarse de manera exponencial, siendo el tema del vídeo el que por complejidad/volumen se lleva la palma. La informática podríamos decir que funciona por capas o por niveles. Nosotros como usuarios vemos el resultado final.

3. KENNETH THIBODEAU, Ph.D. *Building the Archives of the Future: The National Archives & Records Administration's: Electronic Records Archives* [en línea]. http://www.archives.gTIov/era/pdf/thibodeau-planet-storage-2005.pdf.

4. *Ibid*.

– Los datos tienen un formato (xls, doc, etc.).
  – Es necesario un software (o varios) para acceder a ellos.
– Sólo funcionan en una interfaz determinada.
  – Es necesario hardware.
– A veces los documentos se completan a través de la red, por ejemplo la web.

La magnitud del problema en cifras es realmente alarmante. La Unidad de Sistemas de Reproducción de Documentos del Ayuntamiento de Zaragoza da servicio al Archivo, Biblioteca y Hemeroteca. Actualmente la información digital a conservar supera los 5 Tb (en realidad estaramos hablando de casi 10 Tb.) y la previsión de crecimiento anual son, aproximadamente, 2 Tb. Aunque puede parecer mucho, sobre todo cuando no está establecido correctamente un protocolo de actuación, la situación no es comparable a casos como los de la BBC que prevé invertir 88 millones de euros en 10 años para dar soporte a 500 000 horas de video digital de alta calidad, lo que significará 50 Petabytes (1 Ptb = 1000 Tb = 1 000 000 Gb).

Deberemos ser capaces de desgajar las necesidades de preservación de las necesidades de uso actual, sólo así podremos asegurar la P.D.

Hay que tener en cuenta dos cosas:

1. Material de digitalización retrospectiva, es decir, existe un documento analógico.
   - Poca diversidad de formatos, lo que significa que es un problema para todos y que las soluciones, más tarde o más temprano, llegarán.
   - Unificación a formatos "máster", lo que significa transformaciones e implica riesgo de pérdida/transformación de datos.

2. Material "nacido" digital, sin un respaldo analógico primario. Aquí deberemos imponer de manera obligatoria una unificación hacia el uso de unos formatos identificables, cuantos menos sean, mejor resultado obtendremos. Es más fácil encontrar soluciones para un "archivo muerto" que para uno vivo.

3. SOPORTES DIGITALES

Inicialmente es lo que más nos preocupa (es la parte más visible del problema) pero, como he comentado antes, puede tener una solución más sencilla. Se impone lo que se denomina "refreshing" o cambio de soporte. Esto debe ser realizado de una manera regular teniendo en cuenta la naturaleza de los soportes y las condiciones de conservación de los mismos (manipulación, humedad relativa y temperatura). Como problemas más típicos podemos tener:

- La copia de datos de soportes deteriorados.

- ¿Para que nos servirán esos datos en el futuro si no disponemos del hardware y/o el software necesario?

¿Qué hacemos con los datos? Porque eso es lo que debe preocuparnos, y muchas veces lo perdemos de vista. Hay que dejar claro que una cosa es una copia de seguridad y otra bien distinta es un sistema de preservación digital. Como es fácil de entender, el segundo engloba al primero y no al contrario. Como recomendaciones básicas:

- Un protocolo estricto y documentado de todo el proceso.

- Hay que diferenciar los soportes de uso diario de los soportes de preservación digital.

- Hay que usar material de calidad.

- Control de manipulación (con registro de las mismas), de Temperatura y de Humedad relativa.

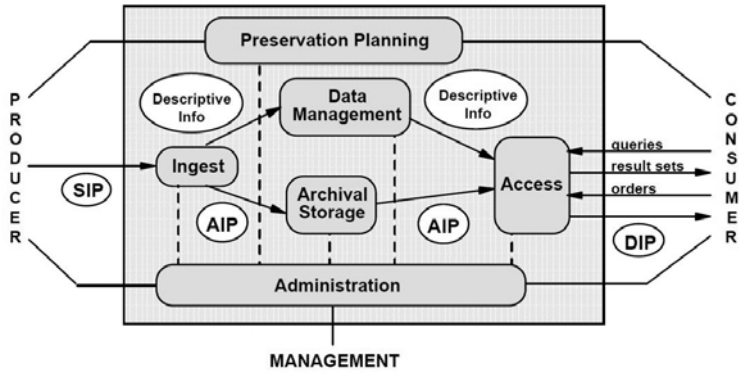
- Control de:
  - CRC= Control de Redundancia Cíclica. Es un tipo de función que recibe un flujo de datos de cualquier longitud como entrada y devuelve un valor de longitud fija como salida.
  - CHECKSUM= Un tipo de Control de Redundancia. El proceso consiste en sumar cada uno de los componentes básicos de un sistema (generalmente cada byte) y almacenar el valor del resultado. Posteriormente se realiza el mismo procedimiento y se compara el resultado con el valor almacenado. Si ambas sumas concuerdan se asume que los datos probablemente no han sido corrompidos.
  - La aplicación práctica es el uso de técnicas criptográficas de comprobación de datos. De manera que usando SHA-1 obtendremos un código alfanúmerico generado a partir de todos los bytes de ese elemento digital. Es en realidad un HASH, es decir, una función o método para generar claves o llaves que representen de manera casi unívoca a un documento, registro, archivo, etc., resumir o identificar un dato a través de la probabilidad. Un **hash** es el resultado de dicha función. Si hay una alteración este código no coincidirá. Eso implica guardar el elemento digital y su código, para futuras comprobaciones. ¿Cuál será nuestro nivel de fiabilidad? Es el momento de empezar a decidirlo.

- Hay que conservar los metadatos, los manuales, controlar las alteraciones bits.

4. OAIS UN MODELO TEÓRICO QUE SE IMPONE

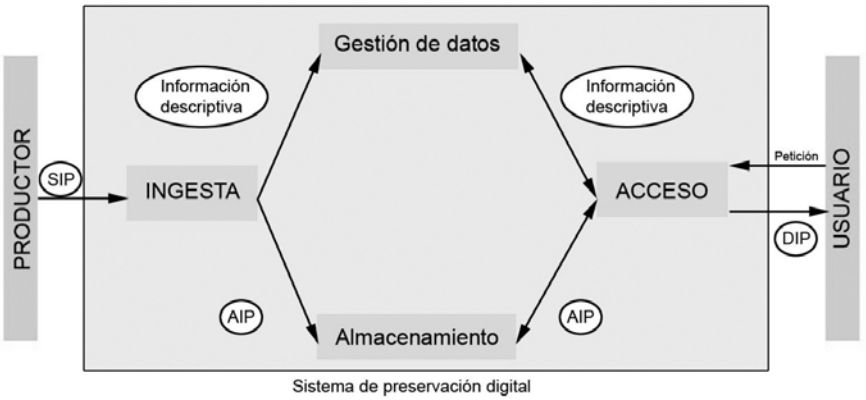
OAIS, o Sistema Abierto de Información en Archivos, no debe confundirnos. No es un software ni un sistema de metadatos. En realidad es un modelo teórico que nos define cómo deberemos organizar un sistema de preservación digital.

Su denominación puede llevar a equívoco en principio, por lo de sistema abierto de archivo, ya que parece relacionarse con la iniciativa de archivos abiertos, pero en realidad lo único en lo que coinciden ambas iniciativas es que tratan de sistemas de archivos porque el sentido de la palabra abierto tiene diferentes significados en uno y en otros. En la iniciativa de OAI se refiere al acceso libre de la información guardada, pero en el modelo OAIS se refiere al carácter no definido del propio sistema de gestión, es decir, su naturaleza flexible y no cerrada para poder adoptarla a las características propias de la organización o de los propios documentos que se van a gestionar. La ISO (International Organization for Standardization), la organización que marca la pauta de cómo se tiene que realizar cualquier proceso en cualquier campo o área para garantizar cierta calidad, encargó al Consultative Committe for Data Space System (CCDSS), otro organismo internacional formado por más de 50 agencias espaciales, la realización de este modelo, y así es como nació. Nos encontramos ante un sistema modular con una serie de elementos imprescindibles o básicos. No es un software que podamos usar para realizar todos los pasos o elementos, sobre todo pensando en grandes instituciones. Veamos el esquema del modelo OAIS[5].



5. COMITÉ FOR SPACE DATA SYSTEMS (CCSDS). *Reference Model for an Open Archival Information System (OAIS). Blue Book*. 2002, January [en línea]. <http://public.ccsds.org/publications/archive/650x0b1.pdf>.

Para simplificar el tema incorporo un esquema simplificado del modelo OAIS[6]:



Sistema de preservación digital

Podemos observar seis bloques destacando la ingesta (la entrada). No debemos olvidar que nos estamos refiriendo a un modelo de preservación digital, por lo que será la entrada de documentos electrónicos al sistema de preservación.

A modo de resumen remarcaríamos que:

- Debe existir un control de entrada: virus, versiones diferentes del mismo modelo, versiones dentro de los diferentes formatos (ejemplo, no es lo mismo un word 6.0 que un word'2007).
- Comprobar que los elementos recibidos son lo que dicen ser (ejemplo, un documento con extensión .pdf que en realidad sea .doc).
- Unificación de formatos.

Del esquema simplificado hay que señalar ciertos términos del modelo anglosajón:

- SIP= Submission Information Package. Sería (resumiendo) el acuerdo de cómo debe ser entregado el documento electrónico (por ejemplo el formato).
- AIP= Archival Information Package. Sería (resumiendo) como se envían los archivos al sistema de almacenamiento o de acceso.
- DIP= Dissemination Information Package. Sería (resumiendo) cómo se realiza la difusión.

6. TÉRMENS GRAELLS, Miquel. *Preservación Digital: Aspectos Técnicos.17-18 de junio de 2008. Asociación Española de Documentación e Información*. Madrid: SEDIC, 2008.

5. EXPERIENCIAS PARA SOLUCIONAR LA INGESTA

De la dificultad de la preservación digital nos puede dar idea el comprobar que la fase más trabajada es la de INGESTA. Todas las demás etapas están en fase beta o interna. La INGESTA es básica para que todo funcione. No es lo mismo ingresar millones de documentos que millones de ficheros. Teniendo en cuenta la tremenda cantidad de formatos existentes hay que saber exactamente qué es lo que nos están entregando. Eso implica dos cosas:

- Conocer en profundidad los diferentes tipos de ficheros existentes y sus variables, siendo las variables lo más confuso por la cantidad de tipologías existentes.
- Adoptar un catálogo cerrado de tipos de ficheros (siendo esto complicado en grandes instituciones).

De todos es conocido que los programas actualmente graban con una extensión por defecto, pero esta extensión puede ser cambiada manualmente (con o sin intención) y además no nos indica esa extensión la tipología o versión dentro de ese formato. Lo importante es cómo está formado internamente. Deberemos *validar* la extensión y *validar* que es un fichero con extensión correcta. Por ejemplo, el sistema debería poder detectar que ese fichero .doc es realmente un Word y además indicarnos qué versión. Sabiendo la versión se puede determinar aplicar políticas para guardarlos o convertirlos. El sistema deberá realizar esas operaciones de manera automática, ya que no estamos hablando de la gestión de unos miles de ficheros, sino de la gestión de millones de ficheros.

En el momento actual casi todos los ficheros funcionan, pero dentro de 30 ó 40 años esto no será así, por lo que ser estrictos se convierte en una ayuda importante para el futuro. Si el fichero está mal (configurado, formado, etc.), ¿para qué guardarlo? Ejemplo, nos pueden enviar un fichero Excel, pero, ¿es un Excel correcto? Deberá ser devuelto a la oficina creadora para que lo envíen según los protocolos establecidos.

Algunas iniciativas internacionales que contemplan el tema de la INGESTA:

- The National Archives (UK) del Reino Unido con el proyecto PRONOM[7]. Este proyecto reúne y hace disponible información técnica acerca de estructuras de formatos de ficheros y productos de software que los soportan. No están todos los formatos del mundo pero sí los más habituales, por ahora los formatos más especializados no están incluidos. Estamos hablando de un catálogo y no de

7. *PRONOM* [en línea]. <http://www.nationalarchives.gov.uk/aboutapps/pronom/>.

una herramienta informática. La información resultante se puede exportar en XML o CSV.

- Library of Congress[8]. Proyecto muy similar al británico de PRONOM: reúne información técnica de los formatos y además explica su experiencia con los mismos. Y al igual que el modelo inglés tampoco cuenta con una herramienta para aprovechar esa información.

- Global Digital Format Registry (GDFR)[9]. Es un proyecto que intenta aunar esfuerzos. Es un registro mundial público en el que incluso participan los fabricantes. Faltan la conexión de los dos modelos anteriores, pero tanto la Library como The Nacional Archives apoyan la iniciativa que surgió inicialmente de la Universidad de Harvard. También apoya este modelo la OCLC (Online Computer Library Center) y la Fundación Andrew W. Mellon. Muchos pensamos que será el modelo elegido por todos en el futuro, con éste o con otro nombre.

- JHOVE (JSTOR/Harvard Objetc Validation Enviroment)[10]. Lo interesante del JHOVE, además de su carácter público y abierto, es que es modular, en código abierto (java) y que encontramos una herramienta informática que realiza tres funciones:

  – Identifica el formato.
  – Valida el formato.
  – Nos da las características del formato. Por ejemplo: gris, grado de compresión, variables usadas a la hora de guardarlo, etc.

  El programa no toma decisiones, sino que nos da un listado XML. Y sobre esto un programa ya puede tomar las decisiones que cada uno estime conveniente.

- DROID (Digital Record Object Identification)[11]. Surge usando las informaciones contenidas en PRONOM. Es una iniciativa anterior a JHOVE y está desarrollada por The National Archives (UK) en JAVA y además es *opensource*.

La preservación digital se organiza siguiendo una estructura o cadena técnica que se refuerza por todo lo visto hasta ahora. De manera que un fichero puede estar formado por varios objetos digitales (por ejemplo si es multipágina) y la representación de

8. LIBRARY OF CONGRESO. *Sustainability of Digital Formats Planning for Library of Congress Collections* [en línea]. <http://www.digitalpreservation.gov/formats/index.shtml>.

9. *GDFR. Global Digital Formats Registry* [en línea]. <http://www.gdfr.info/>.

10. *JHOVE. JSTOR/Harvard object validation enviroment* [en línea]. <http://hul.harvard.edu/jhove/>.

11. *DROID. Digital Record Object Identification* [en línea]. <http://droid.sourceforge.net/>.

esos objetos puede ser variable, por lo que se hace imprescindible entrever el uso de los mismos. Así tendríamos:

- MODS (Metadata Objetc Description Standard)[12], para la descripción intelectual. Son metadatos en XML.
- METS (Metadata Encoding and Transmisión Standard)[13], para la información de empaquetado, la descripción informática, de gestión, derechos, etc.
- PREMIS (Preservation Metadadata Maintenance Activity)[14], para la información técnica de los formatos.

6. UNIENDO LOS ESTÁNDARES

La complejidad del asunto ha conseguido poner de acuerdo a grandes instituciones: el tema es tan complejo que se debe crear una red de responsabilidades compartida. Por ejemplo eso implicaría que si me transfieres los datos estos deben ir correctamente marcados. ¿Cómo veríamos todo lo comentado hasta ahora? Una posibilidad sería la siguiente:

- **Tendríamos un modelo teórico siguiente OAIS**
  - Como lenguaje a la hora de estructurar los contenidos usaríamos XML.
    - El intercambio de contenidos mediante NLM DTD.
    - Para la codificación descriptiva, administrativa y estructural de metadatos usaríamos METS.
      - Diccionario de metadatos: PREMIS
      - Descripción de contenidos: MODS
        - MARC 21
        - Dublín Core

  - Ingesta (entrada) de ficheros.
    - Reconocimiento de formatos: GDFR
    - Validación de formatos: JHOVE, DROID

Una aplicación informática que ya se puede probar es la desarrollada por Nacional Archives de Australia. Se denomina XENA (Xml Electronic Normalising for Archives)[15].

12. *MODS. Metadata Objetc Description Standard* [en línea]. <http://www.loc.gov/standards/mods/>.

13. *METS. Metadata Encoding and Transmisión Standard* [en línea]. <http://www.loc.gov/standards/mets/>.

14. P*REMIS. Preservation Metadadata Maintenance Activity* [en línea]. <http://www.loc.gov/standards/premis/>.

15. *XENA. Xml Electronic Normalising for Archives* [en línea]. <http://xena.sourceforge.net/>.

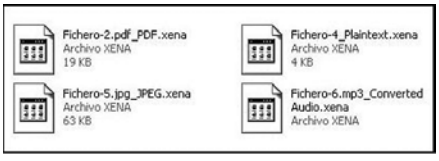
Reconozco que llegados a este momento alguno se ha podido perder dentro de esta "maraña de siglas", por lo que quizá intentando hacer una aproximación práctica se podría arrojar algo de luz al tema. Para ellos vamos a "inventarnos" un ejemplo. Es útil tener visible el esquema simplificado del modelo OAIS que aparece en la comunicación.

- (SIP) El área de urbanismo envía los ficheros al sistema de preservación. Estos ficheros llegarán siguiendo un protocolo establecido, y entre otras cosas cada fichero llevará su código SHA-1 asociado. Este código SHA-1 se obtiene mediante una aplicación informática.

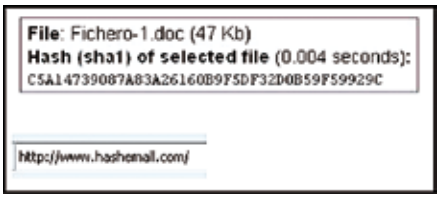
- INGESTA. Ahora tenemos cada fichero con su código SHA-1, por lo que procederemos a archivarlo en una ubicación que todavía no es la definitiva. Ahora en esta ubicación intermedia:
  1. Archivar SIP
  2. Se comprueba que no hay virus.
  3. Se comprueba la integridad del código SHA-1.
     3.1. Se arregla el código SHA-1. Se crea el SIP arreglado, es decir, enlazamos el fichero con el código y entonces procedemos a archivar el SIP arreglado.
  4. Ya tenemos los ficheros con su código HASH y ahora tenemos que validar los formatos con DROID. Del resultado de esta validación podemos tener que:
     4.1. Arreglar extensiones incorrectas, eliminar ficheros que no cumplan el protocolo, etc.
     4.2. Se arregla el código SHA-1. Se crea el SIP arreglado, es decir, enlazamos el fichero con el código y entonces procedemos a archivar el SIP arreglado.
  5. Es en este momento cuando procedemos a normalizar los ficheros con XENA.
  6. Creamos SHA-1 de los formatos normalizados.
  7. Los metadatos son extraídos con Metadata Extraction Tool.
  8. Los metadatos obtenidos se envían a Gestión de Datos.
  9. Se crea AIP para el almacenamiento.
  10. Se envía AIP a almacenamiento.

Ahora procederemos a crear AIP previo al almacenamiento. Recordemos que AIP era como se envían la información al sistema de almacenamiento o de acceso. En este caso tendremos por un lado:

- Los ficheros creados por xena.
  Estos ficheros llevarán extensión .xena



- Y por otro lado el código SHA-1 en XML.



El almacenamiento como tal llevará implícito una serie de pasos:

1. Recibimos de AIP los ficheros tratados.
2. Pasamos a comprobar el SHA-1.
3. Almacenamos los ficheros si el SHA-1 ha sido correcto.
4. Aplicamos técnicas de preservación digital a esos ficheros, que como podemos imaginar pueden ser de diferentes tipologías.
   4.1. Realizamos copia de seguridad en soportes alternativos.
   4.2. Migramos los datos.
   4.3. Actualizamos los metadatos asociados en Gestión de Datos.

El usuario final puede realizar una consulta en el sistema, y desde ese momento el sistema actuaría:

1. Muestra cuál ha sido el resultado de la consulta o búsqueda.
2. Pide al sistema el documento seleccionado.
3. Se recibe el DIP.
4. Se comprueba el SHA-1 almacenado.
5. Se realiza la interpretación de los ficheros seleccionados.
6. Consulta del documento.

El acceso al documento implicaría:

1. Se recibe la consulta del usuario.
2. Se consulta a Gestión de Datos.

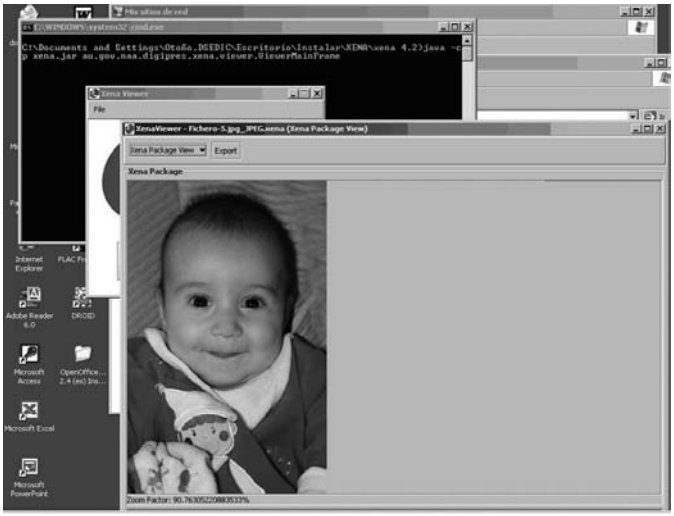
3. Se emite una respuesta que se manda al usuario.
4. Ahora se pide un documento concreto.
5. Esta petición se manda a Almacenamiento.
6. Se recibe el AIP de Almacenamiento.
7. Se convierte en DIP.
8. Y por último se envía al usuario.

Ahora toca crear la AIP de acceso:

1. AIP que estaría compuesta de:
   1.1. Es el fichero .xena.
   1.2. Se genera SHA-1.
   1.3. Empaquetado del AIP.
2. Se envía el conjunto a ACCESO.

Y por último crearíamos la DIP de ACCESO:

1. Se recibe el AIP de Almacenamiento.
   1.1. Comprobación del SHA-1.
   1.2. Desempaquetado.
2. Generamos el DIP.
   2.1. Convertimos el fichero .xena en un fichero de usuario final.
   2.2. Generamos un nuevo SHA-1.
   2.3. Empaquetado.
3. Se envía al consumidor.



Que puede ver la imagen y, si lo desea, grabarla en su equipo en el formato que desee.

Todos estos pasos son realizados por el sistema de manera autónoma y automatizada. La continua generación de HASH (SHA-1) es para garantizar la integridad de la información transmitida y poder comprobarlo en todo momento.

No quisiera terminar sin citar los modelos técnicos de preservación:

- *Refreshing* (recopia): cambio de soporte físico, por ejemplo de CD a DVD.
- Migración: cambio de formato técnico, por ejemplo pasar de Worperfect a OpenOffice.
- Emulación: recreación del entorno de software y hardware, por ejemplo la emulación para los juegos de los AMIGA en los PC actuales.

Aún así hay que aclarar que cualquier migración implica un riesgo de pérdida de información o de funcionalidades.

7. CONCLUSIÓN

Mi idea original era realizar un planteamiento de difusión del estado actual del tema de la preservación digital, pero con un interés fundamental: es necesario que en nuestra Comunidad Autónoma las administraciones implicadas en el patrimonio digital unan sus fuerzas y pongan en marcha lo antes posible un plan de preservación digital, plan que por su diseño y estrategia pueda servir para todos y en el que todos puedan dar su aportación.

Pensar que ya nos caerá la solución encima y quedarnos quietos mientras tanto, es estar abocados a una situación de riesgo que personalmente creo no deberíamos tolerar. Sólo con un planteamiento multidisciplinar de especialidades y de instituciones (no sólo del ámbito público) se puede llegar a conseguir resultados visibles, abandonando los modelos teóricos.

Afortunadamente es el planteamiento que surge más allá de nuestras fronteras, lo que se comprueba al ver como muchos de los proyectos más adelantados son *opensource* y de código abierto.

La preservación absoluta quizá es imposible de garantizar o económicamente es inviable, pero el inmovilismo no es la mejor solución. ¿Afrontaremos el futuro con garantía?

BIBLIOGRAFÍA

- CASTILLO, José Manuel; JORBA, Ferran: "Almacenamiento distribuido y preservación digital". En *Els dipòsits d'e-informació*. Barcelona: CESCA, 2007.
  En línea: <http://www.cesca.es/promocio/congressos/tsiuc2007/FerranJorba.pdf>
- McGOVERN, Nancy. "Aligning Digital Preservation Policies with Community Standards". *ICPSR. Inter-University consortium for political and social research*. 2007 [en línea]. <http://ipres.las.ac.cn/pdf/Nancy%20McGovern%20ipres2007-DPpolicies-mcgovern.pdf>.
- FANNING, Betsy A. "Preserving the data explosion: Using PDF". *Digital Preservation Coalition, 2008* [en línea]. <http://www.dpconline.org/docs/reports/dpctw08-02.pdf>.– GALLOWAY, Patricia. "Preservation of digital objects". *Annual review of information science and technology*. 2004, V. 38, pp. 549-590.
- ROG, Judith. *PDF Guidelines. Recomendations for creation of PDF files for long-term preservation and access*. Koninklijke Bibliotheek, 2007.
  En línea: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/PDF_Guidelines.pdf.
- SALGADO, Cecilia. "Permanencia en CD-R (discos compactos grabables)". *Laboratorio Mexicano de Imágenes. Revista*. 2005 [en línea]. http://www.lmi.com.mx/revista/conservacion/16.html.
- VILMONT, León-Bavi. "Effet des polluants atmosphériques sur les disques compacts". C.R.C.C (Centre de Recherche sur la Conservation des colletions), France, 2000.