EL CONTROL DE LOS FORMATOS EN LA PRESERVACIÓN DIGITAL

Térmens, Miquel
Ribera, Mireia
Universitat de Barcelona. Facultad de Biblioteconomía y Documentación
termens@ub.edu
ribera@ub.edu

Resumen: La preservación de la documentación digital requiere a nivel técnico seguir de forma estricta toda una serie de procedimientos y técnicas. Actualmente el modelo teórico más seguido en preservación digital es OAIS (Reference Model for an Open Archival Information System) que, entre otros, marca los procedimientos de entrada (ingesta) de los ficheros en el sistema. Uno de los aspectos claves de la ingesta es la identificación de los formatos de los ficheros y su validación. En la comunicación se describen las distintas técnicas existentes y la emergencia de los registros de formatos orientados a la preservación, como Pronom y GDFR (Global Digital Format Registry). Por último se presentan las principales herramientas software que se están usando para la comprobación de los formatos de los ficheros DROID, JHOVE, XENA y TrID.

Palabras clave: preservación digital; formatos de ficheros

1. Introducción: preservar ficheros para preservar documentos

El mundo actual tiene la información digital entre una de sus bases más características. La información se consume en forma de series de datos almacenados en bases de datos o bajo la forma de documentos digitales también almacenados en medios informáticos. Una parte de estos documentos acabará siendo conservada a largo plazo para facilitar su reutilización, como testimonio histórico de una época o como registro de unos hechos. La preservación digital se presenta como una tarea importantísima para las empresas y las administraciones públicas, destacando el papel de las instituciones de la memoria: bibliotecas, archivos y museos.

La preservación digital de los documentos y los datos es un objetivo aún no enteramente resuelto en sus procedimientos y técnicas, pero en el que distintas instituciones y empresas están trabajando a nivel mundial y en el que ya se observan resultados plenamente operativos (Keefer, Gallart, 2007).

Esta comunicación analiza una de las primeras problemáticas técnicas –ya parcialmente resuelta– que se deben abordar en cualquier sistema de preservación: la correcta identificación de los formatos de los ficheros a ingresar, pues sólo conociendo perfectamente los objetos de preservación se les podrán aplicar los procedimientos técnicos que aseguren su conservación y validez a largo plazo.

Antes de bajar al nivel de las distintas soluciones técnicas, ha sido necesario llegar a un cierto consenso en la configuración de un esquema teórico que defina y enmarque las principales funciones a las que ha de dar respuesta un sistema pleno de preservación. El esquema que ha recogido más adhesiones y en el que se basan la mayoría de implementaciones actuales es OAIS (Reference Model for an Open Archival Information System), desarrollado por la NASA y publicado en enero de 2002 como estándar CCSDS 650.0-B-1 y aprobado el año siguiente como norma ISO 14721:2003 (Reference, 2002). OAIS considera que un archivo digital se ha de componer de 6 servicios o procesos básicos: adquisición o ingreso (ingesta), conservación o almacenamiento, planificación de la preservación, acceso, gestión de la información y administración. Los documentos entran en el archivo por medio del módulo de ingesta, que comprende todos los procedimientos de control de entrada de los ficheros que se incorporan al sistema y que, entre otras cosas, ha de verificar su procedencia, su integridad, su autenticidad, sus características técnicas y que esté libre de virus. En este estadio es imprescindible identificar de forma segura cuál es el formato técnico de los ficheros; ello es necesario por distintos motivos:

- Entre los expertos existe el consenso de que los sistemas de preservación digital no pueden almacenar todos los formatos que se usan de forma habitual a nivel de gestión, sino que solo tienen capacidad para aceptar un número reducido de formatos previamente escogido debido a sus mejores características técnicas, a su amplia aceptación, o a su condición de estándar abierto. La conversión de los formatos de origen de los ficheros a los formatos de archivo se conoce con el nombre de normalización.
- A lo largo de los años, los ficheros preservados posiblemente sufrirán migraciones de formatos u otras transformaciones técnicas a fin de garantizar su permanencia; para aplicar estos cambios será imprescindible conocer los formatos y otras características técnicas iniciales de los ficheros.
- En grandes sistemas de almacenamiento digital en los que se preserven grandes volúmenes de documentos, los ficheros se organizarán en los distintos soportes de almacenamiento (*racks* de discos, etc.) según sus formatos y tamaños, a fin de facilitar su gestión informática.

Las iniciativas pioneras en la puesta en servicio de sistemas de preservación en archivos y bibliotecas han confirmado la importancia de controlar de forma eficiente los formatos de los ficheros a preservar a fin de conocer con exactitud las características y las problemáticas técnicas de los ficheros que tendrán a su cargo, pues solo de esta forma se podrán planificar y aplicar las políticas adecuadas de preservación que requiera cada caso (Serra, 2008).

Una vez correctamente identificados los ficheros, la información sobre el formato debe ser registrada en el archivo digital como un dato clave a tener en cuenta en los procesos internos de gestión. Las fórmulas para mantener este registro, que se acumula a otros metadatos extraídos de los ficheros, pueden ser variadas, pero el seguimiento de estándares habrá de facilitar en el futuro la interconexión de archivos distintos. Dentro de esta visión, la solución que está teniendo una mayor aceptación es el esquema de metadatos METS, desarrollado por la Library of Congress, cuya finalidad es registrar en formato XML todos los metadatos técnicos de un documento, como identificación y ubicación de los ficheros que lo forman, relaciones entre los mismos (secuencia, manifestaciones, etc.), dependencias entre documentos (concepto de serie, etc.), descripción del contenido (catalogación), condiciones legales de uso, etc.

La obsolescencia tecnológica obligará a los archivos digitales a seguir procesos de vigilancia tecnológica a fin de advertir las necesidades de migraciones de formatos u otras soluciones que se deban aplicar a fin de asegurar la accesibilidad de los documentos. Esta preservación activa también descansará en buena parte sobre las informaciones disponibles de los formatos de los ficheros y se tendrán que documentar las acciones correctoras (migraciones, etc.) que en cada momento se apliquen. El formato de metadatos PREMIS está siendo evaluado para ser utilizado con esta función.

2. Documentos, ficheros y formatos

Los entes productores de la información digital (oficinas administrativas, laboratorios de investigación, artistas digitales, etc.) tienen lógicamente centrada su atención en el concepto de documento: ellos crean documentos, los usan y en el futuro, quizás, continuarán deseando acceder a ellos. Pero en el mundo digital los documentos están formados por ficheros, únicos (como en el caso de una fotografía), o por un conjunto interrelacionado de ellos (como en una página web). Todos los ficheros son en último extremo el resultado de la combinación de ceros y unos, pero estas combinaciones, llamados formatos, pueden ser muy distintas y necesitar de programas y equipamiento informático bien dispares para su descodificación y uso (Abrahms, 2007).

Esta caracterización entre documentos, ficheros y formatos es obvia para los profesionales de la informática y para muchos de la documentación y la archivística, pero no lo es para los usuarios comunes. Las empresas de software, de forma muy especial las creadoras de sistemas operativos, intentan facilitar el trabajo de los usuarios comunes aunque sea a costa de esconder la existencia de los tres niveles mencionados o, como mínimo, de minimizar la presencia de los formatos. Como resultado de ello, muchos usuarios no son conscientes de que un mismo programa puede trabajar con formatos

distintos o de que un mismo documento se puede representar con formatos alternativos. Una consecuencia de este desconocimiento es el cambio accidental de extensiones de ficheros o la inutilización total o parcial de ficheros debido a una manipulación con programas inadecuados.

Volviendo a los archivos digitales cabe preguntarse hasta que punto estas deficiencias en la cultura informática de los productores son generadoras de ficheros defectuosos o problemáticos desde el punto de vista de su preservación a medio y largo plazo. Es evidente que un pequeño porcentaje de los ficheros que se generan y acabaran llegando a los archivos son defectuosos desde el punto de vista técnico, pero es difícil precisar hasta qué punto se trata de un problema residual o de algo más serio, pues incluso habrá quien indicará con sorna que en los manuscritos medievales también se localizan firmas ilegibles o que son numerosos los libros mal impresos o con la compaginación incorrecta.

Archiving, Ingest and Handling Test (AIHT), es un proyecto financiado por la Library of Congress, que se desarrolló el año 2005 con el fin de comprobar distintos aspectos de la problemática inherente a la transferencia de colecciones documentales entre distintos archivos digitales (Abrahms, 2005a; Anderson, 2005; DiLauro, 2005; Nelson, 2005). Analizar la problemática de la identificación de ficheros no era el objetivo central del proyecto pero, de forma sorprendente, acabó siendo uno de sus principales resultados. Los cuatro sistemas de preservación digital que trabajaron en el proyecto (Harvard University, The Johns Hopkins University, Old Dominion University, y Stanford University) acabaron informando que una parte considerable de su tiempo lo emplearon en la identificación de los ficheros que habían de ingresar en los respectivos depósitos. Los informes finales del proyecto AIHT advirtieron que un número pequeño de ficheros con formatos defectuosos puede parecer irrelevante en términos relativos, pero que en el contexto de un sistema automatizado de ingestión de ficheros en un archivo digital invalida la carga automática total y obliga a elegir entre dos dolorosas alternativas: a) corregir de forma manual o semimanual los formatos defectuosos (una tarea que los autores de los informes consideraban como inasumible desde el punto de vista económico y del tiempo a consumir); y b) descartar los ficheros defectuosos y por tanto no preservarlos. Es posible que se puedan articular soluciones intermedias, pero en todo caso el proyecto AIHT mostró dos de los desastres que puede acarrear la generación incorrecta de ficheros: su pérdida a largo plazo o su conservación a unos costes inasumibles.

3. Elementos para la identificación de los ficheros

3.1. Identificación de los formatos

Los ficheros existentes en un soporte informático no se diferencian entre ellos a bajo nivel, pues todos son secuencias de ceros y unos. En cambio, para poderlos editar, visualizar o procesar es necesario conocer el formato con el que han sido gravados. A lo largo de la historia de la informática han surgido diversas iniciativas para identificar de forma rápida y segura el formato de un fichero. Las dos más extendidas son el uso de las extensiones en los nombres de los ficheros y el uso de "*magic numbers*".

El uso de una "extensión", o unas letras estandarizadas tras el punto final del nombre, en la denominación de un fichero ha sido uno de los mecanismos más usados para identificar su formato. Por ejemplo los ficheros gif suelen tener la extensión .gif. Este mecanismo ha sido utilizado desde siempre por los sistemas operativos Macintosh y Windows. La extensión permite una identificación muy rápida del formato, permite hacer listados de los ficheros por formato de forma ágil, es muy transportable, pues la identificación siempre viaja con el fichero, pero es poco fiable y poco informativa, pues se puede cambiar fácilmente y no distingue entre versiones diferentes de un mismo formato.

Actualmente existen diversos servicios online gratuítos que describen el/los formatos asociados a una extensión dada (http://www.file-extensions.org/, http://filext.com/, http://www.fileinfo.net/ y otras).

Los *magic numbers*, por su parte, consisten en un código estándar en la cabecera del fichero que indica el tipo de formato. Por ejemplo los ficheros gif tienen en su cabecera el código "GIF89a" (0x474946383961) o "GIF87a" (0x474946383761). Este mecanismo fue creado en el seno del sistema operativo Unix. Los *magic numbers* son un mecanismo altamente fiable y transportable, pues la

identificación se da junto a los propios datos; no es tan ágil de identificar cómo la extensión pues hay que acceder al fichero propiamente, y no sólo a su nombre. Desafortunadamente no se aplica a los formatos textuales, sólo a los binarios.

Para los ficheros textuales, especialmente para toda la familia de subformatos XML, se usa un mecanismo similar a los *magic numbers*, pero más explicito. Las declaraciones de tipo de documento, o esquema usado, describen con detalle el formato de los datos. Así por ejemplo un gráfico SVG siempre llevará en su cabecera la siguiente declaración <!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 1.1//EN" "http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd">. Los metadatos o declaraciones de tipo de documento son el mecanismo más informativo de los existentes. Suele ser bastante fiables también aunque en el entorno web no es extraño encontrar declaraciones de tipo de documento erróneas. De forma similar a los *magic numbers* no es un mecanismo especialmente rápido ni cómodo para procesar grandes volúmenes de archivos pues obliga a acceder al fichero propiamente dicho.

En algunos casos la identificación detallada del formato se da en metadatos externos al fichero y vinculados a este, como por ejemplo en el marco del Resource Description Framework (RDF). En este caso la identificación del fichero presenta problemas de transportabilidad pues hay que mantener la relación entre dos ficheros físicamente separados.

3.2. Denominación externa de los formatos

Una vez identificados los formatos, aparece la problemática de cómo referirse externamente a los mismos para permitir la interacción entre los ficheros y los sistemas operativos y las aplicaciones. Para automatizar estas relaciones es necesario establecer una forma normalizada para llamar a los distintos formatos, sean estos reconocidos directamente por su extensión o referenciados desde metadatos.

Para esta función tradicionalmente se han usado los MIME types (Fred, 1996a, 1996b) (http://www.iana.org/assignments/media-types/), cuyo origen se remonta a las extensiones del correo electrónico para incluir gráficos, programas, etc. Los MIME types cumplieron su objetivo en los primeros años de Internet, en los que era suficiente con indicar la gran familia del formato (audio, imagen, programa...) y el subtipo (texto, zip, mpeg...); pero se han quedado cortos con la explosión de versiones y subformatos existentes actualmente. Siguiendo con el ejemplo de las imágenes GIF, cuyo media type está definido por la RFC2045 y la RFC2046, éste sólo permite indicar que es una imagen gif, pero no si es la versión 89A (que por ejemplo permite generar gráficos con transparencias) o la versión 87A. Los MIME types son el indicador de formato recomendado por ejemplo en el conjunto de metadatos Dublin Core (http://dublincore.org/documents/dces/).

Para paliar las deficiencias de los descriptores MIME types en el proyecto PRONOM se han creado los Pronom Unique Identifiers – PUIDs– (http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm), que distinguen entre versiones y subformatos. En el Reino Unido se han adoptado como estándar para los metadatos relacionados con el gobierno electrónico (http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=872).

En el ejemplo de los ficheros GIF contamos con un identificador único PRONOM para cada uno de los subformatos existentes (fmt/3 para GIF 1987A; y fmt/4 para GIF 1989a). Como limitación de los PUIDs decir que aún no se han creado PUIDs para muchos tipos de ficheros existentes, especialmente en el ámbito del video y del audio (por ejemplo aún no existe ningún PUID para los ficheros OGG Vorbis).

3.3. Documentos XML bien formados

Los ficheros que ingresan en un sistema de preservación no solo se han de identificar a nivel de formato sino que en algunos casos también se ha de comprobar si su estructura está bien formada, es decir, si sigue de forma correcta las normas técnicas del formato correspondiente.

Los formatos textuales han vivido una expansión sin precedentes gracias a XML. XML ha establecido una sintaxis para crear lenguajes de marcado extensibles de forma muy práctica. Para hacer un documento XML sólo se deben seguir las pautas de apertura y cierre de marcas estructurales, la lógica de anidamiento entre ellas y la sintaxis para los atributos. Con estas simples reglas cada organización o particular puede crear su propio formato textual, con marcas propias según sus necesidades.

A un nivel superior de complejidad XML permite también describir el formato mediante una definición de tipo de documento (DTD) o un esquema (*schema*) en el que se describen los tipos de datos permitidos para cada elemento, la obligatoriedad o no de aparición y la cardinalidad de cada elemento, los atributos existentes... incluso rangos de valores y dependencias entre ellos en el caso de los esquemas.

Hablaremos de un documento XML bien formado si sigue las pautas de sintaxis establecidas por XML, y de un documento XML válido, si además de ser bien formado, se adecua a una definición de tipo de documento o esquema. Los documentos XML bien formados, y aún más los válidos, facilitan enormemente su manipulación posterior y, si fuera el caso, su conversión a nuevos esquemas. La validación de los documentos XML es por tanto de gran importancia en el contexto de la preservación, aún más si tenemos en cuenta que todos los sistemas de metadatos que se están propugnando con este fin (METS, PRONOM, MPEG-21 DIDL...) son desarrollos XML.

3.4. La problemática de los documentos compuestos

Algunos formatos recientes no son más que "wrappers" o envoltorios de un conjunto de ficheros en otros formatos, es decir, documentos compuestos en los que algunas de sus partes tienen su propio formato y características. Cómo ejemplo paradigmático, el formato PDF es un documento compuesto en el que pueden haber imágenes GIF o JPEG, pueden incluirse vídeos u otros elementos multimedia, etc.

En este caso, la identificación del formato, cuando debe servirnos de base para un procesado técnico, debe consistir no sólo en la identificación del formato "wrapper" sino también en la identificación de cada uno de sus componentes.

4. Herramientas para el control de los ficheros

Si en el entorno relativamente cerrado de un ordenador personal ya resulta difícil identificar y realizar las llamadas correctas a los distintos formatos de ficheros, en un gran archivo de preservación la complejidad se dispara. El volumen de ficheros que ingresaran en estos sistemas será enorme y también lo serán las necesidades de control posterior de los mismos para asegurar su pervivencia y legibilidad.

Los proyectos pioneros en preservación digital han identificado que una de sus primeras necesidades a nivel técnico es la de disponer de un sistema fiable y amplio, sino universal, actualmente inexistente, de identificación y denominación de los formatos, incluidas sus diferentes versiones. Entre las numerosas iniciativas que se han desarrollado para dar respuesta a esta necesidad (Brown, 2008) destacan dos. La primera es Pronom (Darlington, 2003) (http://www.nationalarchives.gov.uk/pronom/), un registro de formatos creado por los archivos nacionales del Reino Unido, que documenta las características de cada formato, da indicaciones para su reconocimiento y le asigna un identificador único, el PUID (Pronom Unique Identifier) (Sharpe, 2007), del que ya hemos hablado. La segunda iniciativa es el registro de formatos usados por la Library of Congress, de los Estados Unidos (http://www.digitalpreservation.gov/formats/), de carácter menos técnico y más orientada a dar consejos sobre los formatos más aceptables en un marco de preservación a largo plazo.

Pronom se está consolidando como una herramienta útil e integrable en sistemas informáticos más amplios por medio de herramientas como DROID. Con todo, la capacidad de reconocimiento de Pronom está limitada a unos 150 formatos, que aunque ahora son los de uso más habitual, no representan más que una fracción del total de los existentes. Justamente para lograr un alcance exhaustivo desde la Universidad de Harvard se lanzó la idea de crear un Global Digital Format Registry (GDFR) (http://www.gdfr.info/), dependiente de alguna entidad independiente, y de uso universal (Abrams, 2003, 2005). La idea de crear el GDFR ha recibido el apoyo de distintas instituciones internacionales y actualmente se encuentra en fase de desarrollo.

De momento, la mayoría de las herramientas profesionales de detección de formatos se basan en la información proporcionada por los registros actualmente existentes, principalmente Pronom. Estos programas facilitan las diversas funciones relacionadas con la ingestión siguiendo el protocolo OAIS:

- Identificar, es decir, a partir de un fichero dado indicar a que formato digital corresponde.
- Validar, es decir, a partir de un fichero y unos metadatos identificativos verificar si ambos elementos son consistentes.
- Caracterizar, es decir, dar los datos técnicos descriptivos necesarios para conocer las propiedades más destacables del objeto.
- Normalizar, es decir, convertirlo a otro formato, aceptado por el centro que realiza la preservación.

Entre los programas para la gestión de los formatos digitales con miras a su preservación destacaremos DROID, JHOVE y XENA, creados con este fin, y TrID, de diferente origen, pero muy útil en algunos aspectos.

DROID (Digital Record Object Identification) (http://droid.sourceforge.net), fue creado por los archivos nacionales del Reino Unido en Java, como una herramienta para facilitar la exploración directa del registro Pronom. Este programa, en su versión actual (3.0.0, mayo 2008), realiza sólo la función de identificación de formatos a partir de la codificación interna y de la extensión de los ficheros. Las nuevas incorporaciones de formatos reconocidos se envían a los usuarios como actualizaciones del programa vía web services, de forma automática. DROID permite tratar listas de ficheros (a partir de una lista XML o de una carpeta), identifica los ficheros según su PUID y permite exportar los resultados como CSV. Cómo ya se ha dicho la capacidad de reconocimiento está limitada a unos 150 formatos.

JHOVE (JSTOR/Harvard Object Validation Environment) (http://hul.harvard.edu/jhove), creado por la Harvard University con fines de preservación, realiza las funciones de identificación, validación y caracterización de formatos. Desarrollado sobre Java es un programa modular y extensible, que permite una instalación selectiva de sus funciones. En la parte de caracterización describe con un alto nivel de detalle las características técnicas de la información contenida, y en el caso de formatos compuestos (como PDF) también de sus componentes. A pesar de ser un programa muy robusto y completo, su versión actual (1.1 de febrero de 2008), tiene la limitación que aún trata pocos formatos de ficheros (textuales, JPEG, TIFF, AIFF, WAVE y PDF). Se encuentra en desarrollo una nueva versión, JHOVE2 (http://confluence.ucop.edu/display/JHOVE2Info/Home) (Abrahms, 2008), en la que se separaran las funciones de identificación y validación, y también se facilitará la personalización de la herramienta para su adaptación a los requerimientos de preservación de las distintas instituciones.

XENA (XML Electronic Normalising for Archives) (http://xena.sourceforge.net), es una herramienta creada por los archivos nacionales de Australia que implementa la identificación y la normalización de formatos. La normalización tiene como destino formatos abiertos como la familia de formatos Open Office para documentos de ofimática, FLAC para archivos de audio, etc. Como resultado de la normalización se obtienen ficheros XML con los metadatos descriptivos y que contienen en su interior el contenido binario del archivo. Xena, aunque ya está en la versión 4.2 de Junio de 2008, está muy poco documentado respecto a los demás, pero es un buen ejemplo del proceso completo de ingestión.

TrID (http://mark0.net/soft-trid-e.html) es un programa creado por Marco Pontello para la identificación de formatos, fuera del ámbito de la preservación. TrID es un programa comercial, que se distribuye gratuitamente para uso personal, pero no en código abierto. TrID realiza la función de identificación basándose en los mágic numbers, por lo cual está especialmente indicado para formatos binarios. En la última consulta realizada (25-nov-2008) su base de datos contaba con más de 3000 formatos reconocidos, y los usuarios pueden aumentar esta lista a partir de la utilidad TrIDScan. TrID no da respuestas tan seguras como los programas anteriores, pero tiene la ventaja de su gran colección de formatos reconocidos, y además cuenta con una versión online que no requiere instalación alguna, aunque sólo puede procesar los ficheros de uno en uno. En un experimento de identificación a partir de un fichero mp3, al que se le eliminó la extensión, el único programa de los citados en este artículo capaz de identificarlo fue TrID.

5. Tendencias y líneas de futuro

Las metodologías y las técnicas de preservación digital han avanzado mucho estos últimos años pero aún disponemos de pocos ejemplos de aplicación real, en un ambiente de producción, de estos sistemas. Los proyectos de preservación digital que distintas bibliotecas y archivos nacionales o grandes bibliotecas universitarias están llevando a cabo parten de grandes volúmenes de documentos, pero todos ellos basados en ficheros muy controlados. En este sentido recordemos las estrategias para preservar revistas digitales (básicamente formalizadas por ficheros XML y PDF), colecciones de fotografías (JPEG, JPEG 2000) o documentos históricos (TIFF, PDF); vemos que se trata de ficheros creados de forma eficiente por empresas editoriales o bien el resultado de procesos controlados de digitalización de antiguos fondos en papel. El resultado son ficheros con formatos creados correctamente y que por tanto raramente pueden presentar deficiencias en el momento de la ingestión.

Ahora bien, tan pronto se generalice la administración electrónica y la facturación electrónica entre las empresas privadas (fenómenos que en España se encuentran en plena expansión) la problemática de la preservación se volverá más compleja, con un mayor número de ficheros a tratar y una mayor diversidad en sus formatos.

Este escenario futuro requiere tomar medidas para minimizar los riesgos que lleva asociado.

6. Conclusiones

Los sistemas de preservación digital sólo podrán funcionar de forma eficiente (es decir, con unos costes asumibles y una fiabilidad constatable) si logran normalizar completamente sus procesos y automatizar los procedimientos de ingestión. Aunque esta comunicación sólo se ha centrado en algunos aspectos de la fase de ingestión de documentos, creemos que se ha puesto en evidencia la importancia y la complejidad de la identificación y validación de los formatos.

En estos momentos ya se encuentran disponibles distintas herramientas fiables que permiten poner en marcha el control automatizado de los formatos y, por tanto, éste no ha de ser una barrera que impida construir sistemas de preservación digital. Estas herramientas, como ya se ha indicado, presentan limitaciones en cuanto a su flexibilidad de uso o al número de formatos que identifican, pero estos son obstáculos que serán superados según aparezcan nuevas versiones o herramientas más avanzadas. En este sentido, hay que resaltar que la ampliación de la base de usuarios es quizás la mejor forma para presionar y también para colaborar en la mejora de estas herramientas.

7. Bibliografía

Abrams, Stephen L. (2005a): "Establishing a Global Format Registry". *Library Trends*, 54.1, p. 125-143.

Abrahms, Stephen (2007): "File Formats". *DCC Digital Curation Manual*. 53 p. http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats/file-formats.pdf

Abrahms, Stephen; Chapman, Stephen; Flecker, Dale; Kreigsman, Sue; Marinus, Julian; McGath, Gary; Wendler, Robin (2005b): "Harvard's Perspective on the Archive Ingest and Handling Test". *D-Lib Magazine*. 11(12). http://www.dlib.org/dlib/december05/abrams/12abrams.html

Abrahms, Stephen L.; Seaman, David (2003): "Towards a global digital format registry". *World Library and Information Congress: 69th IFLA General Conference and Council*. Berlin. http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Abrahms, Stephen; Morrissey, Sheelagh; Cramer, Tom (2008): "What? So what?: The next generation Jhove2 architecture for format-aware characterization". *iPres 2008. Proceedings of the fifth international conference onpreservation of digital objects*. London. p. 86-92.

Anderson, Richard; Frost, Hannah; Hoebelheinrich, Nancy; Johnson, Keith (2005): "The AIHT at Stanford University: Automated Preservation Assessment of Heterogeneous Digital Collections". *D-Lib Magazine*. 11(12). http://www.dlib.org/dlib/december05/johnson/12johnson.html

Brown, Adrian (2008): *White Paper: Representation Information Registries.* Planets. 26 p. http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf

Darlington, Jeffrey (2003): "Pronom: a practical online compendium of file formats". *RLG DigiNews*, 7(5).

DiLauro, Tim; Patton, Mark; Reynolds, David; Choudhury, G. Sayeed (2005): "The Archive Ingest and Handling Test: The Johns Hopkins University Report". *D-Lib Magazine*, 11(12). http://www.dlib.org/dlib/december05/choudhury/12choudhury.html

Freed, N.; Innosoft; Borenstein, N.; First Virtual (1996a): *Multipurpose Internet Mail Extensions (MIME). Part One: Format of Internet Message Bodies*. RFC 2045. http://www.isi.edu/in-notes/rfc2045.txt

Freed, N.; Innosoft; Borenstein, N.; First Virtual (1996b): *Multipurpose Internet Mail Extensions (MIME). Part Two Media types*. RFC 2046. http://www.isi.edu/in-notes/rfc2046.txt

Keefer, Alice; Gallart, Núria (2007): *La preservación de recursos digitales: el reto para las bibliotecas del siglo XXI*. Barcelona: Editorial UOC. 232 p.

Nelson, Michael L.; Bollen, Johan; Manepalli, Giridhar; Haq, Rabia (2005): "Archive Ingest and Handling Test: The Old Dominion University Approach". *D-Lib Magazine*, 11(12). http://www.dlib.org/dlib/december05/nelson/12nelson.html

Reference Model for an Open Archival Information System (OAIS), CCSDS, 2002, http://www.ccsds.org/documents/650x0b1.pdf

Serra, Jordi (2008): *Los documentos electrónicos. Qué son y cómo se tratan*. Gijón, Trea. 187 p.

Sharpe, Robert (2007): "An automated framework for the characterisation of records utilising Pronom, Droid and other tools". *Tools and Trends: International Conference on Digital Preservation at the occasion of the retirement of Johan Steenbakkers*. The Hague, Koninklijke Bibliotheek, 1-2 November 2007. http://www.kb.nl/hrd/congressen/toolstrends/presentations/Sharpe.pdf

Searle, Sam; Thompson, Dave (2003): "Preservation metadata: Pragmatic first steps at the National Library of New Zealand". *D-Lib Magazine, 9*(4).
http://www.dlib.org/dlib/april03/thompson/04thompson.html