III Conferecia Bibliotecas y Repositorios Digitales de América Latina (BIREDIAL '13)
VIII Simposio Internacional de Bibliotecas Digitales (SIBD '13)
"ACCESO ABIERTO, PRESERVACIÓN DIGITAL Y DATOS CIENTÍFICOS"
Ciudad de la Investigación, Universidad de Costa Rica, del 15 al 17 de octubre de 2013.
Ponencia

Preservación digital para instituciones de tamaño medio: una adaptación pendiente

Miquel Térmens Graells
Departamento de Biblioteconomía y Documentación. Universidad de Barcelona. España
termens@ub.edu

Palabras clave: preservación digital; OAIS; auditoria; análisis forense; PREMIS; emulación

1. Introducción

Disponer de soluciones que permitan la preservación a largo plazo de los objetos digitales se ha convertido en una necesidad ineludible en el mundo actual. La preservación digital inicialmente se nutre de las técnicas de conservación de los materiales tradicionales y de las técnicas de seguridad informática, pero su marco de trabajo es totalmente nuevo y exige la aplicación de técnicas nuevas, por ello ha sido necesario que se dotase de referencias teóricas a partir de las cuales poder desarrollar las soluciones adecuadas. Este marco teórico apareció el año 2003, plasmado en el documento *Reference Model for an Open Archival Information System* (OAIS), creado por la NASA y refrendado por el Consultative Committee for Space Data Systems (CCSDS). El modelo OAIS definía que elementos funcionales habían de constituir un sistema de preservación digital y se convirtió en norma ISO el año siguiente con el código 14721:2003; posteriormente, en junio de 2012, el CCSDS publicó una revisión que no incorporó cambios substanciales (Reference, 2012). Desde su aparición el modelo OAIS ha tenido una gran aceptación a nivel mundial, como lo demuestra el hecho de que todos los sistemas de preservación actualmente en funcionamiento o en proyecto aseguran que cumplen con este modelo con mayor o menor fidelidad.

Ahora bien, la mayor parte de las experiencias publicadas de preservación digital se refieren a casos de éxito en grandes instituciones –universidades punteras, archivos y bibliotecas nacionales…– o en grandes empresas multinacionales o de servicios informáticos. Los manuales y las principales publicaciones de referencia en la disciplina también tienen este sesgo y a una persona no introducida le dan la impresión de que la preservación digital tiene una gran complejidad técnica, organizativa y económica y que su aplicación sólo está al alcance de grandes corporaciones.

Esta comunicación considera que ésta es una visión equivocada que debe ser rectificada y que por ello es necesario presentar soluciones y metodologías de aplicación factible a pequeña y mediana escala. Se presenta un esquema de actuación para elaborar planes de preservación digital adaptados a instituciones de tamaño mediano. Su aplicación debe hacer posible la construcción de sistemas de preservación digital sostenibles a nivel técnico y económico en este tipo de instituciones, con lo que se facilitará la preservación de datos digitales en un mayor número de situaciones.

2. Metodología

Las propuestas aportadas se han basado en un análisis de la bibliografía disponible, en la experiencia como consultor de implantación de proyectos de preservación digital y en los resultados de distintos proyectos de investigación dirigidos por el autor desde la Universidad de Barcelona en los ámbitos de biblioteca universitaria, repositorios institucionales, hospitales y gestión de documentación cartográfica. Aunque estas experiencias no cubren la totalidad de sectores económicos o de actividad, presentan la suficiente variedad como para destacar rasgos y soluciones comunes.

3. Desarrollo

Los planes de preservación digital adaptados a instituciones de tamaño mediano deberían basarse en los siguientes apartados.

3.1. Utilización de las auditorías de seguridad informática y de preservación digital como metodologías para la autoevaluación de la situación y de las perspectivas de mejora de una institución.

La realización de auditorías es una práctica habitual en los sistemas informáticos, con ellas se puede evaluar de forma objetiva y externa que estos sistemas son gestionados de forma correcta. Este mismo objetivo lleva a demandar que las auditorías también se deberían aplicar de forma obligatoria a los sistemas de preservación digital, dada su dependencia de los componentes informáticos.

Los actuales sistemas de gestión de la seguridad parten de las recomendaciones de la norma británica BS 7799, después reconvertida en la familia de normas internacionales ISO 27000, auditables por un equipo externo. El problema de estas metodologías es que no contemplan los problemas específicos de los sistemas de preservación, como es la problemática tecnológica a largo plazo, la dificultad para asegurar el compromiso institucional o los problemas de financiación. Por ello han aparecido nuevas metodologías de auditoría especializadas en el ámbito de la preservación digital, entre las que destacan: Drambora en el Reino Unido, Nestor 2 en Alemania y TRAC en los Estados Unidos, esta última reconocida como norma ISO/IEC 16363:2012 (Audit, 2011).

Las grandes organizaciones ya han empezado a utilizar los anteriores sistemas de auditoría, en buena medida para justificar sus actuaciones ante sus *skateholders*. En el caso de organizaciones medianas esta rendición de cuentas muchas veces no es lo más importante pues su sistema de preservación aún se encuentras en un estadio incipiente, si es que disponen del mismo. En este caso, las auditorías también tienen un importante papel: el de mecanismo para comprobar el nivel de preparación de la organización para afrontar esta tarea. Con esta orientación hemos conducido diversas auditorías en España:

- 9 Instituciones hospitalarias de la región de Cataluña (España), para evaluar su capacidad para preservar sus archivos de historias clínicas electrónicas (Boté, Térmens, 2011).
- Instituto Cartográfico de Cataluña (Barcelona, España), para evaluar su capacidad para preservar los datos geográficos digitales (geodata) que produce (Térmens, Locher, 2012).
- Repositorio institucional de la Universidad Autónoma de Barcelona (UAB). El Dipòsit Digital de Documents (DDD) de la UAB es el primero de España y el 16º del mundo en el ranking de repositorios (fuente: Ranking web de repositorios, 2013. http://repositories.webometrics.info/es), para evaluar la seguridad de su gestión informática (Térmens, Casaldáliga, Azorín, 2013).

En estas auditorías se han usado diversos instrumentos: un sistema de auditoría de seguridad informática propio de la administración pública española –el Esquema Nacional de Seguridad (ENS)– y los sistemas de auditoría de preservación digital –Drambora y TDR ISO/IEC 16363:2012–. Aunque la experiencia aún no está cerrada pues se está a la espera de realizar nuevas auditorías, ya se pueden extraer algunos resultados:

- A nivel metodológico: es más fácil realizar auditorías sin valor de certificación pues así se rebaja la trascendencia de los resultados, como consecuencia es posible conseguir la participación de todos los agentes implicados (gestores de la documentación, informáticos…).
- A nivel de resultados: se comprueba que a menudo las organizaciones aplican soluciones que externamente se pueden considerar heterodoxas o parciales según las prácticas más comunes, pero que sin embargo dan resultados deseados.
- A nivel de planificación: las auditorías sirven a las organizaciones para tomar conciencia de su situación actual y para marcarse una hoja de ruta de las mejoras que deben implementar para desarrollar un sistema de preservación eficaz.

Las auditorías realizadas han sido un instrumento eficaz y contrastable para comparar la situación de una institución determinada con los requerimientos de preservación que emanan del modelo OAIS y de las buenas prácticas en gestión informática. Así ha resultado fácil comprobar como procedimientos técnicos particulares cumplen perfectamente con los requisitos de preservación digital.

3.2. Adaptación de las especificaciones OAIS a las peculiaridades de estos casos.

OAIS es el actual marco teórico que rige el diseño de sistemas de preservación digital. Entre sus ventajas destacan su carácter abierto, adaptable a cualquier entorno, así como su capacidad para crear flujos de trabajo seguros incluso en entornos complejos. Pero estas mismas cualidades pueden significar un freno para su implantación: la lectura de la recomendación OAIS es difícil para personas no especializadas y provoca una sensación inicial de sistema extremadamente complejo, sólo aplicable en grandes empresas e instituciones; la misma observación de las gráficas que acompañan el documento no hacen más que ahondar en la sensación de complejidad. Esta es ciertamente una visión incorrecta, pues el mismo documento OAIS indica que su estructura debe ser adaptada a cada caso concreto de aplicación.

Las instituciones pequeñas y medianas cuentan con recursos limitados, pero también normalmente con datos de una complejidad limitada o, para ser más precisos, con datos más homogéneos, menos diversos, que en una institución grande como puede ser un archivo o una biblioteca nacional. No son pocos los casos en los que el productor de los datos y el responsable de su preservación a largo plazo son la misma entidad; si es así toda la complejidad del módulo de ingesta de un sistema OAIS puede ser reducida en gran medida, por ejemplo evitando controles sobre la procedencia, la autenticidad o incluso sobre el formato de los ficheros. Estas adaptaciones no se han de entender como una reducción del nivel de rigurosidad en la preservación de los datos almacenados, por el contrario son el resultado de una mejor comprensión de las necesidades de gestión que ellos requieren.

Un buen ejemplo lo tenemos en las instituciones sanitarias, uno de los sectores con una necesidad más clara de preservación de información a largo plazo dentro de un entorno altamente regulado a nivel nacional e internacional; la conservación de las historias clínicas electrónicas de los pacientes es de obligado cumplimiento en todos los países siguiendo un esquema establecido de datos –HL7– así como un formato técnico –DICOM–. Estamos hablando de datos de gran trascendencia para la salud de las personas, que incluyen informaciones muy sensibles (enfermedades, predisposición genética, antecedentes de filiación, mapa genómico…) sometidas a la legislación de protección de datos personales. Además se trata de bases de datos de un gran volumen y con un crecimiento constante. A pesar de estos antecedentes externos, un análisis (Boté, 2012) de este caso respecto a las previsiones del modelo OAIS de preservación ha demostrado que el esquema funcional a aplicar para la preservación de las historias clínicas es relativamente sencillo y no tiene nada que ver con la complejidad que dan a entender algunos gráficos OAIS; los módulos de ingesta, gestión de datos y de consulta, por ejemplo, quedan reducidos a la mínima expresión, a pesar de tratarse de los más complejos dentro de un modelo OAIS. Así, la preservación de las historias clínicas no se aleja mucho de ser simplemente un problema de seguridad informática y de disponer de la capacidad financiera para sostener la infraestructura de almacenamiento; en definitiva, organización y dinero.

Si esta simplificación es posible en los archivos de hospitales, nada impide que también se pueda aplicar a otros sectores con datos menos críticos.

3.3. Adaptación de las técnicas del análisis forense digital para agilizar la ingestión de grandes volúmenes de datos.

Las técnicas de análisis forense digital, inicialmente circunscritas al ámbito de la investigación criminalística, ya se están aplicando como técnicas centrales de la arqueología digital, es decir, para la recuperación de documentos conservados en soportes obsoletos (Kirschenbaum, Ovenden, Redwine, 2010; Lee, Kirschenbaum, Chassanoff, Olsen, Woods, 2012).

Una nueva visión del análisis forense (John, 2012) muestra que se están abriendo nuevos ámbitos para su aplicación. Uno de éstos es su utilización para la ingestión normalizada de grandes volúmenes de datos, en especial aquellos procedentes de proyectos de investigación, siguiendo las prescripciones de los planes de gestión de datos (*data management plans*) que se están empezando a aplicar en determinados países. Siguiendo esta línea, aún es posible concebir un uso más extremo de las técnicas de análisis forense, en este caso para facilitar la ingestión de grandes volúmenes de datos de procedencia externa. La curación de datos digitales a escala individual es la más precisa y flexible, pero también la más costosa. En instituciones sin grandes capacidades técnicas y económicas quizás sea más conveniente que determinados datos sean preservados en bloque y no a nivel individual, mediante el empaquetado de los objetos individuales y sus correspondientes metadatos; estas agrupaciones sin duda requerirán una menor intervención humana y facilitaran su conservación.

*3.4. Aprovechamiento de los servicios en la nube (*cloud storage*) para dar seguridad al almacenamiento.*

La replicación de las copias de seguridad o incluso de sistemas de preservación enteros es una de las bases ineludibles de la seguridad de los datos. Que estas copias se encuentren en ubicaciones separadas e incluso alejadas entre sí no es un problema para grandes corporaciones, que disponen de distintos edificios o de instalaciones en distintas poblaciones. Pero para instituciones pequeñas o medianas este es un requerimiento difícil de cumplir. Afortunadamente la actual expansión de los sistemas basados en la nube pueden aportar una alternativa.

Normalmente se piensa en las soluciones *cloud* para situaciones con grandes necesidades de computación, de escalabilidad de sistemas o grandes necesidades de almacenamiento. Para instituciones pequeñas y medianas el *cloud storage* puede ser una alternativa viable técnicamente y barata para aumentar la seguridad de los datos almacenados, consiguiendo disponer con facilidad de una segunda copia remota. Un ejemplo lo son los productos Amazon S3 y, aún más, Amazon Glacier, que ofrecen un gran nivel de seguridad de los datos bajo la forma de una solución SaaS, en la que la seguridad deja de ser una inversión costosa inicialmente y se convierte en un servicio que se paga mensualmente.

Es cierto que estas soluciones en la nube aún no están plenamente adaptadas para realizar las funciones de preservación, pues exigen que a partir de sus API se desarrollen los servicios que permitan la subida y la monitorización de los ficheros almacenados. Pero asimismo es cierto que estos scripts se pueden programar sin excesivo esfuerzo y que con un buen análisis de la problemática se puede reducir la necesidad de interacciones input/output, siempre costosas en un entorno *cloud*.

3.5. Revisión de la aplicación dada al esquema de metadatos PREMIS.

Los metadatos PREMIS (Premis, 2012) han conseguido categorizar y normalizar el registro de las principales propiedades y eventos que atañen a la vida de un objeto digital, pero la aplicación de este esquema de metadatos es compleja, pues requiere desarrollar procesos que lo alimenten con los datos necesarios. En algunas situaciones, que PREMIS esté soportado bajo un esquema XML no ayuda a facilitar su implementación o su integración con otras aplicaciones ya existentes.

El resultado es que PREMIS está siendo menos usado de lo necesario, a pesar de que las instituciones reconocen la bondad de sus objetivos. Por ello es necesario encontrar vías alternativas que permitan registrar los mismos datos que PREMIS pero con un esfuerzo menor. Hay que volver a los inicios y preguntarse ¿qué es lo importante de PREMIS? ¿para qué fue creado? La respuesta es clara: para documentar los eventos que puedan tener alguna trascendencia para la curación a largo plazo de los objetos digitales. PREMIS normaliza esta documentación y prevé ya un número importante de estos eventos a documentar, pero no entra en consideraciones de si todo su esquema es de aplicación en un entorno concreto, esta determinación deben realizarla los responsables técnicos de este entorno. La realidad nos muestra que en un gran número de situaciones, quizás la mayoría, el número de eventos a documentar será muy pequeño debido a un conjunto de razones:

- tenemos una dispersión baja en el número de los formatos a preservar, de los que conocemos con bastante exactitud su comportamiento técnico a medio plazo;
- la ingestión está muy controlada pues los objetos proceden en su mayoría de un entorno de gestión propio de la misma organización que preserva;
- el archivo de preservación no tiene un volumen o un mandato legal que le obliguen a mantener una vigilancia tecnológica para avanzarse a las posibles incidencias de los formatos almacenados, en especial a sus necesidades de migración o de obtención de nuevos visores para su consulta, por el contrario, se trata de archivos que simplemente seguirán las buenas prácticas que observarán en las instituciones de preservación de referencia en su área, como pueden ser archivos y bibliotecas nacionales.

En estas circunstancias lo más adecuado es tomar PREMIS como una guía de actuación, a adaptar, y no como un esquema normativo de obligado cumplimiento. Un ejemplo de este enfoque lo tenemos en el sistema de preservación COFRE, desarrollado en la Biblioteca de Catalunya, de Barcelona, España (Pérez, Serra, 2010). El equipo de esta biblioteca nacional llegó a la conclusión que estaba fuera de su capacidad económica desarrollar un sistema de preservación digital con todos los

parámetros que formalmente se podrían esperar, pero que por el contrario podían ser capaces de crear otro sistema más simple pero igualmente robusto. Una de los puntos clave de este sistema fue la sustitución de una base de datos XML PREMIS por otra con tecnología MySQL y con un número reducido de campos, los estrictamente necesarios para la gestión eficaz de los objetos almacenados.

Soluciones como la descrita permiten adelgazar los requerimientos técnicos para el registro de los metadatos de preservación y reducir el número de metadatos que se han de introducir directamente o importar de otras fuentes.

Otra estrategia consiste en registrar los eventos a nivel de series de objetos y no a nivel de objetos individuales. Hemos de recordar que estos objetos muchas veces han sido generados de forma automática y según el mismo procedimiento técnico, es el caso de todos los procesos de digitalización masiva en archivos y bibliotecas. Luego estos mismos objetos también suelen ser ingestados en lotes en el sistema de preservación. Ante esta situación lo más sensato también es registrar sus metadatos de preservación de forma conjunta, o sea, en un solo registro, y no como un apunte individual para cada objeto. Este cambio de escala en el nivel donde se registran los metadatos provoca una disminución dramática en los metadatos a conservar y en los medios a disponer para llevarlo a cabo.

3.6. Valoración de las soluciones de virtualización como solución aplicable para el acceso a datos obsoletos.

La migración de ficheros de un formato obsoleto a otro más actual es una de las principales técnicas de preservación digital pero también una sobre las que se presentan mayores objeciones respecto a su sostenibilidad a largo plazo. Entre otros problemas, se alerta sobre el alto coste de realización de las migraciones y como la realización de sucesivos procesos de migración puede dar como resultado la acumulación de errores en los ficheros finalmente conservados. Como alternativa se propugna la preservación de los formatos originales y, posteriormente, facilitar el acceso a los mismos mediante la emulación de su entorno informático original (Rothenberg, 1995).

Aunque la emulación es una solución técnica en la que están trabajando diversos proyectos de investigación, aún no disponemos de demasiados ejemplos de su implementación real aplicada a la preservación digital. En cambio la utilización de máquinas virtuales ha tenido una gran expansión en los centros de datos de gestión, facilitando la escalabilidad de los sistemas, el mejor aprovechamiento del hardware disponible y la replicación de servidores. Sin la virtualización actualmente no serían posibles muchos de los servicios disponibles en la nube, en especial soluciones de *cloud computing* como AWS EC2.

Este éxito de la virtualización permite aventurar que en el futuro su variante la emulación será fácilmente aplicable a la preservación digital. A ello hay que añadir una rebaja en las necesidades de migración de formatos que hasta ahora se habían predicho. Lo podemos comprobar observando la disponibilidad aún de formatos

teóricamente ya obsoletos, como Lotus 123, WordPerfect DOS o dBase III. El resultado conjunto de estos dos fenómenos es una rebaja en el peligro de llegar a almacenar formatos obsoletos y, como resultado, una rebaja en las necesidades de tener activadas las alertas tecnológicas. Posiblemente en muchos casos el mayor peligro no esté en la no visualización de los datos, si no en la no conservación de materiales auxiliares necesarios para su correcta utilización o interpretación, como pueden ser manuales de uso o, aún más, listados de códigos o claves usadas para codificar datos.

4. Conclusiones

La aplicación de soluciones técnicas y metodológicas como las que aquí se han presentado convierte en viable la ejecución de planes de preservación en instituciones medianas. Estas soluciones no invalidan el marco teórico aportado por el modelo OAIS o su aplicación en grandes instituciones, al contrario, hacen que sea posible su adaptación a otros ámbitos.

Sin duda será necesario que las instituciones, antes de aplicar las soluciones planteadas, analicen las consecuencias de las distintas alternativas y sean plenamente conscientes de las renuncias que toda simplificación conlleva.

5. Discusión

Es necesario rebajar las tensiones técnicas que se dan en las instituciones medianas al no poder aplicar sistemas tecnológicamente avanzados de preservación digital como los que parece propugnar una visión simplista del modelo OAIS. El problema reside en que estas tensiones que a menudo llevan a la falta de acción pues en estos momentos las dificultades técnicas se están utilizando como perfecta excusa para no actuar, para no decidir; esto es la indican frases del tipo: “la preservación digital está más allá de nuestras posibilidades técnicas y económicas”. Si por el contrario se aplica el sentido común en el análisis del problema, como ha intentado demostrar esta comunicación, será posible encontrar soluciones prácticas para actuar de forma decidida en este reto ineludible que es la preservación digital. Una vez rebajados los requerimientos técnicos, pronto se verá que el foco se habrá de poner en la toma de decisiones, más concretamente en el compromiso institucional, que es el que llevar a la acción.

En las pocas páginas disponibles en esta comunicación solo ha sido posible presentar de forma somera las principales líneas de acción que sustentan nuestra visión de que es posible y necesario un nuevo enfoque de la preservación digital basado en soluciones realistas y pragmáticas. Este planteamiento debe ser trasladado a casos concretos para poder evaluar su eficacia y descubrir cuáles pueden ser sus limitaciones. Por ello es necesario que nuevos estudios analicen la problemática de la preservación digital en escenarios distintos y permitan la comparación de sus resultados.

6. Agradecimientos

Esta investigación se ha realizado en el marco del proyecto *El acceso abierto (open access) a la ciencia en España: análisis del grado de implantación y de la sostenibilidad de un nuevo modelo de comunicación científica*. 2012-2014. Plan Nacional I+D+i, código CSO2011-29503-C02-01. También ha contado con el soporte del Grupo de Investigación "Cultura y contenidos digitales: aspectos documentales, políticos y económicos".

7. Bibliografía

Audit and certification of trustworthy digital repositories. ISO 16363:2012. Washington: Consultative Committee for Space Data Systems, 2011. http://public.ccsds.org/publications/archive/652x0m1.pdf

Boté Vericad, Juan-José: *Propuesta de un modelo de preservación digital para pequeñas y medianas instituciones sanitarias*. Barcelona, Universitat de Barcelona, 2012. 353 p. http://hdl.handle.net/10803/96254

Boté, Juanjo; Térmens, Miquel: "Trac y ENS en la Auditoría de Preservación Digital de los Archivos Sanitarios". *XIV Congreso Nacional de Informática de Salud. Inforsalud 2011*. Madrid, Sociedad Española de Informática de la Salut, 2011. p. 155-159.

John, Jeremy Leighton: *Digital Forensics and Preservation*. Digital Preservation Coalition. Technology Watch Report, 2012. http://www.dpconline.org/component/docman/doc_download/810-dpctw12-03pdf

Kirschenbaum, Matthew G.; Ovenden, Richard; Redwine, Gabriela: *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources, 2010. http://www.clir.org/pubs/reports/pub149/pub149.pdf

Lee, C.A., Kirschenbaum, M., Chassanoff, A., Olsen, P., & Woods, K.: "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions". *D-Lib Magazine*. 18(5/6), 2012.

Pérez, Karibel; Serra, Eugènia: *Repositori de preservació digital de la Biblioteca de Catalunya. Informe descriptiu i de situació*. Barcelona, Biblioteca de Catalunya, 2010. 24 p. http://hdl.handle.net/2072/97251

PREMIS data dictionary for preservation metadata. Version 2.2. 2012. http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

Reference Model for an Open Archival Information System (OASIS). Recommended

Practice. Magenta Book. Washington DC: Consultative Committee for Space Data Systems (CCSDS), June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

Rothenberg, Jeff, Ensuring the Longevity of Digital Documents. *Scientific American*. 272 (1) (1995).

Térmens, Miquel; Casaldàliga, Núria; Azorín, Cristina: “Evaluación con el Esquema Nacional de Seguridad (ENS): la aplicación en el repositorio institucional de la UAB”. *XIII Jornadas Españolas de Documentación. Fesabid 2013*. Toledo, Fesabid, 2013. p. 176-179.

Termens, Miquel; Locher, Anita E.: "Digital preservation audit on spatial data: a practical experience". *SOMAP 2012: service-oriented mapping*. Wien: Jobstmedia, 2012