METADATOS EN EL MUNDO BIBLIOTECARIO:

Teoría, práctica y aplicaciones prácticas en el entorno digital (preservación digital y linked data)

Profesor: Isabel Bordes Cabrera

INDICE

1. METADATOS EN PRESERVACIÓN DIGITAL

INTRODUCCIÓN

La preservación digital se refiere a las diferentes estrategias que se siguen para asegurar que los objetos digitales, de cualquier tipo, que tenemos en nuestras instituciones y que hoy son accesibles, lo sigan siendo el día de mañana.

Seguramente todos hemos vivido la experiencia de encontrarnos con que ya no podemos acceder a cierta información digital que teníamos guardada ó que creíamos a salvo en páginas web de instituciones sólidas[1]. Por ejemplo, la que teníamos en discos que ya no se utilizan, como los antiguos discos floppy. O que al insertar un CD o DVD en un lector ya no se pueden leer. También es posible que tengamos ficheros en formatos tan antiguos (por ejemplo antiguas versiones del Office) que ya no se pueden leer con las nuevas actualizaciones.

Ilustración 1: **Cartel en la Library of the Congress anunciando el cierre de la biblioteca con motivo del *shutdown* del gobierno federal**

[1] El reciente cierre del Gobierno Federal (1 octubre de 2013) afectó incluso a la página web de la Library of Congress. Esto puede ser un verdadero riesgo para la preservación digital pues precisamente en la biblioteca tienen sede de grupos de trabajo para el mantenimiento de modelos de metadatos claves en este campo (METS, PREMIS…). Durante los días de cierre no se podía acceder a esquemas de validación de los modelos (a no ser que se tuvieran copias personales), directrices y lista de comprobación en torno al uso combinado de PREMIS y METS…etc.

Esto que a pequeña escala lo hemos experimentado todos, se convierte en un verdadero problema para las instituciones patrimoniales. Éstas, tienen la misión de asegurar que la información que guardan esté disponible a largo plazo y, en el campo de lo digital, el largo plazo es complicado de conseguir.

Los problemas técnicos que implica la preservación digital son muchos: obsolescencia de formatos, software, hardware y/o soportes; fragilidad de los soportes digitales y la propia volatilidad de la información digital. A todos ellos, se une otro muy importante, y es que los proyectos de digitalización se enfocan (aunque esto cada vez está más superado) desde la perspectiva de la preservación del documento físico. Típicamente, igual que la microfilmación, la digitalización se planteaba como una medida de preservación de los documentos. La digitalización, presentaba además la ventaja de que permite una difusión infinitamente superior a la del microfilm.

De este modo, en la digitalización primaba muchas veces la preservación del documento físico y la difusión de los objetos digitales. La preservación del nuevo objeto digital quedaba como algo relativamente menor. Sin embargo, sólo esta preservación asegura verdaderamente la perdurabilidad en el tiempo de la información contenida en el soporte físico.

La preservación digital no es algo que afecte sólo a las bibliotecas, sino a todos los sectores que utilizan tecnologías digitales. De hecho, el modelo OAIS de preservación digital que veremos más adelante lo diseñó la NASA, tras descubrir que no eran capaces de acceder a información de misiones espaciales por encontrarse en formatos y soportes obsoletos.

En un intento por dar una dimensión real a la explosión de información en formato digital a la que nos enfrentamos, y la cantidad de almacenamiento disponible, en el informe *Diverse exploding digital universe* (IDC, 2008) se recogen unos gráficos de lo más elocuentes:

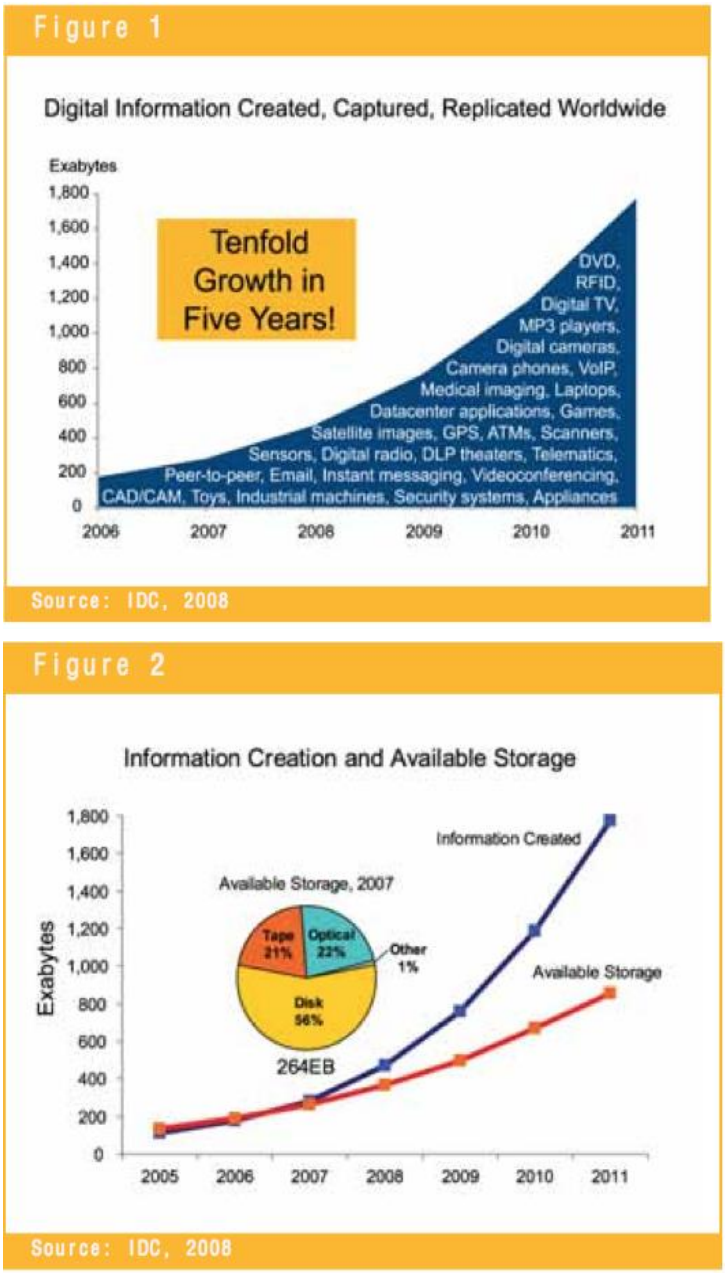

Ilustración 2: Crecimiento de información digital vs. almacenamiento disponible

En el momento de redacción de dicho informe se estimó que la información digital se multiplicaría por 10 de 2005 a 2011[2], estos datos enfrentados al crecimiento del almacenamiento disponible (ilustración 2) ponen de relieve un desfase a todas luces evidente. Lógicamente estas estimaciones no dejan de ser un intento por entender a qué nos enfrentamos cuando empezamos a abordar el tema de la preservación digital, ya que conocer los datos reales a cada momento sería una tarea prácticamente imposible, comparable quizá a la carrera de Aquiles y la Tortuga. Por ejemplo, al hacer estos cálculos, los autores no podían prever las inundaciones que en 2011 afectaron seriamente a Tailandia y que podían poner en peligro la fabricación de dispositivos de almacenamiento (Post de 4 de noviembre de 2011 en el Blog Bits del New York Times), elemento clave de la preservación digital.

Como ya hemos dicho la preservación de información digital es un tema que afecta a todos los sectores; si bien es verdad, que cada tipo de objeto digital muestra una pérdida de interés en su contenido distinto según el campo en el que nos movamos. Tal y como puede verse en el siguiente gráfico[3], aquellos gestionados por las bibliotecas digitales son los que tienen una vida media mayor:

[2] Tal y como se explica en el informe: "para obtener estos datos se estimó el número de "unidades de información" creados en un año (archivos, imágenes, canciones, minutos de vídeo, llamadas per cápita, paquetes de información), se convirtió esas unidades de información en MB asumiendo parámetros como la resolución, la compresión y la utilización de las mismas; por último, se estimó el número de veces que una unidad de información podría replicarse, compartirse o almacenarse.

[3] Extraído del Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, el cual toma a su vez el gráfico de una presentación inédita de Nowell, L. (2008). Data Preservation and Access: A Global Challenge. Unpublished Presentation. National Science Foundation.

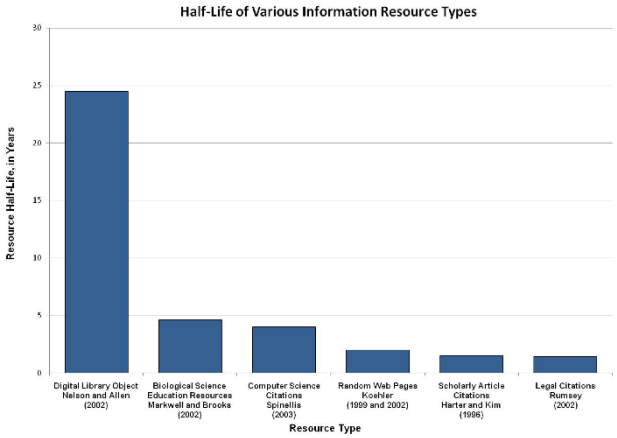

Ilustración 3: Vida media de los recursos (años) en función del tipo de recurso

Y es que, lo cierto es que en instituciones patrimoniales la información digital presenta una necesidad de conservación a largo plazo mucho mayor que la de otros ámbitos, por la propia naturaleza de estas instituciones cuya misión es preservar el conocimiento.

En cualquier caso, también hay que tener en cuenta que no por ser digital hay que preservarlo todo. La facilidad que existe hoy día a la hora de producir información digital no debe confundirnos y hacernos creer que todo lo que tiene esta "dimensión digital" debe de conservarse. Nicholas Joint lo expresa, a nuestro modo de ver, de una manera bastante acertada:

> *[...] "We never tape-recorded […] conversations between academics on campus in the 1960s. […] why should we worry about losing the equivalent email versions of these conversations, just because they're written down and make us think they are like hard copy letters of Leibniz or Newton? […]We are letting "format confusion" distort our understanding of information management [...]* [4]

De hecho, una de las primeras preguntas a las que tendremos que enfrentarnos a la hora de poner en marcha un plan de preservación (y que entronca directamente con esta observación): ¿qué podemos/debemos y vamos a preservar dentro de nuestra colección digital?

[4] "Choosing between print of digital collection building in times of financial constraint". Library Review vol. 58, nº4, 2009.

La preservación digital es un asunto complejo porque intervienen muchos factores: requiere conocer la distinta tipología de objetos digitales que existen, los diferentes riesgos que amenazan su perdurabilidad y las distintas estrategias de salvaguarda de la información digital.

Además de conocimientos técnicos, la preservación digital tiene un gran componente de gestión de proyecto a largo plazo. En las instituciones patrimoniales se está convirtiendo en una línea de actuación estratégica clave. Decisiones como qué preservar, cuántos recursos invertir en la preservación, cómo financiarla, etc. superan los aspectos meramente técnicos pero son claves para plantear la preservación a largo plazo de objetos digitales.

En 1997, se utilizó el término Digital dark age: para describir un posible futuro en el que resulte imposible leer documentos históricos por estar almacenados en formatos obsoletos[5]. Evitar que esto suceda dependerá en gran medida de que desde las instituciones encargadas de preservar la memoria se tomen medidas fundamentadas y técnicamente viables. Medidas que deben estar acompañadas de la convicción de que esta nueva realidad no puede ser desatendida y necesita recursos suficientes para poder llevarla a cabo con los niveles de éxito necesarios.

1.1. Conceptos básicos en torno a la Preservación digital

1.1.1. Definición de preservación digital

La **Preservación digital** (ALCTS, 2007)[6]: es el conjunto de políticas, estrategias y acciones para garantizar el acceso al contenido digital (ya sea de origen o por conversión

[5] Se utilizó por primera vez en la 63rd Conference de la IFLA por parte de Terry Kun (A Digital Dark Ages? Challenges in the Preservation of Electronic Information)

[6] ALCTS: Association for Library Collections and Technical Services
http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.cfm

de formatos), manteniéndolo independiente de los fallos que pueda tener el soporte que lo contenga y de los cambios tecnológicos. Y todo ello con el fin de que a lo largo del tiempo sea posible una interpretación correcta del contenido "autenticado".

Estas políticas, estrategias y acciones que menciona esta definición deben de girar en torno a tres ejes fundamentales:

1. La **creación de un contenido** basado en especificaciones técnicas claras y completas que permitan generar ficheros máster fiables junto con la producción de los metadatos necesarios para su gestión;
2. Asegurar la **integridad del contenido**, es decir, que los objetos digitales estén correctamente identificados, que se documente cualquier cambio que sufra y se proteja de virus;
3. Garantizar el **mantenimiento del contenido** en el tiempo mediante la infraestructura adecuada y diferentes estrategias (migración, emulación, refresco… que veremos en el tema).

Los últimos dos puntos (integridad y mantenimiento en el tiempo) están íntimamente relacionados porque en realidad es imposible hablar de mantenimiento del contenido si se pierde la integridad. Es decir, por mucho que tengamos "guardado" algo, si no sabemos qué es, o si es algo diferente de lo que creemos, o si son ficheros corruptos… no podemos decir que lo estamos manteniendo. Por tanto la integridad es parte inherente al mantenimiento del contenido.

1.1.2. Evolución del concepto de preservación digital

El concepto de preservación digital nació como una consecuencia lógica a la imparable evolución de las nuevas tecnologías y nuestra cada vez mayor dependencia de ellas: en 1923 se inventa la máquina Enigma para transcribir lenguajes codificados; en 1951 el mercado cuenta con el primer ordenador comercial de IBM; en 1969 surge el primer nodo de ARPANET experimento militar precursor de Internet; en 1981 aparece los discos 3 ½"; en 1984 se confirma la existencia de virus informáticos; en 1993 surge la versión pública de la www…). Sin lugar a dudas, a partir de los años 90, ya no hay marcha atrás:

- comienzan a surgir los primeros intentos de normalización en torno a lo digital, nacen publicaciones especializadas (1995, D-Lib Magazine),
- surgen repositorios digitales especializados (1995, JSTOR archivo digital para publicaciones de investigación)
- empiezan a generalizarse las respuestas institucionales con reuniones de expertos. Se inicia la aparición de recomendaciones por parte de organismos clave en el panorama internacional, destacando el punto de partida que supuso el informe "Preserving digital information" de la RLG y la Comission on Preservation and Access (CPA). Este informe constituye una referencia fundamental en el que se recogen problemas y recomendaciones en torno a la preservación digital tanto para materiales escaneados como para aquellos conocidos como "born digital".

Relativamente pronto se empieza a ser consciente de la urgencia de dar respuesta al reto de garantizar el acceso a largo plazo y de poner en marcha acciones colectivas.

Así pues, a partir del año 2000 vemos cómo:

- hay cada vez una mayor especialización en cuanto al enfoque de los problemas.
- Surgen las iniciativas de normalización y la publicación de mejores prácticas (PREMIS; TRAC; PRONOM; GDFR…)[7]
- La Comisión Europea comienza a financiar proyectos en torno a la preservación digital (Planets; Keep…)

Para seguirle el pulso a la complicada aunque relativamente breve historia del concepto de preservación digital, y a las iniciativas surgidas en torno a este tema, resulta extraordinariamente útil la cronología elaborada por la Universidad de Cornell en torno a la Preservación digital http://www.dpworkshop.org/dpm-eng/timeline/popuptest.html. Este recurso pedagógico tal y como se explica en la web:

> "destaca eventos, proyectos, publicaciones y cambios tecnológicos clave que afectan a la utilización de la tecnología digital y a los esfuerzos destinados a su preservación"

[7] PREMIS: PREservation Metadata; TRAC: Trustworthy Repositories Audit and Certification; PRONOM: ;GDFR: Global Digital Format Registry; UDFR: Universal Digital Format Registry…

Esta cronología no sólo permite tener una visión global, sino que tiene un formato que permite acceder de manera selectiva a otras seis facetas de la misma historia:

- desarrollos generales = eventos en la historia de la preservación y la tecnología digital
- protocolos y formatos = a medida que la tecnología ha evolucionado, también lo han hecho los estándares y protocolos
- redes = puntos clave en torno a las redes computacionales
- hardware y software = evolución de los ordenadores y las soluciones de software
- soportes = todo en torno a los soportes de almacenamiento externo
- crisis y obsolescencia = eventos que afectarían negativamente a la preservación digital
- respuesta organizativa = acciones llevadas a cabo por los gobiernos y las organizaciones como reacción a la proliferación de la información digital.

1.1.3 *Document Like Object* (DLO)

Un concepto importante a la hora de hablar de Preservación es el de *Document Like Object*. Con esta expresión nos referimos a la **unidad documental o al documento digital mínimo, que forma parte de una colección digital, al cual se le aplican metadatos para su descripción y recuperación.** Es decir, es aquello que se quiere preservar.

Ilustración 4: : Composición de un documento entendido como objeto digital

Los DLOs tienen una composición más o menos compleja, pero que de manera extraordinariamente esquemática puede "reducirse a":

- **un identificador unívoco y preferiblemente persistente.** Este podría ser, por ejemplo: una **URL** (*uniform resource locator*), un **DOI** (*digital object identifier*), un **PURL** (*persistent uniform resource locator*). En el caso actual de la Biblioteca Digital Hispánica (BDH)[8], como identificador podría considerarse el PID (persistent identifier) que el SGOD (Sistema Gestor de Objetos Digitales) Digitool da a cada uno de los objetos del repositorio.

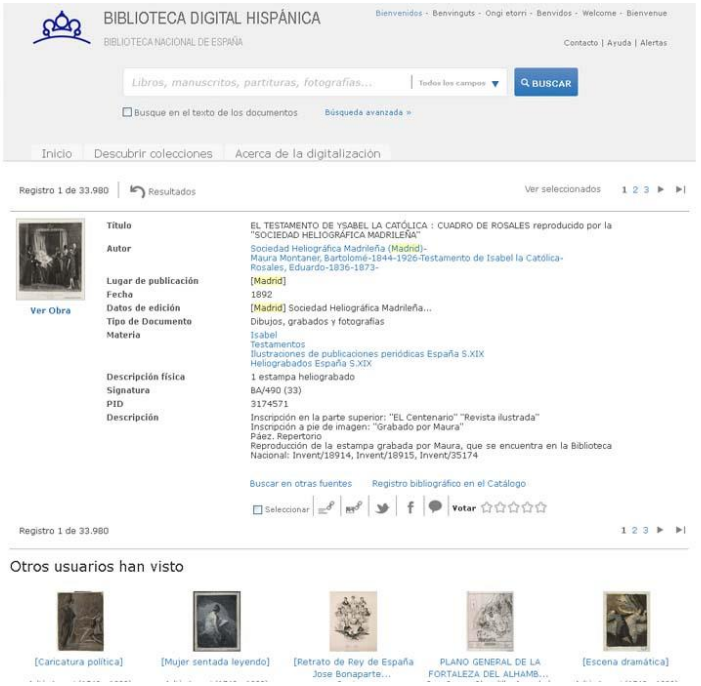

Ilustración 5: visualización detallada de un registro en Biblioteca Digital Hispánica

- unos **metadatos** descriptivos, administrativos y estructurales
- **el propio contenido** (en el caso de la ilustración sería la imagen del recto del folio 157 del Códice Madrid II de Leonardo da Vinci conservado en la BNE bajo la signatura MSS/8936)

[8] Biblioteca digital de la Biblioteca Nacional de España (BNE), nacida en 2008 al amparo de un convenio de digitalización masiva firmado entre la BNE y Telefónica Patrocinios.

- **elementos de seguridad adicionales** mediante lo que conocemos como Medidas Tecnológicas de Protección (MTP) [9]

Es importante darse cuenta de que los DLOs, aunque dependen de elementos informáticos, precisando de un entorno de lectura y/o uso, en realidad son independientes del soporte físico. Los DLOs son la información a preservar, independientemente de su soporte.

Es fundamental tener presente estas características ya que, si no, corremos el riesgo de aplicar a nuestras colecciones digitales (consciente o inconscientemente) nociones que derivan de la experiencia con los libros: fijos, estables y duraderos. De hecho, si hay algo que singulariza al DLO es la falta de fijeza, estabilidad y durabilidad de los soportes, así como la ruptura del vínculo que existe entre los datos y los soportes (Alice Keefer y Miguel Térmens, 2010) . La volatilidad de los soportes digitales, en último término determina la facilidad con la que pueden alterarse y eliminarse datos almacenados. Así pues, no basta con centrar los planes de preservación a largo plazo en los soportes digitales, éstos y otros aspectos (como pueden ser los propios formatos de los archivos) deben de ser gestionados correctamente, reemplazándose de manera sistemática y programada[10].

1.1.4 Tipología de recursos digitales

De manera muy genérica (puesto que este módulo se centra más en los metadatos) queremos hacer hincapié en la variada tipología de recursos digitales, lo cual supone un reto continuo para la preservación digital. Esta puntualización la hacemos sólo para que seáis conscientes de que la preservación digital implica una vigilancia tecnológica constante para conocer las características relevantes sobre nuestros

[9] estas medidas son las utilizadas por los titulares de derechos de propiedad intelectual con el fin de controlar: el acceso, la reproducción y la comunicación de los DLOs ej: limitar la funcionalidad de una obra, sistemas anticopia, sistemas de acceso controlado mediante contraseña, software de tipo *shareware*… Estas MTP, ó TPM en inglés, deben distinguirse (aunque algunos autores no lo hagan) del concepto de *Digital Rights Management* (DRM). Las DRM son medidas tecnológicas de identificación y seguimiento de las obras en el entorno digital, con el fin de facilitar la explotación de los derechos de propiedad intelectual patrimoniales ej: certificados digitales, sistemas de estenografía…

[10] Smith, A. Access in the future tense, 2004, p.4.

recursos digitales, sus formatos, la tecnología necesaria para su visualización, reproducción…

Y es que sólo conociendo los recursos digitales que queremos preservar, podremos tomar decisiones sobre los esquemas y modelos de metadatos que debemos utilizar.

Según el método de creación, podemos hablar de:

Born digital, son aquellos que ya nacen como digitales, sin una manifestación analógica previa.

Ilustración 6: ejemplo de objetos nacidos digitales

Los crean personas, entidades, la comunidad científica; pueden adoptar miles de formatos y están en evolución constante.

Ej.: páginas web, bases de datos, imágenes médicas, documentos de paquetes de ofimática, blogs…

Aquellos que nacen por digitalización, parten de una manifestación analógica, se obtienen mediante el uso de escáneres, cámaras digitales y otras técnicas, pero hablamos siempre de un número limitado de técnicas de reproducción. Además originan un nº "reducido" de formatos, facilitando en cierta medida la preservación a largo plazo.

Ilustración 7: Ejemplos de objetos que se originan en procesos de digitalización

El método de creación puede influir tanto en la gestión, como en las decisiones de preservación que se tomen respecto a un determinado tipo de DLO. Su método de

creación puede determinar un determinado formato de archivo para el que nuestra institución pueda tener mayor o menos capacidad técnica de gestión. Por ejemplo, no estamos preparados igual para conocer y preservar un tiff, que un archivo de Garage Band (programa de edición musical de Apple), especialmente en instituciones que sólo se manejan en entornos de Windows.

Según el origen de los DLOs podemos hablar de: *recursos de creación interna* y *recursos de creación externa*. Reconocer esta tipología es reconocer distintos niveles de voluntad, capacidad y/u obligación legal de una institución de cara a garantizar el acceso a largo plazo. Por ejemplo, poder llevar a cabo una migración de formato o incluso de soporte (estrategias habituales en materia de preservación) no es algo tan sencillo en un recurso digital protegido por propiedad intelectual[11] y que, como tal, puede contar con Medidas Tecnológicas de Protección anticopia.

1.1.5 Vulnerabilidad de la información

Como ya os hemos comentado la información digital es muy vulnerable y a esta vulnerabilidad contribuyen varios factores agrupables en las siguientes categorías:

- Obsolescencia tecnológica
  - Formatos y software
  - Hardware y soportes
- Fragilidad de los soportes digitales
- Volatilidad de la información digital

La obsolescencia tecnológica se refiere a la incapacidad de uso de determinados elementos informáticos (hardware, software, soportes), por ejemplo, por desaparición de los medios que permitían el acceso a los mismos. Esta incapacidad lo que da lugar es a recursos digitales inutilizables, también llamados *orphans* en la literatura especializada anglosajona (nada que ver con el concepto de obra huérfana, referido a obras de las que se desconocen los titulares de los derechos de propiedad intelectual).

[11] Por ejemplo, una base de datos en dvd

Ilustración 8: La obsolescencia tecnológica puede afectar al hardware, software y los soportes de información, i.a.

Son muchas las razones que condicionan la obsolescencia tecnológica, y aunque no directamente relacionado con la preservación digital, uno de los aspectos que podría relacionarse con este tema es la obsolescencia planificada de los productos que adquirimos en la sociedad de consumo en la que nos encontramos[12].

La obsolescencia tecnológica puede afectar, por un lado, a los formatos[13], al software; pero también al hardware y los soportes.

Existen muchos factores que contribuyen a la obsolescencia de un formato (actualizaciones de software, formatos que quedan superados y resulta incompatible con software actual, mercado tecnológico que puede originar interrupción de desarrollo de un formato). La falta de documentación sobre un determinado formato, puede llevar también a su obsolescencia. Recientemente, y a través de distintas iniciativas internacionales se ha hecho un esfuerzo de catalogación, documentación, estudio de relaciones, variaciones que ha dado lugar a herramientas que permiten la identificación,

[12] Aunque no está relacionado directamente con preservación digital os recomendamos el vídeo "the story of stuff" de Annie Leonard, que habla de la obsolescencia programada, demostrando que este no por ser una realidad, es algo sostenible.

[13] Se entiende por formato aquellos fundamentos estructurales y organizativos que permiten la construcción de un archivo válido, así como el desarrollo de software que permita decodificarlo y presentarlo (en pantalla, a una impresora o cualquier otro dispositivo). Todos esos fundamentos deben de quedar recogidos en documentos que conocemos como especificaciones

caracterización y validación de formatos (registros de formatos como PRONOM, UDFR; herramientas libres de validación como JHOVE, JHOVE2…)[14].

En lo que respecta al software, todo software está sujeto a la obsolescencia, y como consecuencia todo archivo también lo está. El seleccionar un programa más o menos estable supone un menor riesgo de obsolescencia a corto plazo. Sin embargo, se corre el riesgo de perder la adaptación al entorno tecnológico (CPU, sistemas operativos, esquemas de codificación, protocolos de transferencia de datos…)

A la hora de decantarnos por un formato u otro, conviene analizar con qué tipo de especificaciones cuenta dicho formato. En el tutorial de la Universidad de Cornell podéis ver un análisis de los pros y contras de uno y otro (http://www.dpworkshop.org/dpm-eng/oldmedia/obsolescence1.html

- especificaciones propietarias y cerradas (psd de photoshop, .doc…)

Ilustración 9: Ejemplos de especificaciones propietarias y cerradas

- especificaciones propietarias y abiertas (adobe, flash player, dejavu…)

Ilustración 10: Ejemplos de especificaciones propietarias y abiertas

[14] Existen también herramientas desarrolladas a nivel privado como TrID de Marco Pontello; o páginas web como file-extensions.org que permiten extraer algo de información a partir de la extensión de los archivos a nivel usuario no experto. Sin embargo, son menos versátiles que iniciativas como la de PRONOM, de uso obligatorio en la preservación digital a nivel institucional.

- especificaciones no propietarias y abiertas (mepg, xml, jpeg…)

Ilustración 11: Ejemplos de especificaciones no propietarias y abiertas

En los últimos cincuenta años el hardware ha estado sometido a un vertiginoso ritmo de obsolescencia, y el panorama actual no es indicativo de que esta tendencia vaya a cambiar. En general se han mejorado parámetros como la potencia, la velocidad, la eficacia, y el coste vs. valor; todo ello ha llevado a dispositivos más rápidos, productivos, de mayor capacidad, con nuevas funcionalidades (producción de audio-vídeo, imagen digital, navegación web…).

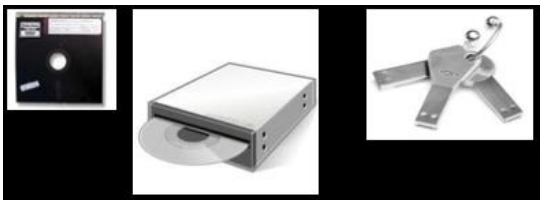

Ilustración 12: Ejemplos de algunas de las soluciones de hardware que se han ido sucediendo en materia de almacenamiento digital

En lo que respecta, a los soportes la evolución también es constante[15]; en general las tendencias del mercado apuntan a: menor tamaño, mayor capacidad, y menor precio por unidad de almacenamiento…En general todos estos cambios no son sino un factor más de obsolescencia.

Pero además, los soportes físicos en los que almacenamos la información digital, no sólo son susceptibles a quedar obsoletos por el avance de la tecnología sino que, además, su condición física hace que resulten inestables pues tanto el uso, como las

[15] Para tener una visión más exhaustiva recomendamos que se eche un vistazo a la "cámara de los horrores" (recurso elaborado por la Universidad de Cornell) en la que pueden verse los distintos tipos de soportes (perforados, discos, cintas y "en estado sólido"), sus características, su tiempo de permanencia en el mercado…En este recurso se incluyen desde soportes ya extintos, a soportes que "únicamente" pueden considerarse como que están en riesgo.

condiciones ambientales pueden suponer un riesgo de pérdida de la información que en ellos reside.

En los depósitos de nuestras instituciones deben de convivir materiales muy diversos. El problema que deriva de esta circunstancia es que, lógicamente, no todos ellos precisan de las mismas condiciones de Humedad (H) y Temperatura (T). Por tanto, dependiendo de los medios con los que contemos (no siempre serán posibles depósitos especializados según el soporte a preservar) deberemos llegar a una solución de compromiso. Existen mucha literatura especializada al respecto, si bien, por su claridad y solvencia científica recomendamos la guía rápida elaborada por el IPI (Image Permanence Institute): IPI Media Storage Quick Reference.

Del hecho de que cada soporte tenga unas condiciones de humedad y temperatura idóneas para su almacenamiento, es fácil deducir que la esperanza de vida de un determinado soporte variará en función de las condiciones finalmente seleccionadas. Esta idea es precisamente la que Jones y Beagrie transmitían en la tabla que figura a continuación, y que queda recogida en el manual The Preservation Management of Digital Materials: the handbook (2001):

Device	25RH 10°C	30RH 15°C	40RH 20°C	50RH 25°C	50RH 28°C
D3 magnetic tape	50 years	25 years	15 years	3 years	1 year
DLT magnetic tapecartridge	75 years	40 years	15 years	3 years	1 year
CD/DVD	75 years	40 years	20 years	10 years	2 years
CD-ROM	30 years	15 years	3 years	9 months	3 months

Ilustración 13: Longevidad de soportes digitales en función de las condiciones de almacenamiento (humedad relativa y temperatura). Jones y Beagrie, 2001

1.1.6 Autenticidad y propiedades significativas de los DLOs

En materia de preservación digital no basta con preservar los datos, los soportes que los contienen, los software que interpretan los formatos en los que se codifican esos datos, y/o los hardware que soportan ese software, además hay que tener en cuenta su **autenticidad**. La UNESCO en sus Guidelines for the Preservation of Digital Heritage (2003) habla además de la importancia de asegurar la autenticidad de la información

digital que preservamos. La autenticidad de un objeto se desprende del grado de confianza con el que podemos asegurar que un DLO es realmente lo que pensamos que es, y tendría dos componentes:

- **identidad**, cualidad por la que un DLO es lo que dice ser y por la que no puede confundirse con ningún otro.
- **integridad**, cualidad por la que sabemos que un DLO no ha sido modificado hasta el punto de que pueda variar su significado

Ilustración 14: En preservación digital debemos no sólo garantizar el acceso a los datos. Sino que además debemos garantizar la autenticidad de los datos (ilustrador: Joe)

La autenticidad puede verse comprometida ya sea por acciones malévolas, negligentes o fortuitas; o por una falta de control de las versiones de un DLO que nos impide saber cuál es la que tiene valor por ser la definitiva y validada (por ejemplo, un documento de procedimientos, durante su preparación contará con multitud de versiones, siendo auténtica únicamente la última versión).

A la hora de preservar un DLO deberemos analizar qué merece la pena, o dicho de otro modo, qué es lo que precisamos preservar de él realmente, para garantizar su accesibilidad a largo plazo. Como ya hemos visto todo DLO tiene sus propiedades, de las que destacaremos y ejemplificaremos las siguientes:

- **Contenido** = información que transmite el documento ej: el contenido intelectual que intenta transmitir este módulo3

- **Contexto** = entorno físico/intelectual del DLO que permite valorar su utilidad, audiencia…ej: saber que este documento se concibió para el

módulo 3 del curso sobre uso de metadatos en biblioteconomía, permitiría valorar la utilidad de este DLO para una determinada audiencia,

- **Estructura** = la organización del contenido ej: cada uno de los epígrafes en los que está dividido este módulo

- **Aspecto** = la presentación del DLO: colores, imágenes, tipo de letra ej: el documento de este módulo 3 se ha redactado en Microsoft office para Macintosh, y se ha utilizado la tipografía arial con un cuerpo 11… En él se han intercalado figuras, tablas para ilustrar su contenido y/o amenizarlo.

- **Comportamiento** = experiencia del usuario frente al DLO ej: no será lo mismo si se visualiza este documento desde un ordenador, donde el estudiante podrá acceder a los enlaces incluidos, que si se almacena la misma información en un documento de texto plano sin hipervínculos.

A partir de esta enumeración de propiedades, podemos hablar de las propiedades significativas (*significant properties*) o elementos esenciales de un DLO, que no son otras sino las merecedoras de preservación digital. Estas no son constantes, y debe priorizarse su conservación en función:

- la comunidad designada de usuarios que acceden y van a acceder a dicha información
- la capacidad económica y técnica de la institución que pretende preservar el DLO
- y demás recursos de los que disponga dicha institución

1.2. Estrategias de preservación

Dependiendo de la literatura consultada, la clasificación de las estrategias planteadas para hacer frente a la preservación digital puede variar ligeramente, solapándose, e incluso produciendo algo de confusión entre algunas de las categorías. La clasificación que aquí presentamos, es un extracto de lo que puede encontrarse en el manual de la Digital Preservation Coalition (ya mencionado anteriormente).

A grandes rasgos pueden hablarse de estrategias primarias y estrategias secundarias:

- **Primarias,** son aquellas adoptadas por repositorios a medio-largo plazo para materiales digitales sobre los que asumen la responsabilidad de preservación. Son fundamentalmente:
  - migración
  - emulación.
- **Secundarias,** son aquellas adoptadas por repositorios a corto-medio plazo ó sobre aquellos materiales con un interés más transitorio. Son estrategias que:
  - que pueden preceder a las primarias
  - que pueden fortalecer a las primarias

  no implican necesariamente inferioridad, es más, la utilización de medidas preventivas adecuadas (ej: almacenamiento, mantenimiento) junto con algunas de estas estrategias secundarias (ej: adhesión a estándares) pueden reducir la necesidad de aplicar medidas primarias (ej: migración)

  Son fundamentalmente las siguientes:

  - Back up
  - Refreshing
  - Utilización de soportes duraderos
  - Preservación de tecnología
  - Adhesión a estándares
  - Compatibilidad retrospectiva
  - Encapsulación
  - Utilización de identificadores permanentes
  - Conversión a formatos analógicos estables
  - Arqueología digital

No vamos a detallar nada sobre estas estrategias pero si alguna no sabéis en qué consiste os recomendamos acudir al manual citado. En cualquier caso, queríamos mencionar este tema ya que cada uno de estos procesos podría constituir un evento al

que sometemos un objeto digital; evento sobre el que querremos almacenar información en el modelo de metadatos que hayamos elegido utilizar (METS, PREMIS…).

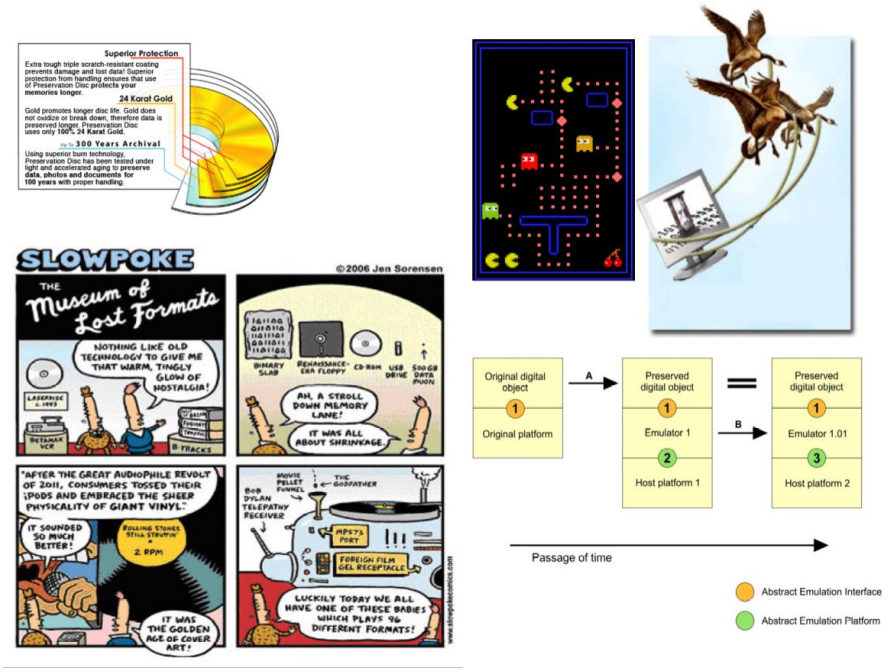

Ilustración 15: Figuras relacionadas con algunas de las estrategias de preservación que existen y/o se pueden adoptar, generalmente de manera complementaria (de arriba abajo y de izquierda a derecha: uso de soportes duraderos como el CD de oro, emulación de entornos,

1.3. Control de formatos

No queríamos abordar el tema de los metadatos sin daros unas mínimas nociones sobre el control de formatos; y es que en la preservación digital de un DLO es básico:

- Verificar su procedencia
- Verificar su autenticidad (identidad e integridad)
- Verificar que sus características técnicas concuerdan con nuestros requisitos (ej: resolución, profundidad de bits…)
- Verificar que se encuentra libre de virus

- Identificar el formato[16] de los DLOs

Sólo conociendo las características y la problemática técnica que pueden tener los archivos, se podrán poner en marcha políticas de preservación consecuentes (Serra, 2008).

En lo que respecta a la identificación de formatos debe decirse que en general se tiende a la **normalización**, esto facilita la gestión de los DLOs, además de que los sistemas de preservación digital no aceptan todos los formatos. Esto es algo que puede suponer un reto en el archivo de la web, en las recolecciones de los dominios pueden aparecer hasta más de 300 formatos de archivos susceptibles de ser conservados. La identificación de formatos es imprescindible, sólo así podremos tomar decisiones correctas respecto a medidas de preservación digital, como pueden ser la **migración**. En nuestras colecciones digitales, tendemos a manejar grandes volúmenes de DLOs y, tanto su almacenamiento, como su gestión se ve facilitada si se hace en función de formatos y tamaños.

Una vez se ha identificado un DLO, toda la información que se deriva de las verificaciones debe quedar registrada como metainformación. La forma más aceptada hoy día para registrar esta información sería el esquema de metadatos **METS** (mantenido por la Library of Congress). La finalidad de este esquema es básicamente registrar en XML:

- todos los metadatos técnicos del documento (identificación, ubicación de los ficheros que lo forman, metadatos EXIF)

- relaciones entre ficheros

- dependencias entre documentos

- descripción del contenido (por ejemplo, la catalogación)

- condiciones legales de uso

[16] Recordamos aquí el concepto de formato: principios estructurales y organizativos de acuerdo a los que se presentan los DLOs. Dichos principios se recogen en las especificaciones.

La implantación de planes de preservación lógicamente prevé que se tengan que poner en marcha **medidas correctoras,** como podría ser por ejemplo, el cambio de un formato de archivo por otro. Esas medidas correctoras también deben de quedar registradas, pero esta vez en otro modelo de esquema de metadatos: PREMIS[17].

En 2005 el proyecto AIHT (Archiving, Ingest and Handling Test) financiado por la Library of Congress, exploró la problemática de la transferencia de archivos entre distintos repositorios, una tarea común dentro de los planes de preservación digital. En este estudio se observó que se dedicaba mucho tiempo en identificar los ficheros, pero no sólo eso, además muchos de esos ficheros presentaban formatos defectuosos. Las alternativas ante estos formatos erróneos sería: bien la corrección manual/semimanual; o bien el descarte de los mismos y la decisión de no preservarlos. Ninguna de las alternativas es muy halagüeña y, en último término, lo que están logrando es que se vea dificultada la automatización de las transferencias de archivos (o la carga de un sistema de preservación digital).

Existen varias iniciativas que se han ido sucediendo y que tienen como fin la identificación rápida y segura del formato de un fichero:

1. **Extensión** Son las letras estandarizadas que figuran al final del nombre de un archivo siguiendo a un punto …
   Es la forma más común de identificar los formatos y la que han utilizado siempre los sistemas operativos más extendidos (Macintosh, Windows)

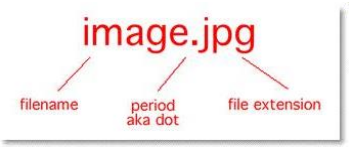

Ilustración 16: Extensión de un archivo

   El problema es que es poco fiable, fácilmente manipulable y poco informativo. Además no distingue versiones de los formatos
   Ej: .jpeg, .tiff, .pdf

[17] En los puntos siguientes se detallarán conceptos básicos sobre PREMIS y otros modelos de metadatos, además del modelo de referencia por excelencia que es OAIS. De momento, mencionar que la LC también menciona la posibilidad de incluir los archivos PREMIS en el archivo de metadatos METS.

Hay algunas páginas web que de manera más o menos sencilla permiten saber a qué formato nos enfrentamos a partir de la extensión del archivo (filextensions.org, FileInfo.com, PC Pitstop…).

Ilustración 17: Ejemplos de servicios gratuitos para la identificación de formatos a través de la extensión de un archivo.

2. ***Magic numbers***

   Es un código estándar que figura en la cabecera del fichero y que indica el tipo de formato de que se trata. Nació en el entorno de Unix.

   Es muy fiable y fácilmente transportable (está incluido en el propio fichero). Sin embargo, sólo puede utilizarse en ficheros binarios, no sirve, por ejemplo, para ficheros textuales

   Ej: “GIF89a” = 0x474946383961
   “GIF87a”= 0x474946383761

3. **DTD ó Declaraciones de Tipo de Documento**

   Son documentos en los que se describe con todo detalle el formato de los datos, es el más informativo y son bastante fiables (aunque en la red pueden encontrarse alguno erróneos)
   Ej:en un fichero SVG en su cabecera se recogerá:
   <!DOCTYPE svg PUBLIC “-//W3C//DTD SVG 1.1//EN” “http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd“>

4. **Metadatos externos al fichero**

   Casos en los que la información sobre el formato del fichero se da en otro independiente y con el que está vinculado. Esto plantea problemas de transportabilidad, puesto que no debe perderse nunca esta vinculación.

No basta con identificar el formato, también deberá existir una forma de poder referirse externamente a los formatos. Esta referencia permite que los ficheros interactúen con los sistemas operativos y las aplicaciones. Además, para que esta relación se pueda automatizar deberán ser métodos de referencia estandarizados:

1. **MIME Types** (Media Types), es el medio utilizado tradicionalmente y que surgieron como una extensión del correo electrónico. Son capaces de indicar una gran familia de formato (audio-imagen-programa) y un subtipo (texto, zip, mpeg…). Sin embargo, han quedado cortos a la hora de especificar **versiones** y **subformatos**.
   Es el indicador de formato en el modelo de metadatos Dublin Core (DC).
   Ej: para GIF su MIME type está definido por la RFC20245 y RFC2046, sin embargo no es posible especificar sus versiones (89A, 87A)
2. **PUID (Pronom Unique Identifier)**
   Es el identificador que el directorio de formatos PRONOM ha dado a cada subformato. El inconveniente es que no todos los formatos tienen un PUID asignado
   Ej: fmt/3 = GIF 1987A
   fmt/4 = GIF 1989a

En cualquier caso, cuando gestionamos DLOs en sistemas de preservación digital (y en realidad en cualquier otro sistema), no basta con identificar el formato de que se trata. Deberemos saber también si ese fichero cumple las normas técnicas definidas para el formato. Para ello las distintas herramientas que hay disponibles manejan las Document Type Definition (DTD) de los formatos redactadas en XML. La utilización de este lenguaje de marcado en este tipo de documentos permite definir:

- el tipo de datos permitidos para un elemento
- la obligatoriedad ó no de algún elemento
- los atributos
- el rango de valores

La mayoría de los sistemas de metadatos relacionados con la preservación digital son desarrollos XML. De manera que analizando esos XML pueden concluirse dos cosas sobre un fichero con un determinado formato:

- que está **bien formado**, cuando se cumplen las reglas de sintaxis definidas por el XML

- que es **válido**, si además se adecúa a una DTD en concreto

Sólo los documentos bien formados y válidos, facilitan la manipulación y conversión, labores fundamentales en el contexto de la preservación digital. Así pues han surgido varias herramientas cuya finalidad es verificar estos documentos XML (JHOVE, DROID, XENA…).

Ante la ingente variedad de formatos que se suceden en el mundo digital, una de las primeras necesidades que surgió para hacer frente a su gestión fue la creación de un **registro universal**, que recogiera información de todos ellos. En 2003 los Archivos Nacionales del Reino Unido crearon el primer registro, conocido como PRONOM. A cada uno de los registros se le asigna un nº o **PUID**, que permite distinguir las versiones de los formatos y los subformatos. Presenta carencias en el ámbito del vídeo y del audio, y "únicamente" acoge 150 formatos distintos. En 2008 surgió otra iniciativa de la Biblioteca de la Universidad de Harvard, el NARA y la OCLC, con financiación de la Andrew W. Mellon Foundation. Se trataba del Global Digital Format Registry (GDFR). En abril de 2009 PRONOM y GDFR, se unieron con el fin de constituir el Universal Digital Format Registry (UDFR). El UDFR se presentó en la asamblea general del "International Internet Preservation Consortium", celebrada en la Library of Congress, en abril y mayo 2012. Al contrario que PRONOM (que se trata de una base de datos relacional) UDFR expresa la información de forma semántica usando la tecnología OntoWiki. Otras aplicaciones externas pueden interrogar la base de datos y exportar su contenido.

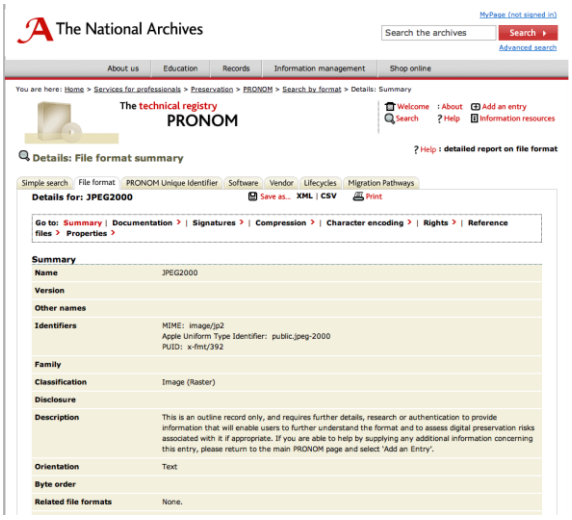

Ilustración 18: vista del registro de jpeg2000 en Pronom

Todas las herramientas utilizadas para el control de formatos en el campo de la preservación digital se basan en estos registros. Las funciones que suelen desempeñar estas herramientas son:

- **Identificación**, dado un fichero determinan de qué formato se trata
- **Validación**, dados unos ficheros y unos metadatos, confirman que hay coherencia entre ellos
- **Caracterización**, obtiene metadatos técnicos descriptivos necesarios con el fin de conocer las propiedades más importantes de un determinado archivo
- **Normalización**, proceden a la conversión de un formato a otro aceptado en función de unos parámetros definidos en una política determinada.

A continuación recogemos sólo algunas características de las herramientas de control de formatos más extendidas:

DROID (Digital Record object Identification)

Desarrollado por lo Archivos Nacionales del Reino Unido.

Sólo identifica según el PUID.

Exporta los resultados en fichero .csv

A 14 de febrero de 2011 identificaba 150 formatos.

XENA (XML Electronic Normalising for Archives)

Desarrollado por los Archivos Nacionales de Australia

Identifica y normaliza formatos abiertos ej: open office

Se obtiene como resultado un fichero xml con contenido binario

TrID

Es comercial, aunque gratuito para uso personal.

Se basa en los *magic numbers*

Reconoce más de 3000 formatos, pero las respuestas no son 100% fiables

JHOVE/JHOVE2 (JSTOR/Harvard Object Validation Environment)

Identifica, valida y caracteriza los archivos

Está escrito en java, es modular y extensible

Muy robusto y completo

De momento no trabaja con muchos formatos (txt, jpeg, tiff, pdf…). El Jhove 2, permitirá separar las funciones de identificación, validación

Las distintas herramientas que van surgiendo pueden solaparse en cuanto a funcionalidad, el decantarse por el uso de una u otra será una decisión que deberá documentarse y hacerse según nuestras necesidades.

A modo de ejemplo, en la siguiente tabla se recoge una comparativa entre tres de estas herramientas (DROID, JHOVE y la desarrollada por la Biblioteca Nacional de Nueva Zelanda), tras la realización de algunas pruebas con una batería de formatos realizadas por una empresa canadiense centrada en soluciones *open source* de preservación digital (www.artefactual.com).

File	Description	DROID	NLNZ	JHOVE	C
0239.mpg	MPEG-2 video file	• Identified format? NO • Identified version? NO • DROID: 0239.mpg	• Identified format? YES • Identified version? NO • NLNZ: 0239.mpg	• Identified format? NO • Identified version? NO • JHOVE: 0239.mpg	
2001-04-LegendsPBP_512kb.mp4	MPEG-4 video file	• Identified format? NO • Identified version? NO • DROID: 2001-04-LegendsPBP _512kb.mp4	• Identified format? NO • Identified version? NO • NLNZ: 2001-04-LegendsPBP_512kb.mp4	• Identified format? NO • Identified version? NO • JHOVE: 2001-04-LegendsPBP_512kb.mp4	
A08917.TIF	TIFF image file	• Identified format? YES • Identified version? NO • DROID: A08917.TIF	• Identified format? YES • Identified version? YES • NLNZ: A08917.TIF	• Identified format? YES • Identified version? YES • JHOVE: A08917.TIF	
accelerando.com	Website consisting of html, xhtml, css, jpg, gif and png files	• Identified formats? YES • Identified versions? PARTIALLY • DROID: accelerando.com	• Identified formats? PARTIALLY • Identified versions? PARTIALLY • NLNZ: accelerando.com	• Identified formats? PARTIALLY • Identified versions? PARTIALLY • JHOVE: accelerando.com	
artefactual.com.zip	Zip archive containing html, css, jpg, gif and png files	• Identified format? YES • Identified version? NO • DROID: artefactual.com.zip	• Identified format? YES • Identified version? NO • NLNZ: artefactual.com.zip	• Identified format? NO • Identified version? NO • JHOVE: artefactual.com.zip	
Basic search.odt	OpenOffice.org Writer file with inserted png file.	• Identified format? NO • Identified version? NO • DROID: Basic search.odt	• Identified format? YES • Identified version? YES • NLNZ: Basic search.odt	• Identified format? NO • Identified version? NO • JHOVE: Basic search.odt	
ct000654.jp2	JPEG2000 image file	• Identified format? YES • Identified version? N/A • DROID: ct00654.jp2	• Identified format? NO • Identified version? N/A • NLNZ: ct000654.jp2	• Identified format? YES • Identified version? N/A • JHOVE: ct000654.jp2	
DadClip_64kb.mp3	MPEG-1 Audio Layer-3 file	• Identified format? YES • Identified version? YES • DROID: DadClip_64kb.mp3	• Identified format? YES • Identified version? YES • NLNZ: DadClip_64kb.mp3	• Identified format? NO • Identified version? NO • JHOVE: DadClip_64kb.mp3	
DemoANSI.txt	US_ASCII text file	• Identified format? NO • Identified version? NO • DROID: DemoANSI.txt	• Identified format? YES • Identified version? NO • NLNZ: DemoANSI.txt	• Identified format? YES • Identified version? YES • JHOVE: DemoANSI.txt	
DemoUTF-8.txt	Unicode text file	• Identified format? NO • Identified version? NO • DROID: DemoUTF-8.txt	• Identified format? YES • Identified version? NO • NLNZ: DemoUTF-8.txt	• Identified format? YES • Identified version? YES • JHOVE: DemoUTF-8.txt	
Free_to_use_10_sec.aif	AIF audio file	• Identified format? YES • Identified version? N/A • DROID: Free_to_use_10_sec.aif	• Identified format? YES • Identified version? N/A • NLNZ: Free_to_use_10_sec.aif	• Identified format? YES • Identified version? N/A • JHOVE: Free_to_use_10_sec.aif	
Holdings.png	PNG image file	• Identified format? YES • Identified version? YES • DROID: Holdings.png	• Identified format? YES • Identified version? NO • NLNZ: Holdings.png	• Identified format? NO • Identified version? NO • JHOVE: Holdings.png	
Ica-atom-technical-architecture-2008-06.gif	GIF image file	• Identified format? YES • Identified version? YES • DROID: Ica-atom-technical-architecture-2008-06.gif	• Identified format? YES • Identified version? YES • NLNZ: Ica-atom-technical-architecture- 2008-06.gif	• Identified format? YES • Identified version? YES • JHOVE: Ica-atom-technical-architecture- 2008-06.gif	
Ica-atom-technical-architecture-2008-06 (copy).jpg	Corrupted GIF image file (file extension changed to jpg)	• Identified format? YES • Identified version? YES • DROID: Ica-atom-technical-architecture-2008-06 (copy).jpg	• Identified format? YES • Identified version? YES • NLNZ: Ica-atom-technical- architecture-2008-06 (copy).jpg	• Identified format? YES • Identified version? YES • JHOVE: Ica-atom-technical- architecture-2008-06 (copy).jpg	
inkscape_wallpaper___blue_ by_ryanlerch.svg	SVG image file	• Identified format? YES • Identified version? YES • DROID: inkscape_wallpaper___blue_ by_ryanlerch.svg	• Identified format? NO • Identified version? NO • NLNZ: inkscape_wallpaper___ blue_by_ryanlerch.svg	• Identified format? NO • Identified version? NO • JHOVE: inkscape_wallpaper___ blue_by_ryanlerch.svg	

Ilustración 19: Resultados de validación, identificación y caracterización con DROID, NLNZ y JHOVE (consultable en: https://wiki.artefactual.com/wiki/Test_File_Results)

En el campo de la preservación digital no dejan de surgir nuevas herramientas de software, y no siempre su instalación/utilización resultan intuitivas. El grupo Preserva digital, formado por docentes e investigadores de la Universidad de Barcelona, ha preparado una serie de tutoriales que pueden ser de gran utilidad para acercarse a algunas de estos complementos: http://bd.ub.edu/preservadigital/tutoriales[18]

1.4. Modelo Open Archival Information System (OAIS)

Como ya hemos explicado la preservación digital es un tema que afecta a todos los sectores; para muestra el caso de la NASA. Y es que la Agencia Espacial se ha visto enfrentada a la pesadilla de muchos: la pérdida de datos en formato digital (concretamente las cintas magnéticas con datos recogidos por las sondas *Viking* en los 70). Si estos datos no hubieran existido en formato impreso, por ejemplo, en un estudio realizado en los años 90 no se hubiesen podido refrendar las evidencias de vida en Marte. Ante un problema de esta magnitud la NASA respondió desarrollando un modelo teórico en el que se integrasen y explicasen los requisitos que debería cumplir cualquier sistema de preservación.

El modelo fue discutido y aprobado dentro del *Council of the Consultative Committee for Space Data Systems* (CCSDS), el organismo encargado de desarrollar los estándares de datos de las principales agencias del espacio a nivel mundial, tomando el nombre de *Reference Model for an Open Archival Information System* (OAIS) y siendo publicado en enero de 2003. OAIS se convirtió en norma ISO el año siguiente con el código 14721:2003.

Este sistema OAIS es un archivo, compuesto por una organización de personas y sistemas que han aceptado la responsabilidad de preservar la información y hacerla disponible para una **comunidad designada de usuarios.** Implica una serie de responsabilidades que hacen que este "archivo" sea diferente de otros tipos de archivo. El término "abierto" se utiliza en el sentido de que estas recomendaciones y estándares

[18] Todos estos manuales están optimizados para Windows 7.

se desarrollen en foros abiertos, no que el acceso al archivo sea sin restricción. Está pensado para la Preservación digital a largo plazo; y aunque puede servir para un archivo físico está centrado en información digital.

Este modelo de referencia establece una serie de entidades a las que asigna determinadas funciones relacionadas con la preservación del archivo de información. Entre dichas se entidades incluyen la "ingesta", el almacenamiento, la gestión de datos, el acceso y la difusión.

Establece un conjunto mínimo de responsabilidades para que un archivo pueda ser llamado OAIS; aquel que pretende preservar la información para el acceso y uso de una Comunidad Designada de usuarios.

Para que un **Objeto de información** sea preservado correctamente, es crítico que el OAIS identifique y comprenda el **objeto de datos** y la **información de representación** asociada. En lo que a información digital se refiere esto significa que el OAIS debe identificar claramente los bits y la representación de información que se aplica a estos bits.

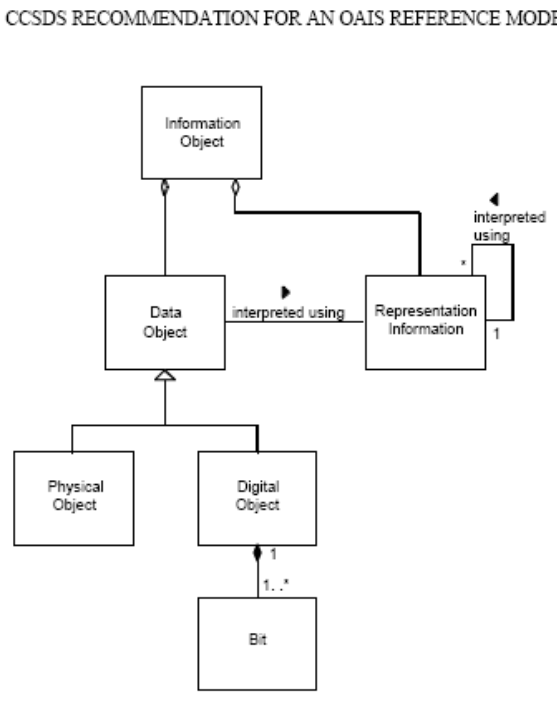

Ilustración 20: El objeto de información según OAIS

El Objeto digital, tal y como se describe en la figura anterior compone de una o más secuencias de bits. El objetivo de la IR es convertir la secuencia de bits en una información más significativa. Lo hace describiendo el formato o la estructura de los datos, conceptos que aplicados a la secuencia de bits lo convierten en valores más significativos como caracteres, números, píxeles, tablas, etc. Estos tipos de datos, las agregaciones de estos datos, y las reglas que mapean los tipos de datos subyacentes con conceptos más abstractos necesarios para entender el Objeto Digital se conocen como **información de la estructura** de la IR.

Sin embargo, esta información de la estructura casi nunca es suficiente; también es necesario acompañar a la IR de **información semántica**, como la lengua en la que esa información está escrita.

Por último, es posible que la IR contenga referencias a otras IR. Si a esto unimos el hecho de la IR es también un Objeto de Información que puede tener su propio objeto digital y otras IRs asociadas, podemos tener como resultado una **Red de Representaciones**.

Como puede verse son muchos tipos de información los implicados en la preservación a largo plazo dentro del modelo OAIS. Cada uno de estos tipos puede verse como un Objeto de Información en la medida en que contiene un objeto de datos y una IR para comprender los datos. Los diferentes tipos de información del OAIS son:

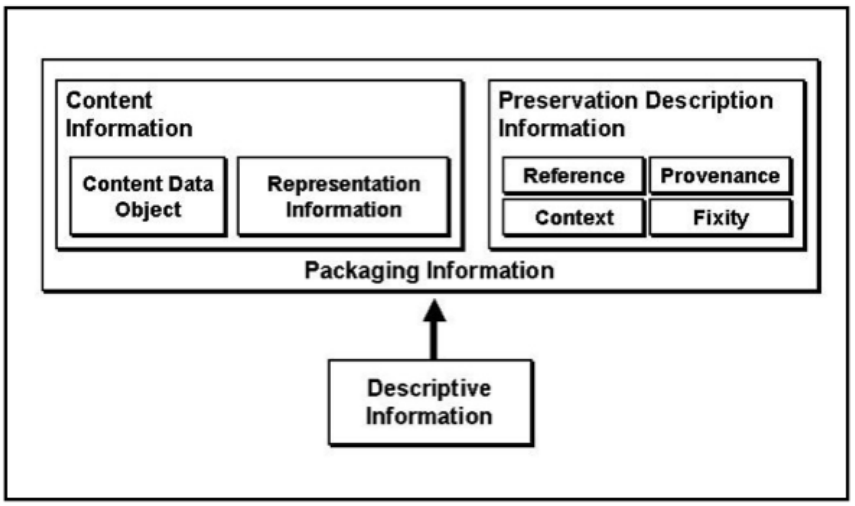

Ilustración 21: Tipos de información identificados en OAIS

- **Información de contenido**: Es el conjunto de información que originalmente se quiere preservar. Decidir qué es Información de contenido y qué no, puede no ser tan evidente y puede ser necesario negociarlo con el Productor. La información de contenido, que es un Objeto de Información, es el objeto contenedor de datos con su IR.
- **Información descriptiva de Preservación (PDI):**Además de la información de contenido el Archivo debe incluir información que permitirá la comprensión de esta información de contenido durante un tiempo ilimitado. Los objetos de información que permiten esto es la PDI. Está centrada específicamente en describir el estado pasado y presente de la información de contenido, asegurando que es identificable unívocamente y que no ha sido alterada sin conocimiento. La información que debe recoger en principio se puede dividir en:
    - **Información de referencia:** Identifica y en su caso, describe, uno o más mecanismos utilizados para asignar identificadores a la Información de Contenido. También provee los identificadores necesarios para que los sistemas externos se refieran a esta Información de Contenido.
    - **Información del contexto**: Documenta las relaciones entre la Información de contenido y su entorno. Incluye por qué la Información de Contenido fue creada y cómo se relaciona con otros objetos de información de contenido externos.
    - **Procedencia**: Documenta la historia de la Información de Contenido. El origen y cualquier cambio que se haya producido. Puede ser visto como una forma de información del contexto.
    - **Integridad.** Está referida a los mecanismos de autentificación para garantizar que la información contenida no se ha alterado (comprobación, firma digital, etc)

Para hablar de la información que se envía, se gestiona y/o se recibe en un archivo OAIS se maneja además el concepto de **paquete de información**. El modelo reconoce 3 tipos: paquete de información de envío **(SIP: Submission Information Package)**; paquete de información de archivo **(AIP: Archival Information Package**) y paquete de información de difusión **(DIP: Difusion Information Package)**.

El intercambio de estos paquetes de información se produce entre agentes (productor, consumidor) y las entidades que antes hemos mencionado (ingest, gestión de datos, almacenamiento de archivo, planificación de preservación, administración y acceso). En la documentación sobre OAIS también se habla de la migración de la información digital a nuevos medios y soportes, de los modelos de datos para representar la información, del papel del software en la preservación de la información y del intercambio de información entre archivos. En este módulo nos vamos a centrar en los modelos de metadatos más comúnmente relacionados con la preservación digital: METS, PREMIS[19]

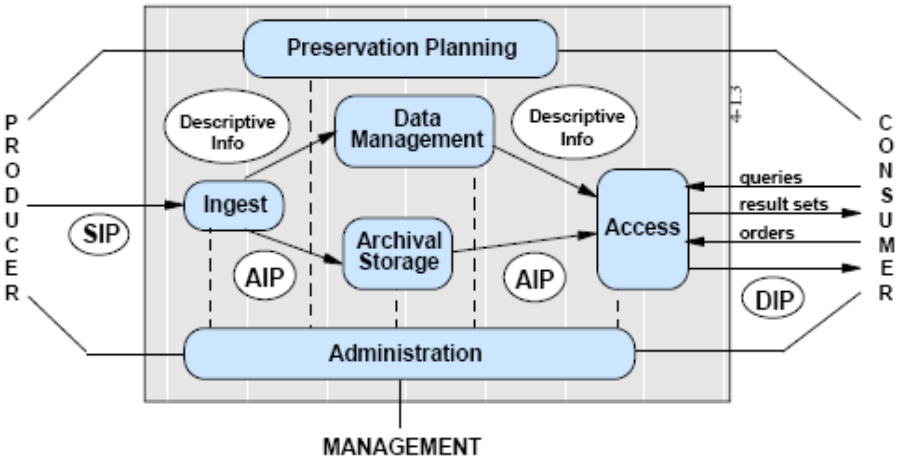

Figure 4-1: OAIS Functional Entities

Ilustración 22: Esquema de alto nivel que recoge los elementos básicos del OAIS

Para una explicación más detallada del modelo OAIS (representado de manera esquemática en el diagrama que os incluimos arriba) lo mejor es manejar la propia norma ISO. Sin embargo, no se debe perder nunca la idea de que se trata de un acercamiento teórico.

[19] Los MIX, ideales para guardar información técnica sobre imágenes digitales, ya los habéis tratado en el módulo 1. En cualquier caso los volvemos a mencionar de manera muy somera más adelante.

1.5. Metadatos en torno a la preservación digital

Desde que en 2005 el PREMIS Data Dictionary for Preservation metadata ganara el premio de Preservación digital Internacional, se ha avanzado mucho en el tema de metadatos de preservación, pasándose de abordar asuntos conceptuales a detalles más relacionados con la implementación. Así lo explican Lavoie y Gartner en la 2ª edición sobre metadatos de preservación editada por la Digital Preservation Coalition[20].

En este informe se abordan los avances experimentados desde que PREMIS se convirtiera en estándar internacional y entre los que incluyen:

- Las revisiones de PREMIS, actualmente en su versión 2.2 (más adelante veremos las características básicas de este modelo y las claves para entender el diccionario)
- La expansión del uso de METS como entorno de almacenamiento y relación con otros tipos de metadatos. Al final del tema veremos brevemente las recomendaciones al respecto de la Library of Congress(LoC) y un ejemplo práctico con las decisiones tomadas al respecto por parte de la BNE
- El desarrollo de herramientas en torno a PREMIS: JHOVE, DROID, PREMIS in METS Toolbox que valide la conformidad conforme a las recomendaciones/*checklist* que propone la Library of Congress
- La creación y mantenimiento de un registro de implementaciones PREMIS; en el que se recogen las funciones asignadas al modelo PREMIS y las entidades del mismo que utilizan, así como la arquitectura de despliegue de los PREMIS (fundamentalmente METS). En este sentido, Lavoie y Gartner, reconocen que ahora que empieza a existir conciencia de uso y conocimiento sobre metadatos de preservación, lo que falta es documentación sobre mejores prácticas, e incluso análisis que evalúen los costes y beneficios derivados del uso de los metadatos de preservación. Es decir, ver cómo efectivamente el

[20] 'Preservation Metadata' Technology Watch Report

utilizar estos modelos ayudan en la toma de decisiones y flujos de trabajo de la preservación digital.

- El Comité Editorial de PREMIS ha promovido la creación de una ontología PREMIS OWL; esto permitirá expresar la semántica del Diccionario de datos en RDF. La LoC ha desarrollado una serie de vocabularios controlados en SKOS (Simple Knowledge Organization System).
- Por último, el proyecto TIPR (Towards Interoperable Preservation Repositories) ha trabajado en un protocolo que permita el intercambio de metadatos de preservación entre repositorios.

Por todo lo dicho hasta ahora deduciréis que hablar de metadatos de preservación es hablar de PREMIS; sin embargo aprovecharemos para dar algunos apuntes básicos sobre un modelos estrechamente relacionados con PREMIS: METS y MIX (en este último caso será apenas un recordatorio ya que lo habéis visto en el módulo 1).

El concepto de metadatos de preservación no encaja 100% en ninguna de las categorías de metadatos habituales (descriptivos, estructurales o administrativos); de hecho más bien se encuentran a caballo de todos ellos. De ahí que Lavoie y Gartner los definan genéricamente como los metadatos que posibilitan la preservación a largo plazo. Con ellos fundamentalmente lo que se trata es de dotar al objeto digital preservado de un marco de información que permanecerá ligado a él a lo largo del tiempo y que abarca:

- Procedencia del objeto, un historial del objeto desde su creación a todos los cambios a los que se somete con fines de preservarlo a largo plazo. Es información que permite fijar la autenticidad[21], integridad[22] del objeto.
- Información sobre la gestión de los derechos del objeto
- Entorno técnico e interpretativo sobre el objeto que permitirá en última instancia acceder a él, representarlo y utilizarlo

[21] cualidad que garantiza que un objeto es “lo que dice ser” y “que no se confunde” con ningún otro

[22] cualidad que permite garantizar que un objeto no ha experimentado ninguna modificación que cambie su significado.

El tema de los metadatos de preservación pasó rápidamente de la teoría a la práctica y en ese lapso de tiempo surgieron muchas iniciativas que se han solapado y retroalimentando. Aunque en este modulo nos centremos en PREMIS y las soluciones más directamente relacionadas con dicho modelo, Lavoie y Gartner resumen así alguna de las otras iniciativas:

- Preservation Metadata for Digital Collections impulsado por la National Library of Australia[23]. Este modelo se diseñó para objetos digitalizados y para los nacidos digitales y admitía tres niveles de granularidad descriptiva (colección, objeto y archivo). No se abarcaba nada sobre estrategias de preservación
- Proyecto CEDARS (CURL Exemplars in Digital Archives)[24] nació como proyecto piloto para un archivo digital y los elementos que incluía podían aplicarse a cualquier nivel de descripción
- Proyecto NEDLIB (Networked European Deposit Library)[25]; en esta iniciativa se definió un conjunto de metadatos de preservación esenciales; hizo especial énfasis en superar la obsolescencia tecnológica. Sus elementos se definieron a un alto nivel para así maximizar su aplicabilidad independientemente de formatos y tipos de objetos.
- El modelo desarrollado por la National Library of New Zealand (NLNZ)[26]; en este caso podría decirse que se trata de un punto de partida para implementar sistemas responsables de la recolección y gestión de metadatos de preservación.

[23] En el siguiente enlace encontraréis más detalles sobre los elementos que incluye este modelo: http://pandora.nla.gov.au/pan/25498/20020625-0000/www.nla.gov.au/preserve/pmeta.html

[24] Modelo consultable en: http://www.webarchive.org.uk/wayback/archive/20050410120000/http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html

[25] http://www.kb.nl/sites/default/files/docs/NEDLIBmetadata.pdf

[26] http://www.natlib.govt.nz/downloads/metaschema-revised.pdf

Los tres primeros modelos eran muy teóricos, el último, aunque tuvo sus bases en estos antecedentes, ya estaba más en línea con la planificación e implementación de sistemas archivo digital.

Podéis leer más a fondo sobre estas iniciativas en los enlaces propuestos.

1.5.1 Metadata Encoding Transmission Standard (METS)

METS es un modelo de metadatos más complejo en comparación con otros modelos como MARCXML, Dublin Core que hacen lo que siempre se ha hecho en una biblioteca describir documentos: dar datos sobre su autoría, título, materia, edición, etc. Como ya habéis visto en el primer módulo METS es una norma que permite expresar metadatos descriptivos, administrativos y estructurales relativos a objetos en una biblioteca digital. Se basa en los esquemas del lenguaje XML. Esta norma la mantiene la Biblioteca del Congreso y está siendo desarrollada como iniciativa de la *Digital Library Federation*.

Como ya os dijimos os recomendamos consultar el tutorial de la Library of Congress[27], un documento que hemos utilizado como guía para este epígrafe.

El modelo de metadatos METS tiene una gran riqueza expresiva y flexibilidad, apoyada además en el lenguaje xml, esto hace que sea el contenedor ideal para recoger información de muy distinta naturaleza y aplicable a las bibliotecas digitales (metadatos descriptivos, técnicos, de derechos, estructurales…). Como dice Guenter[28] esta flexibilidad no sólo permite satisfacer muchas necesidades en cuanto a metadatos se refiere, sino que además implica que su implementación (al igual que ocurre con PREMIS) obliga a tomar una serie de decisiones, no sólo sobre qué esquema de metadatos utilizar, sino además sí incluir o enlazar al mismo y en qué nivel de estructura expresarlo, lo cual depende de la estructura del objeto representado por un documento

[27] http://www.loc.gov/standards/mets/METSOverview_spa.html#MHead

[28] http://www.dlib.org/dlib/july08/guenther/07guenther.html

METS.

Como ya se os mencionó, un documento METS consta de siete secciones.

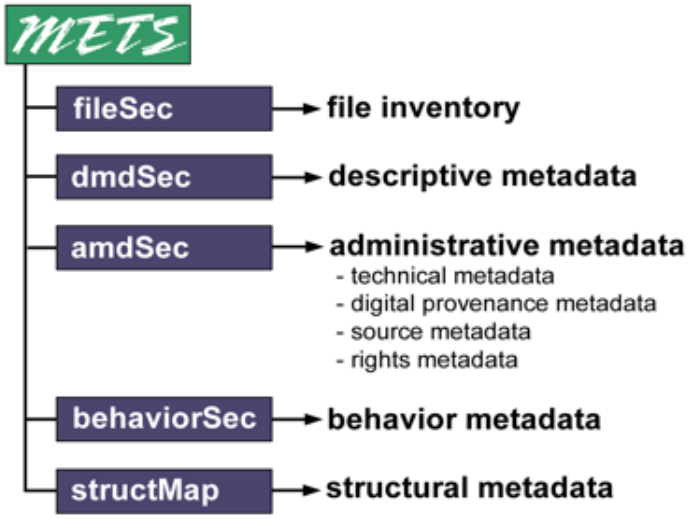

Ilustración 23: Las principales secciones de METS y el tipo de metadatos que recoge cada uno

Las siete secciones de METS permite englobar metadatos con distintas funciones.

A continuación, nos vamos a centrar a continuación en cada una de ellas:

1.- Cabecera METS

El elemento Cabecera METS (METS Header) permite registrar - dentro del propio documento METS - unos mínimos metadatos descriptivos sobre el propio documento METS. Estos metadatos incluyen la fecha de creación del documento METS, fecha de última modificación y estado. También se puede registrar el nombre de uno o más agentes que han desempeñado alguna función en el ciclo de vida del documento METS, especificar dicha función y añadir una breve nota sobre estas actividades. Finalmente, se puede registrar una variedad de identificadores alternativos para el documento METS adicionales al identificador principal.

El siguiente fragmento recoge un ejemplo de Cabecera METS:

```
<metsHdr CREATEDATE="2003-07-04T15:00:00"
RECORDSTATUS="Complete"> <agent ROLE="CREATOR"
TYPE="INDIVIDUAL">

    <name>Jerome McDonough</name>

  </agent>

  <agent ROLE="ARCHIVIST" TYPE="INDIVIDUAL">

    <name>Ann Butler</name>

  </agent>

</metsHdr>
```

En este ejemplo el elemento <metsHdr> contiene dos atributos: CREATEDATE y RECORDSTATUS. Indican respectivamente la fecha y hora en que se creó el documento METS y su estado. Se listan dos agentes que han trabajado en este documento: la persona responsable de su creación y un archivero responsable del material original. Los atributos ROLE y TYPE del elemento <agent> toman sus valores de vocabularios controlados. Los valores permitidos para el atributo ROLE son: "ARCHIVIST," "CREATOR," "CUSTODIAN," "DISSEMINATOR," "EDITOR," "IPOWNER" y "OTHER." Los valores permitidos para el atributo TYPE son: "INDIVIDUAL," "ORGANIZATION" y "OTHER."

2.- Metadatos Descriptivos:

La sección Metadatos Descriptivos consiste en uno o más elementos <dmdSec> (Descriptive Metadata Section). Cada elemento <dmdSec> puede: a) contener un puntero a metadatos externos (elemento <mdRef>); b) contener metadatos internamente (dentro de un elemento <mdWrap>), o c) combinar estas dos opciones. Esto quiere decir, que la parte descriptiva no tiene por qué estar en el propio fichero, se puede indicar una relación con otro fichero externo (por ejemplo un fichero con una descripción en Dublin Core).

Metadatos descriptivos externos (mdRef): un elemento mdRef recoge una URI en la que se pueden recuperar metadatos externos. Por ejemplo, la siguiente referencia apunta a un instrumento de descripción externo para un objeto digital:

<dmdSec ID="dmd001"> <mdRef LOCTYPE="URN" MIMETYPE="application/xml" MDTYPE="EAD" LABEL="Berol Collection Finding Aid">urn:x-nyu:fales1735</mdRef>

</dmdSec>

El elemento <mdRef> de este <dmdSec> contiene cuatro atributos. El atributo LOCTYPE especifica el tipo de localizador que se usa; los valores aceptados para LOCTYPE son: 'URN,' 'URL,' 'PURL,' 'HANDLE,' 'DOI,' y 'OTHER.' El atributo MIMETYPE especifica el tipo MIME de los metadatos descriptivos externos, y MDTYPE a qué tipo de metadatos se hace referencia. Los valores aceptados para MDTYPE incluyen MARC, MODS, EAD, VRA (VRA Core), DC (Dublin Core), NISOIMG (NISO Technical Metadata for Digital Still Images), LC-AV (Library of Congress Audiovisual Metadata) , TEIHDR (TEI Header), DDI (Data Documentation Initiative), FGDC (Federal Geographic Data Committee Metadata Standard [FGDC-STD-001-1998]), y OTHER. LABEL ofrece un mecanismo para describir estos metadatos para aquellas personas que vean el documento METS.

Metadatos descriptivos internos (mdWrap): el elemento mdWrap contiene los metadatos dentro del propio documento METS. Estos metadatos podrán ser:

1. Metadatos codificados en XML, en cuyo caso se indicará que pertenecen a un espacio de nombres distinto de METS, o

2. Metadatos en cualquier otro formato binario o textual (no XML), siempre que los metadatos se codifiquen en Base64 y se escriban dentro de un elemento <binData> contenido dentro del elemento mdWrap.

Los siguientes ejemplos muestran el uso del elemento mdWrap:

<METS:dmdSec ID="DMD2">

<mdWrap MIMETYPE="application/mab" MDTYPE="OTHER" LABEL="MAB Record">

<binData>

00471nM2.01010024 h001 66230 002 19941207000000.0 003 20070608000000.0 030 zz5d||rz||||7 050 ||||||||||||| 051 n|||||| 077 c0 100

Oertel, Christian Gottfried 331 VollstÉandiges corpus gravaminum evangelicorum 359 An das Licht gestellet von Christian Gottfried Oertel 410aRegensburg 412aNeubauer 501 Erschienen: 1 (1771) - [8] (1775). - Bd. [8] im Verl. Montag, Regensburg, erschienen 710 Corpus Evangelicorum / Gravamen 902 |Corpus Evangelicorum 902 |Gravamen

</binData>

</mdWrap>

</METS:dmdSec>

<METS:dmdSec ID="DMD3">

<mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core Metadata">

<xmlData>

<dc:title>Vollständiges corpus gravaminum evangelicorum</dc:title>

<dc:creator>Oertel, Christian Gottfried</dc:creator>

<dc:date>1 (1771) - [8] (1775)</dc:date>

<dc:publisher>Montag, Regensburg</dc:publisher>

<dc:type>text</dc:type>

</xmlData>

</mdWrap> </METS:dmdSec>

3.- Metadatos administrativos

Los elementos <amdSec> contienen los metadatos administrativos correspondientes a los archivos que conforman el objeto digital, y también los del material original a partir del cual se creó la representación digital. En los documentos METS hay cuatro tipos de metadatos administrativos:

- Metadatos técnicos (información relativa a la creación del archivo, su formato y características de uso),
- Metadatos sobre derechos de propiedad intelectual (copyright e información sobre licencias),
- Metadatos sobre el origen (metadatos descriptivos y administrativos sobre

el documento origen a partir del cual se ha generado el objeto digital), y

- Metadatos sobre la procedencia digital (información sobre la relación entre el documento original y su representación digital, incluyendo la relación entre copias maestras y derivadas, migraciones y transformaciones realizadas sobre los archivos desde su digitalización inicial).

Cada uno de estos cuatro tipos de metadatos administrativos tienen un elemento propio dentro de la sección <amdSec>: <techMD>, <rightsMD>, <sourceMD>, y <digiprovMD>. Todos pueden repetirse.

Los elementos <techMD>, <rightsMD>, <sourceMD> y <digiprovMD> tienen el mismo modelo de contenido que <dmdSec>: pueden contener un elemento <mdRef> para apuntar a metadatos administrativos externos, un elemento <mdWrap> para incorporar metadatos administrativos dentro del propio documento METS, o combinar ambas opciones.

4.- Sección Archivo

La sección archivo (<fileSec>) contiene uno o más elementos <fileGrp>. Estos agrupan archivos relacionados entre sí. Un <fileGrp> reúne todos los archivos que conforman una misma versión electrónica del objeto digital. Por ejemplo, puede haber elementos <fileGrp> para las miniaturas, las copias master (alta resolución) de las imágenes, la versión en pdf, etc.

El siguiente ejemplo muestra una sección Archivo para un registro sonoro del que hay tres versiones: una trascripción codificada en TEI, una copia master audio en formato WAV y una versión audio derivada de la anterior en formato MP3:

<fileSec>

<fileGrp ID="VERS1"> <file ID="FILE001" MIMETYPE="application/xml" SIZE="257537"

CREATED="2001-06-10"> <FLocat LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.xml</FLocat>

</file> </fileGrp> <fileGrp ID="VERS2">

<file ID="FILE002" MIMETYPE="audio/wav" SIZE="64232836" CREATED="2001-

05-17" GROUPID="AUDIO1"> <FLocat LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.wav</FLocat>

</file> </fileGrp> <fileGrp ID="VERS3" VERSDATE="2001-05-18">

<file ID="FILE003" MIMETYPE="audio/mpeg" SIZE="8238866" CREATED="2001-05-18" GROUPID="AUDIO1"> <FLocat LOCTYPE="URL">http://dlib.nyu.edu/tamwag/beame.mp3</FLocat>

</file> </fileGrp>

</fileSec>

En este caso, <fileSec> contiene tres elementos <fileGrp>, uno para cada versión del objeto. El primero es una trascripción codificada en XML, el segundo es una versión audio en formato WAV, y el tercero una versión audio en formato MP3. Aunque en este ejemplo puede no parecer necesario utilizar elementos <fileGrp> para las distintas versiones del objeto, <fileGrp> sería mucho más útil si el objeto consistiese en un gran número de imágenes escaneadas, o si cada versión del objeto constase de un mayor número de archivos. En estos casos, ser capaz de agrupar los elementos <file> en distintos <fileGrp> facilita la identificación de los archivos que pertenecen a cada versión.

En definitiva, esta parte de un documento METS es algo así como un inventario de todos los objetos que componen el objeto digital. Un ejemplo sencillo podría ser imaginar el inventario de todas las fotografías que componen un álbum, o cada uno de los volúmenes de una obra multivolumen (como una enciclopedia). Como vamos a ver ahora mismo, en el Mapa estructural se dará el orden y la jerarquía de todos estos objetos.

5.- Mapa estructural

La sección Mapa Estructural de un documento METS define una estructura jerárquica que puede presentarse a los usuarios para navegar a través del objeto digital. El elemento <structMap> establece esta jerarquía como una serie de elementos anidados. Cada cuenta con atributos que especifican de qué tipo de división se trata; también puede contener múltiples punteros METS (<mptr>) y punteros a archivos (<fptr>) para identificar los contenidos correspondientes a esa sección. Los punteros METS apuntan a documentos METS aparte que contienen la información sobre los archivos relevantes para la sección . Son útiles cuando se codifican grandes

colecciones de materiales (por ejemplo, una revista completa) y se quiere mantener el tamaño de cada documento METS relativamente pequeño. Los punteros a archivos indican qué archivos (o en ciertos casos, qué grupos de archivos o partes de un archivo) previamente declarados en la sección <fileSec> del documento METS se corresponden con la sección representada por el elemento .

A continuación se muestra un ejemplo de mapa estructural:

```
<structMap TYPE="logical">  <div ID="div1" LABEL="Oral History: Mayor
Abraham Beame"

TYPE="oral history">  <div ID="div1.1" LABEL="Interviewer Introduction"

ORDER="1">  <fptr FILEID="FILE001">

<area FILEID="FILE001" BEGIN="INTVWBG" END="INTVWND"

BETYPE="IDREF" /> </fptr>

<fptr FILEID="FILE002">  <area FILEID="FILE002" BEGIN="00:00:00"
END="00:01:47"

BETYPE="TIME" /> </fptr>

<fptr FILEID="FILE003">  <area FILEID="FILE003" BEGIN="00:00:00"
END="00:01:47"

BETYPE="TIME" /> </fptr>

</div>  <div ID="div1.2" LABEL="Family History" ORDER="2"> <fptr
FILEID="FILE001">

<area FILEID="FILE001" BEGIN="FHBG" END="FHND" BETYPE="IDREF" />

</fptr>  <fptr FILEID="FILE002">

<area FILEID="FILE002" BEGIN="00:01:48"END="00:06:17" BETYPE="TIME" />

</fptr>  <fptr FILEID="FILE003">

<area FILEID="FILE003" BEGIN="00:01:48" END="00:06:17" BETYPE="TIME" />

</fptr> </div>

<div ID="div1.3" LABEL="Introduction to Teachers' Union" ORDER="3">
```

<fptr FILEID="FILE001"> <area FILEID="FILE001" BEGIN="TUBG" END="TUND"

BETYPE="IDREF" /> </fptr>

<fptr FILEID="FILE002"> <area FILEID="FILE002" BEGIN="00:06:18" END="00:10:03"

BETYPE="TIME" /> </fptr>

<fptr FILEID="FILE003"> <area FILEID="FILE003" BEGIN="00:06:18" END="00:10:03"

BETYPE="TIME" /> </fptr>

  </structMap>

Este mapa estructural se corresponde con un registro sonoro (una entrevista al Alcalde Abraham Beame de la ciudad de Nueva York) e incluye tres subsecciones: una introducción por parte del entrevistador, la historia familiar por parte del Alcalde Beame, y una discusión de cómo llegó a participar en el sindicato de maestros de Nueva York. Cada una de estas subsecciones o divisiones está enlazada con tres archivos (los mismos que usamos en el ejemplo anterior): una transcripción XML, un archivo audio master y uno correspondiente a una versión derivada. El elemento hijo <area> se usa en cada <fptr> para indicar que la división se corresponde únicamente con una parte del archivo al que se hace referencia, y con él se identifica la parte exacta del archivo. Por ejemplo, la primera división (la introducción por parte del entrevistador) está enlazada a un fragmento del archivo con la transcripción XML (FILE001) que se encuentra entre las dos etiquetas del archivo cuyos atributos ID recogen los valores "INTVWBG" y "INTVWND". También está enlazado a los dos archivos de audio; en esos casos, en lugar de especificar valores del atributo ID para acotar el fragmento, su inicio y fin se indica en forma de tiempo HH:MM:SS. Así, la introducción del entrevistador se puede encontrar en los archivos de audio en los fragmentos que comienzan en el instante 00:00:00 y que tienen una duración de 00:01:47.

Siguiendo con el ejemplo que hemos puesto antes de una enciclopedia: En la parte descriptiva del documento METS tendíamos el título de la obra, sus datos de edición, etc. En la Sección de Archivo tendríamos la relación de todos los volúmenes que forman la obra completa. En el Mapa Estructural se indicaría el orden

correspondiente a cada uno de esos volúmenes.

El resultado final de esto se puede ver en algo como lo siguiente (tomado de la Biblioteca Digital Hispánica de la Biblioteca Nacional de España):

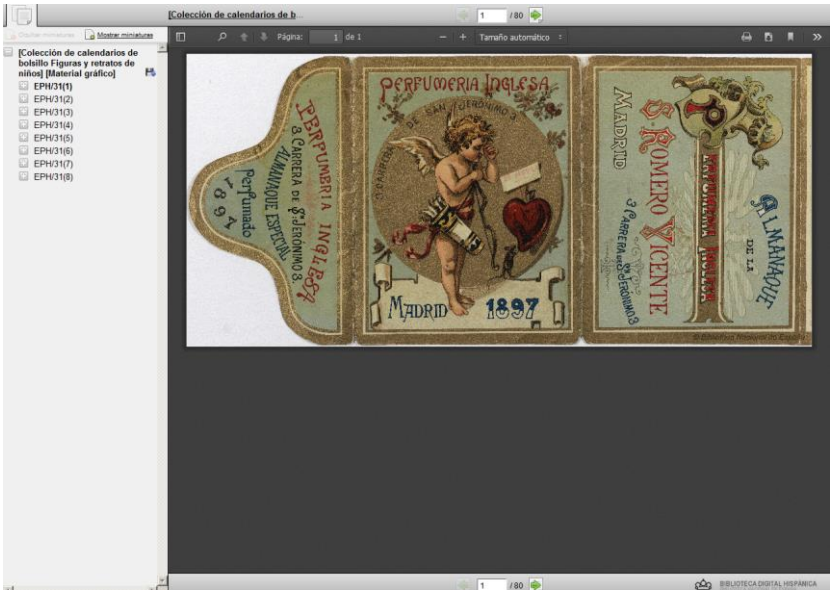

Ilustración 24: Ejemplo de un archivo METS en el visor de BDH

En este caso se trata de una colección de documentos efímeros (calendarios). Gracias a que la estructura está codificado en METS se puede presentar la obra y a continuación ordenados cada una de las imágenes que componen esa colección de calendarios.

6.- Enlaces estructurales

Permite registrar la existencia de hiperenlaces entre las secciones del mapa estructural. Tiene gran valor cuando se usa METS para archivar sitios web.

7.- Sección Comportamiento

Se puede usar para vincular comportamientos ejecutables con los contenidos del documento METS. Cada comportamiento tiene una definición de interfaz y un "mecanismo" que identifica un módulo de código ejecutable que implementa y ejecuta el comportamiento definido de forma abstracta por la interfaz.

Como ya os comentamos una de las grandes ventajas de METS es la posibilidad de incluir información sobre la estructura de los documentos. Pero además, su gran versatilidad permite además gestionar información de gran utilidad no sólo para utilizarse en la preservación a largo plazo; de hecho, es apto para su uso en un sistema OAIS (Open Archival Information System)

Además de su utilidad para expresar la estructura de los documentos, METS tiene muchas posibilidades para almacenar información relevante de cara a la preservación a largo plazo de los objetos digitales (fechas de creación y modificación de los registros, relaciones entre copias maestras y documentos derivados, etc.) que lo hacen apto para utilizarlo también como conjunto de metadatos en proyectos de preservación digital. De hecho, METS podría utilizarse dentro del modelo de preservación OAIS (Open Archival Information System). Más adelante hablaremos sobre el uso combinado entre METS y PREMIS; una tendencia que han adoptado algunas bibliotecas, entre ellas la Biblioteca Nacional de España.

1.5.2 Preservation Metadata: Implementation Strategies (PREMIS)

PREMIS es un modelo de metadatos orientado a la descripción de objetos digitales con vistas a su preservación digital a largo plazo. Como ya os adelantamos, la Biblioteca del Congreso cuenta con una guía muy útil escrita por Priscilla Caplan, en ella se explican los fundamentos básicos para comprender un modelo que de primeras pueda resultar demasiado complejo. Este documento está traducido al español[29]; las explicaciones que figuran a continuación están extraídas de esta guía.

El mantenimiento de PREMIS se realiza a través de la *PREMIS Maintenance*

[29] CAPLAN, Priscilla: *Entender Premis*. Traducido por en 2009 por María Luis Martínez Conde: http://www.mcu.es/bibliotecas/docs/MC/PREMIS/Contenido.pdf

Activity, que patrocina la Biblioteca del Congreso[30]. PREMIS como tal es, como ya os dijimos, el nombre de un Grupo de Trabajo. Sin embargo, cuando se habla de "PREMIS" se asimila que hablamos del Diccionario de datos fruto del trabajo de este Grupo[31]. En el diccionario no veréis un esquema XML, en cualquier caso PREMIS se diseñó pensando en la versatilidad y funcionalidad del lenguaje XML, y de hecho existe un esquema XML para PREMIS[32].

PREMIS está pensado para la preservación digital a largo plazo, y en esa línea destacamos cuatro ideas que señala Caplan en su trabajo:

- Un recurso debe almacenarse de manera segura para que nadie pueda modificarlo inadvertidamente (o malintencionadamente).
- Los ficheros deben almacenarse en soportes que puedan leer los ordenadores actuales.
- Transcurrido un período largo de tiempo, incluso los formatos de fichero más comunes pueden convertirse en obsoletos, lo que significa que las aplicaciones actuales no pueden reproducirlos. Los gestores de la preservación deben emplear *estrategias de preservación* que garanticen que los recursos se puedan seguir utilizando. Esto puede significar la transformación de los antiguos formatos en otros nuevos equivalentes (*migración*), o la imitación del antiguo entorno de reproducción en el nuevo hardware y software (*emulación*)[33]. Tanto las estrategias de emulación como las de migración requieren metadatos sobre los formatos de los ficheros originales y los entornos de hardware y software que los soportan.
- Las acciones de preservación pueden implicar modificaciones de los

[30] http://www.loc.gov/standards/premis/

[31] www.loc.gov/standards/premis/v2/premis-2-0.pdf Este diccionario también está disponible en español gracias a la traducción de Bárbara Muñoz de Solano y Palacios y Lorea Elduayen Pereda en: http://www.loc.gov/standards/premis/PREMIS_es.pdf.

[32] http://www.loc.gov/standards/premis/schemas.html

[33] En el punto 1.2 os mencionamos todas las posibles estrategias que se pueden aplicar. Sin embargo, os recomendamos que consultarais el manual de la DPC para obtener información detallada sobre cada una de ellas.

recursos originales o cambios en su modo de reproducción. Esto puede poner en duda la autenticidad del recurso. Los metadatos pueden ayudar a soportar la autenticidad del recurso mediante la documentación de la *procedencia digital* de dicho recurso – su cadena de custodia y el historial de cambios autorizados.

Para hacer todo esto, el Diccionario de datos PREMIS define un conjunto de *unidades semánticas* fundamentales que deben entender los repositorios para llevar a cabo sus funciones de preservación. Hay que destacar que el Diccionario excluye voluntariamente algunos metadatos como son:

- Los metadatos de un formato específico, es decir, los metadatos que pertenecen solo a un formato de fichero o a una clase de formato como audio, video o gráficos de vectores.
- La implementación de metadatos específicos y reglas de negocio, es decir, los metadatos que describen las políticas o prácticas de un repositorio en particular, por ejemplo, cómo proporciona dicho repositorio el acceso a los materiales.
- Los metadatos descriptivos. Aunque la descripción de los recursos es, obviamente, relevante para la preservación, pueden utilizarse varios estándares independientes para este objetivo, como MARC21, MODS y Dublin Core.
- La información detallada sobre el soporte o el hardware. De nuevo, aunque asimismo está claro que son relevantes para la preservación, estos metadatos deben ser definidos por otras comunidades.
- La información sobre agentes (personas, organizaciones o software) distintos de los mínimos necesarios para la identificación.
- La información sobre derechos y permisos, excepto los que afectan directamente a las funciones de preservación. Si se tienen presentes todos los metadatos que necesita una organización que gestiona un repositorio de preservación, PREMIS puede considerarse el subconjunto que se define en el centro, que no está relacionado con la recuperación y el acceso ni se propone definir los metadatos detallados de un formato específico. Solamente define los metadatos que se necesitan, por lo general, para llevar

a cabo las funciones de preservación de todos los materiales.

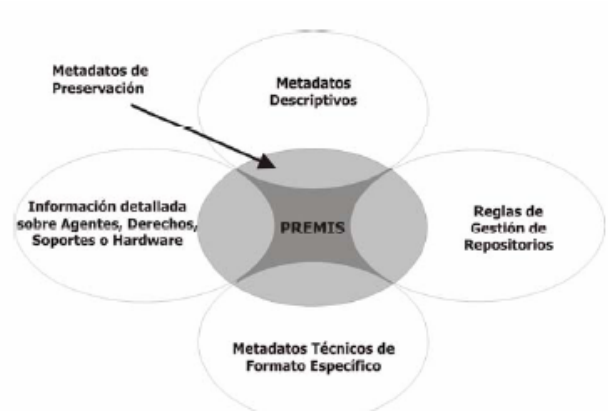

Ilustración 25: El modelo de metadatos PREMIS entendido únicamente como un "set" de todos los metadatos de preservación posibles (Tomado del informe "Understanding PREMIS").

Otra cosa que destaca Caplan en su trabajo es que el Diccionario PREMIS define unidades semánticas y no metadatos. Las unidades semánticas son piezas de información o de conocimientos y los elementos de metadatos son concreciones de esas unidades semánticas. En realidad es la misma idea que cuando decíamos que el Diccionario PREMIS no es lo mismo que el esquema PREMIS, aunque de hecho, la mayoría de las veces se usan como sinónimos.

> *"Así, para ser un purista de PREMIS, hay que pensar en términos de unidades semánticas bastante abstractas. Las unidades semánticas de PREMIS presentan un mapeo directo con los elementos de metadatos definidos en el esquema PREMIS XML, y pueden tener un mapeo menos directo con los metadatos de otro esquema."*

De estas unidades semánticas algunas se definen como **contenedores**. Se trata de aquellas que no tienen un valor en sí mismas sino que existen para agrupar unidades semánticas relacionadas. Por ejemplo, donde quiera que se registre un identificador en PREMIS debe especificarse de qué tipo de identificador se trata (por ejemplo, "DOI", "ISBN", "asignado por el sistema local"). Los contenedores proporcionan una estructura

jerárquica al Diccionario de datos que en la versión 2.0[34] se refleja en la numeración de las unidades semánticas:

1.1 objectIdentifier (identificador del objeto) (O, R)

1.1.1 objectIdentifierType (tipo de identificador del objeto) (O, NR) 1.1.2 objectIdentifierValue (valor del identificador del objeto) (O, NR)

Este extracto del Diccionario de datos muestra a simple vista que la unidad semántica *objectIdentifier* (identificador del objeto) es obligatoria (O) y repetible (R). Puesto que existen unidades semánticas definidas bajo ella, se puede deducir que el *objectIdentifier* (identificador del objeto) no tiene un valor en si mismo sino que sirve como contenedor de los elementos componentes *objectIdentifierType* (tipo de identificador del objeto) y *objectIdentifierValue* (valor del identificador del objeto). Puesto que el *objectIdentifierType* (tipo de identificador del objeto) y el *objectIdentifierValue* (valor del identificador del objeto) son no repetibles (NR) dentro del contenedor, hay que repetir toda la estructura del contenedor para registrar dos identificadores distintos.

Otro concepto que nos explica Caplan es el de **contenedores de extensión**. Estos son un tipo especial de contenedor que no tiene subunidades definidas bajo si mismo. Se ha diseñado para disponer de un lugar donde registrar los metadatos no-PREMIS. En este sentido, PREMIS puede extenderse para incluir metadatos que se encuentran fuera de su alcance u otros no incluidos en el Diccionario de datos. Los contenedores de extensión llevan "Extension" (Extensión) en la última parte de su nombre.

Por ejemplo, los metadatos técnicos de un formato específico no se incluyen en PREMIS, pero es una información muy importante para la preservación digital. El contenedor de extensión 'extensión de las características del objeto' (*objectCharacteristicsExtension*) proporciona un lugar en el que registrar los metadatos

[34] Ya hay una versión 2.2, y en julio de 2013 el Comité Editorial de PREMIS ya ha aprobado una serie de cambios que una vez actualizados darán lugar a la versión 3.0. Podéis consultar estos cambios en: http://www.loc.gov/standards/premis/changes-3-0.html

técnicos definidos por otros diccionarios de datos, por ejemplo el estándar Z39.87 para describir el mapa de bits de las imágenes.

Si se está familiarizado con XML, será obvio que el Diccionario de datos PREMIS se ha diseñado de manera que sea compatible con XML. Las unidades semánticas de PREMIS pueden implementarse como elementos XML; las unidades contenedoras son elementos que únicamente toman otros elementos como contenido y las unidades de extensión son contenedores de los elementos definidos por un esquema externo.

En resumen, el Diccionario de Datos define unidades semánticas, algunas de las cuales son contenedoras (de elementos o de extensión).

1.5.2.1 Entidades

Todas las unidades semánticas de PREMIS (o elementos) se estructuran para definir diferentes tipos de Entidades. Las Entidades son diferentes *cosas* que se relacionan con la preservación digital. PREMIS define cinco tipos de Entidades: Entidades Intelectuales, Objetos, Agentes, Acontecimientos y Derechos. A esto es a lo que se llama el modelo de datos de PREMIS.

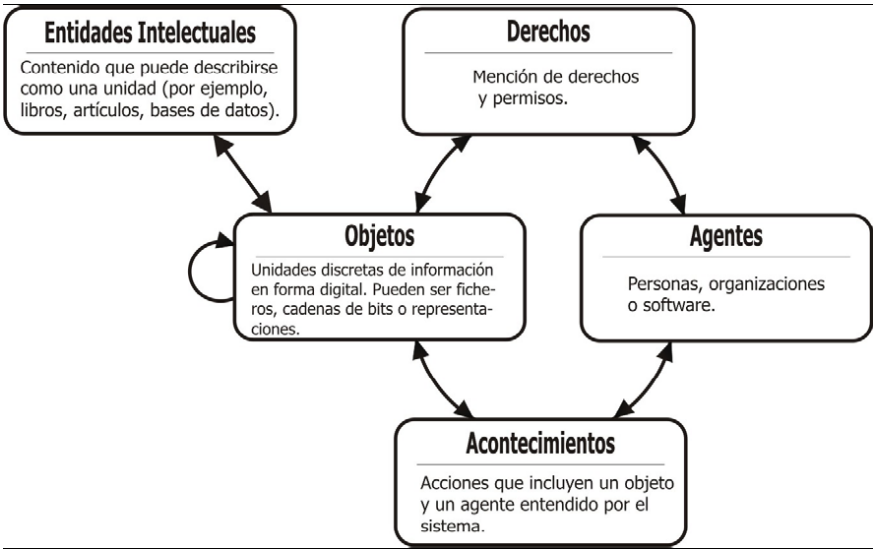

Ilustración 26: El modelo de datos PREMIS define o puede definir 5 tipos de entidades.

Entidad Intelectual

Las *Entidades Intelectuales* son conceptuales y pueden denominarse "entidades bibliográficas". PREMIS define la Entidad Intelectual como "un conjunto de contenido que se considera como una sola unidad intelectual para los propósitos de gestión y descripción: por ejemplo, un libro, un mapa, una fotografía o una base de datos". En realidad, PREMIS no define los metadatos correspondientes a las Entidades Intelectuales porque existen muchos estándares de metadatos descriptivos entre los que se puede elegir.

PREMIS establece que en un sistema de preservación un objeto debe estar asociado a la entidad intelectual que representa mediante la inclusión de un identificador de dicha entidad en los metadatos del objeto. Así, por ejemplo, si estamos preservando un ejemplar de *Buddishm: The Ebook: an Online Introduction* podemos utilizar el ISBN como enlace a la Entidad Intelectual en la descripción PREMIS del e-libro.

Entidad Objeto

Los *Objetos* son lo que realmente se almacena y gestiona en un repositorio de preservación. La mayor parte de PREMIS se dedica a describir objetos digitales. La información que se puede registrar incluye:

- El identificador único del objeto (tipo y valor),
- Fijeza de la información, como la suma de verificación (mensaje cifrado) y el algoritmo utilizado para obtenerla,
- El tamaño del objeto,
- El formato del objeto, que puede especificarse directamente o mediante un enlace a un registro de formatos,
- El nombre original del objeto,
- Información sobre su creación,

- Información sobre los inhibidores,
- Información sobre sus propiedades significativas,
- Información sobre su entorno,
- Dónde y en qué soporte se almacena,
- Información sobre la firma digital,
- Relación con otros objetos y otros tipos de entidades.

En realidad, PREMIS define tres tipos diferentes de objetos: objetos *fichero*, objetos *representación* y objetos *cadena de bits*.

- El *objeto fichero* es exactamente como suena: un fichero de ordenador, por ejemplo, un fichero PDF o un JPEG.
- El *objeto representación* es el conjunto de todos los objetos fichero que se necesitan para reproducir una Entidad Intelectual. Por ejemplo, supongamos que se quiere preservar una página web. Hay bastantes posibilidades de que la página de inicio que se ve en el navegador esté formada por diferentes tipos de ficheros –uno o más ficheros HTML, unas cuantas imágenes TIFF o JPEG, quizá un pequeño audio o una animación Flash. Probablemente también utiliza una hoja de estilo para su visualización. Un navegador reúne todos estos ficheros para reproducir la página de inicio y visualizarla de manera que si un repositorio quiere preservar una página web que se pueda reproducir tiene que entender todos estos ficheros y saber cómo reunirlos. El objeto representación permite al repositorio no solo identificar el conjunto de ficheros relacionados sino también describir las características de la totalidad (e.g., la página web como un todo) que pueden ser diferentes de las de sus partes.
- Los *objetos cadena de bits* son subconjuntos de ficheros. Un objeto cadena de bits se define como datos (bits) dentro de un fichero que a) presenta propiedades comunes para los propósitos de preservación, y

b) no puede ser autónomo sin añadir un fichero cabecera u otra estructura. Así, por ejemplo, si se dispone de un fichero en formato AVI (audio-video intercalado) puede que se quiera diferenciar la cadena de bits del audio de la cadena de bits del vídeo y describirlos como objetos cadena de bits independientes.

Algunas unidades semánticas definidas en el Diccionario de datos PREMIS se pueden aplicar a los tres tipos de objetos, mientras que otras solo se aplican a uno o dos tipos de objetos. El hecho de tener distintos tipos de objetos obliga a pensar en lo que se describe y a ser lo más preciso posible, lo que es importante para el proceso automático.

Entidad Acontecimiento

La entidad Acontecimiento agrega información sobre acciones que afectan a los objetos del repositorio. Un registro preciso y fiable de los acontecimientos es crítico para el mantenimiento de la procedencia digital de un objeto lo que, a su vez, es importante para demostrar la autenticidad del objeto. La información sobre los acontecimientos que se puede registrar incluye:

- El identificador único del acontecimiento (tipo y valor),
- El tipo de acontecimiento (creación, ingesta, migración, etc.),
- La fecha y hora en la que ocurrió el acontecimiento,
- La descripción detallada del acontecimiento,
- El resultado codificado del acontecimiento,
- Una descripción más detallada del resultado,
- Los agentes implicados en el acontecimiento y sus funciones,
- Los objetos implicados en el acontecimiento y sus funciones.

Entidad Agente

Los *Agentes* son actores con funciones en los acontecimientos y en las menciones de derechos. Los agentes pueden ser personas, organizaciones o aplicaciones de software. PREMIS solo define el número mínimo de unidades

semánticas necesarias para identificar los agentes puesto que existen varios estándares externos que se pueden utilizar para registrar información más detallada. (Como muestra de ellos véase "Metadata standards and specifications for describing people and their interests" en www.ukoln.ac.uk/metadata/resources/people.) Un repositorio puede elegir entre utilizar un estándar independiente para registrar información adicional sobre los agentes o utilizar el identificador del agente para apuntar a la información registrada externamente. El Diccionario de datos incluye:

- Un identificador único para el agente (tipo y valor),
- El nombre del agente,
- La designación del tipo de agente (persona, organización, software).

Siempre que se haga referencia a un agente en relación con un acontecimiento o con una mención de derechos, debe registrarse también la función del agente. Cualquier agente puede tener varias funciones. Por ejemplo, yo podría ser el autor y el derechohabiente de una obra, el autor (pero no el derechohabiente) de una segunda obra y el depositario de una tercera. En el modelo PREMIS un repositorio debe asignarme un identificador único y debe consignar el identificador del registro de cualquier acontecimiento o mención de derechos en el que soy un agente, junto con mi función en ese contexto particular. La función de un agente en relación con un acontecimiento o una mención de derechos se considera una propiedad de la entidad acontecimiento o de la entidad derechos y no del propio agente.

Entidad Derechos

La mayor parte de las estrategias de preservación implican la creación de copias idénticas y versiones derivadas de los objetos digitales, acciones que están restringidas por la ley del copyright a los derechohabientes. La *entidad Derechos* agrega información sobre los derechos y permisos que son directamente relevantes para preservar los objetos del repositorio. Cada una de las menciones de derechos de PREMIS constata dos cosas: las acciones a las que tiene derecho el repositorio y las bases para la reclamación de ese derecho.

Por ejemplo, un repositorio puede albergar la versión escaneada de un libro que se

publicó en 1848 y se encuentra, por lo tanto, en el dominio público. El repositorio puede actuar sobre su versión digital sobre la base del estado del copyright del ítem. Otro repositorio alberga un objeto copiado de un CD publicado en el que la licencia de uso individual permite hacer copias de seguridad, pero restringe el acceso y el uso.

La información que puede registrarse en una mención de derechos incluye:

- El identificador único de la mención de derechos (tipo y valor),
- Si la base para la reclamación de los derechos es el copyright, una licencia o una ley,
- Información más detallada sobre el estado del copyright, los términos de la licencia o la ley, si es aplicable,
- La(s) acción (es) que permite la mención de derechos,
- Cualquier restricción sobre la(s) acción(es),
- Los derechos otorgados o el período de tiempo durante el que se aplica la mención,
- El (los) objeto(s) a los que se aplica la mención,
- Los agentes implicados en la mención de derechos y sus funciones.

Nos queda ver propiamente los elementos que se incluyen dentro del Diccionario de Datos. Esto lo podéis consultar directamente en el propio diccionario. Para facilitar esta consulta Caplan explica cómo se definen las unidades semánticas dentro del Diccionario.

Partimos del ejemplo de la definición de la unidad semántica size (tamaño). En el Diccionario encontramos lo siguiente:

Unidad semántica	**1.5.3 size (tamaño)**		
Componentes semánticos	ninguno		
Definición	El tamaño en bytes del fichero o cadena de bits almacenados en el repositorio		
Fundamentos	El tamaño es útil para asegurar el número correcto de bytes de almacenamiento que se han recuperado y que una aplicación tiene espacio suficiente para mover o procesar los ficheros. También puede utilizarse cuando se factura por el almacenamiento.		
Restricciones de los datos	entero		
Categoría del objeto	**Representaciones**	**Fichero**	**Cadena de bits**
Aplicabilidad	No aplicable	Aplicable	Aplicable
Ejemplos		2038937	
Repetibilidad		No repetible	No repetible
Obligatoriedad		Opcional	Opcional
Notas de creación/mantenimiento	Obtenidas automáticamente del repositorio		
Notas de uso	La definición de esta unidad semántica como tamaño en bytes hace innecesario registrar una unidad de medida. Sin embargo, para el propósito de intercambio de datos la unidad de medida debe ser establecida o entendida por ambas partes.		

Ilustración 27: Entra de la unidad semántica "size" en el diccionario de metadatos PREMIS

La entrada del Diccionario de datos incluye la definición del elemento y la razón (fundamentos) por la que se incluye entre los metadatos fundamentales de PREMIS así como ejemplos y notas sobre cómo obtener y utilizar el valor. Se pretende que todos ellos ayuden a los implementadores a utilizar adecuadamente el elemento.

Las dos filas “Categoría del objeto” y “Aplicabilidad” se utilizan conjuntamente para mostrar si la unidad semántica es adecuada para describir representaciones, ficheros y/o cadenas de bits. Aquí se presenta el tamaño como perteneciente únicamente a ficheros y cadenas de bits. Finalmente, hay un conjunto de reglas de uso: “Restricción de los datos”, “Repetibilidad”, “Obligatoriedad”.

Las restricciones de los datos especifican restricciones sobre los valores que puede tener una unidad semántica. En este ejemplo, el valor del tamaño puede ser un entero.

Otra restricción común de los datos es que el valor puede tomarse de un vocabulario controlado.

A veces los términos del vocabulario se especifican en el Diccionario de datos y a veces no, pero, en cualquier caso, debe registrarse el nombre del vocabulario utilizado. En el Diccionario de datos no se definen unidades semánticas para los nombres de vocabularios, pero la Actividad de Mantenimiento de PREMIS está desarrollando un esquema XML para solucionarlo.

La repetibilidad indica si se puede repetir la unidad semántica. La obligatoriedad indica si un valor es obligatorio (requerido) u opcional para la unidad semántica.

Vamos a ver ahora la definición de una unidad contenedora:

Unidad semántica	**1.5 objectCharacteristics (características del objeto)**		
Componentes semánticos	1.5.1 compositionLevel (nivel de composición) 1.5.2 fixity (fijeza) 1.5.3 size (tamaño) 1.5.4 format (formato) 1.5.5 creatingApplication (aplicación creadora) 1.5.6 inhibitors (inhibidores) 1.5.7 objectCharacteristicsExtension (extensión de las características del objeto)		
Definición	Propiedades técnicas de un fichero o cadena de bits aplicables a todos o a la mayoría de los formatos.		
Fundamentos	Hay algunas propiedades técnicas importantes que se aplican a objetos en cualquier formato. La definición detallada de las propiedades de un formato específico queda fuera del alcance de este Diccionario de datos aunque dichas propiedades pueden incluirse en *objectsCharacteristicsExtension* (extensión de las características del objeto).		
Restricciones de los datos	Contenedor		
Categoría del objeto	**Representación**	**Fichero**	**Cadena de bits**
Aplicabilidad	No aplicable	Aplicable	Aplicable
Repetibilidad		Repetible	Repetible
Obligatoriedad		Obligatorio	Obligatorio
Notas de uso	Las unidades semánticas incluidas en *objectsCharacteristics* (características del objeto) deben tratarse como un conjunto de información que pertenece a un único objeto a un único nivel de composición [*compositionLevel*]. Las características del objeto pueden repetirse.		

Ilustración 28: Entrada del diccionario PREMIS para la unidad contenedora "objectCharacteristics"

Aquí vemos el principio de la entrada del Diccionario de datos para las características del objeto [*objectCharacteristics*], la unidad contenedora del tamaño. Se puede decir que se trata de un contenedor porque tiene componentes semánticos y la restricción de los datos es "contenedor". Obsérvese que los componentes semánticos incluidos pueden ser unitarios, como el tamaño, o contenedores en si mismos, como el formato.

PREMIS es, por todo lo dicho, el modelo de metadatos típicamente asociado a la preservación digital a largo plazo. Se trata de un Diccionario (con su correspondiente esquema XML) en el que se definen unidades semánticas relativas a cinco entidades que se consideran claves para la preservación digital: Entidades intelectuales, Objetos, Acontecimientos, Derechos y Agentes.

Como ya hemos comentado anteriormente, resulta interesante reflexionar sobre la compatibilidad entre METS y PREMIS, dado que ambos presentan elementos concurrentes. Su integración no es complicada pero requiere tomar una serie de decisiones sobre cómo estructurar el documento XML que incluya los datos recogidos en ambos modelos. La analizaremos más adelante, al final de este módulo.

1.5.3 Metadata for Images in XML Schema (MIX)

El modelo de metadatos MIX, ya lo habéis visto y se os ha comentado está pensado para la definición de imágenes digitales ráster[35], con el fin de permitir a los usuarios desarrollar, intercambiar e interpretar archivos de imágenes digitales.

El diccionario de datos[36] (donde se definen cual es necesario consignar para la

[35] Imagen ráster o de barrido: Se produce al dividir una imagen en puntos mediante una retícula, y posteriormente, asignar un valor a cada punto para indicar su color (definición tomada del módulo de autoformación de SEDIC sobre la digitalización de documentos; disponible en: http://www.sedic.es/autoformacion/digitalizacion/index.htm)

[36] http://www.niso.org/kst/reports/standards/kfile_download?id%3Austring%3Aiso-8859-1=Z39-87-2006.pdf&pt=RkGKiXzW643YeUaYUqZ1BFwDhIG4-24RJbcZBWg8uE4vWdpZsJDs4RjLz0t90_d5_ymGsj_IKVaGZww13HuDlSn6cvwjex0ejiIKSaTYlErPbfamndQa6zkS6rLL3oIr

gestión de imágenes digitales) es una norma ANSI/NISO Z39.87-2006, y la información principal el estándar se puede encontrar en la pagina de la Biblioteca del Congreso[37], donde también se puede acceder al esquema MIX en XML[38].

En el Diccionario se emplean dos tipos de elementos: contenedores de datos y elementos de datos. Los contenedores de datos son agrupaciones semánticas de dos o más elementos de datos relacionados, contenedores o sub-contenedores. Los elementos de datos son la parte componente del Diccionario de datos y se utilizan para registrar los valores específicos de los datos.

Como ya visteis el Diccionario describe elementos para definir cinco tipos de informaciones: las relativas al objeto digital; a la imagen digital; a la captura de la imagen; a la evaluación de la calidad de la imagen; y la relativa al historial de cambios a los que se somete dicha imagen.

[37] http://www.loc.gov/standards/mix//

[38] http://www.loc.gov/standards/mix/mix20/mix20.xsd

1.5.4 Combinación de modelos METS y PREMIS y ejemplos de aplicación (el caso de la BNE)

En 2008 la LoC publicó un sencillo documento en el que daban directrices sobre cómo codificar PREMIS en ficheros METS para facilitar el intercambio de información (básicamente lo que preservación conocemos como SIP y DIP). Sin embargo, no se descarta que esta codificación pueda utilizarse también para fines de archivo (AIP).

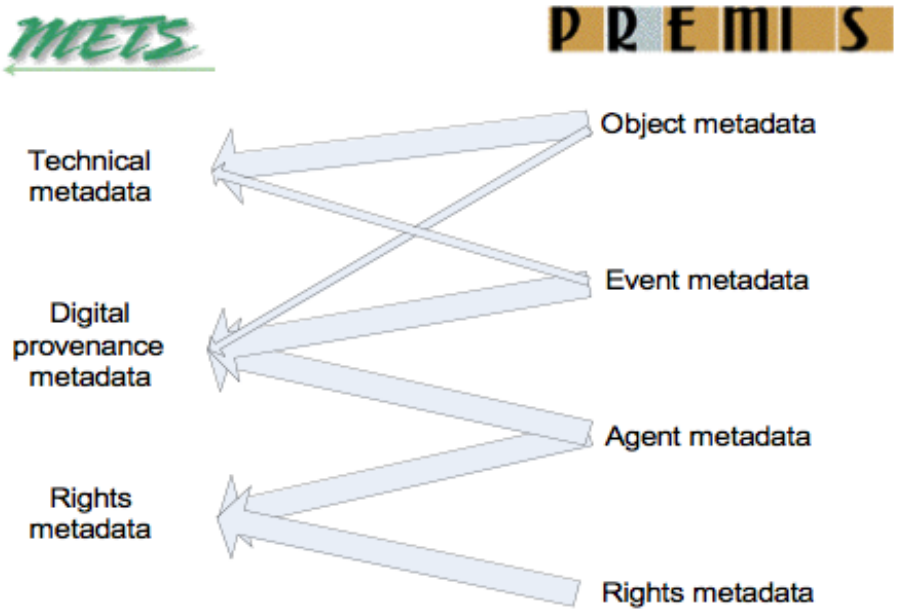

Ilustración 29: Esquema que sintetiza las posibilidades de combinación entre los modelos de metadatos METS y PREMIS

En esta pequeña guía tras una mínima descripción de los grandes apartados que tienen uno y otro modelo, agrupan las decisiones a tener en cuenta en 7 apartados principalmente:

1. <u>En qué secciones de METS utilizar PREMIS</u>

   Una opción sería incluir todos los metadatos PREMIS que hayamos decidido incluir en nuestro perfil de metadatos de preservación dentro de una misma sección amdSec. Concretamente se incluirían en digiProvMD con el elemento <premis> actuando de contenedor.

   La otra sería incluir los elementos en distintas secciones según las siguientes

premisas:

- La entidad premis:object puede incluirse en techMD o digiProvMD. Se decida una u otra cosa, lo importante es que a través del StructMap y de la Sección de archivo quede claro los archivos a los que se hace referencia y qué metadatos se aplican a cada uno.
- La entidad premis:event se incluiría en digiProvMD
- La entidad premis:rights se incluiría en rightsMD
- La entidad premis:agent se incluiría en digiProvMD (si se refiere a un evento) o en rightsMD (si se refiere a una declaración de derechos)

2. El número de secciones a utilizar

   Se puede ó bien utilizar una única sección amdSec en la que se repitan los distintos subelementos METS (techMD,…) ; o repetir la amdSec por cada subelemento METS.

   La entidad agente de premis relacionada con un evento o los derechos, debe ir en la sección digiProv o rightsMD que le corresponda.

   Los metadatos técnicos procedentes de distintos esquemas pueden incluirse en secciones techMD independientes o incluirse bajo objectCharacteristicsExtension de la entidad objeto

3. Uso de un contenedor PREMIS

   Si se distribuyen los metadatos PREMIS entre las subsecciones administrativas de METS no hace falta de utilizar un contenedor METS. Sólo se utiliza si se agrupa todo el PREMIS, y como ya hemos señalado se incluiría en digiProvMD.

4. Manejo de redundancias entre PREMIS y METS

   Sendos modelos pueden presentar redundancias; así, por ejemplo, los metadatos técnicos pueden expresarse bien a través de METS, PREMIS (ej: el tamaño del archivo (SIZE) en PREMIS se incluye en <size> bajo <objectcharacteristics>, en METS es un atributo de <file> dentro de la sección <fileGrp>, pero además es un elemento en MIX.

   Aunque sea una decisión a tomar por cada institución, y deba tenerse en cuenta la finalidad última de nuestros datos (visualización para METS; preservación para PREMIS), conviene señalar que PREMIS suele ser más expresivo en lo que a metadatos técnicos se refiere, no tanto estructura (uno de los puntos fuertes de

METS).

Las decisiones finales sobre la forma de hacer frente a estas redundancias debe quedar reflejado en un perfil.

5. Manejo de relaciones estructurales en METS (StructMap) y PREMIS (structural relationship elements)

   En resumen se recomienda que las relaciones jerárquicas estructurales se expresen mediante elementos anidados dentro del esquema PREMIS.

6. Elementos de identificación de METS y PREMIS

   Tanto en METS como en PREMIS existen identificadores. En PREMIS ayudan a identificar entidades (objectIdentifier, eventIdentifier…) y enlazar unas entidades con otras, además de enlazar elementos relacionados. Por su parte, METS cuenta también con ID/IDrefs que ayudan a enlazar archivos y los metadatos que se refieren a ellos.

   Entre las claves a tener en cuenta a la hora de crear IDs, no se deben perder de vista algunas pautas básicas: deben ser unívocos, persistentes y deben establecerse las relaciones necesarias para que no quede ningún elemento huérfano ej: no deberá existir ningún elemento techMED que no tenga su referencia en el StructMap/fileSec.

Vamos a ver de cerca un ejemplo relacionado con un fichero METS-PREMIS de la BNE. Podéis encontrar el archivo entero en la documentación adjunta a este módulo[39]. Este archivo contiene información sobre una obra descrita en MARC y de la cual tenemos:

- 2 archivos máster (FPM000001, FPM000002),
- 2 archivos máster cortados (FPC000001, FPC000002),
- 2 archivos máster editados (FPE000001, FPE000002)

[39] EJERCICIO_PREMIS_ALUMNO.xml. Si bien, encontraréis que falta alguna información en algún elemento, circunstancia que se señala con "XXXXXXX". Esto es así porque este xml lo utilizaréis como base para la práctica asociada al módulo.

- y un archivo de difusión en formado pdf (FPP000001).

Para analizar este ejemplo vamos a hacer uso de otro documento muy útil también disponible en la página de Premis de la LoC: *A checklist for documenting PREMIS-METS decisions in a METS profile*. Como se adelanta en su introducción, esta lista fue concebida para ayudar en la implementación de PREMIS en METS, o al menos considerarla. En 13 puntos de la lista se facilitan ejemplos de la Biblioteca Nacional de Australia (en inglés NLA), la Universidad de Illinois, La Universidad de California en San Diego y la de Southhampton. Nosotros iremos recorriendo cada uno de los puntos aportando además imágenes del perfil diseñado por la BNE[40]. Conviene no olvidar que, por el momento, la BNE sólo tiene a punto un perfil, queda mucho por hacer y, por ejemplo, debería adaptarse para otros materiales ej: archivos de audio, audiovisuales, ficheros nacidos digitales…

1. **¿Cómo se relaciona el perfil que estamos diseñando con el resto de perfiles METS?**

   Un repositorio digital, por lo general, necesitará varios perfiles. Por lo tanto, resulta muy útil declarar en ellos su aplicación y la posible relación con el resto.

   Por ejemplo, en el perfil de la BNE se explica que, de momento, éste sólo se ha pensado para aquellos objetos que proceden del proyecto de digitalización masiva.

```
<METS_Profile xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:xlink="http://www.w3.org/1999/xlink" xmlns="http://www.loc.gov/METS_Profile/"
        xmlns:MARC="http://www.loc.gov/MARC21/slim"
        xsi:schemaLocation="http://www.loc.gov/METS_Profile/ http://www.loc.gov/standards/mets/
profile_docs/mets.profile.v1-2.xsd">
        <URI LOCTYPE="URL"
                >http://www.libnova.es/profiles/clientes/perfil_mets_BNELibnovaProfile2011-01.xml</URI>

        <title>Perfil de la Biblioteca Nacional de España creado por libnova orientado a la
preservación
                de los objetos digitales generados en el proyecto de digitalización masiva</title>

        <abstract>El perfil ha sido creado con el propósito de resultar directamente utilizable o
                fácilmente convertible en procesos de preservación posteriores. No se ha definido
teniendo
                en cuenta fines relacionados con la diseminación de contenido.</abstract>

        <date>2011-12-27T11:31:18</date>
```

Ilustración 30: Ejemplo del perfil METS-PREMIS de la BNE en el que se especifica la relación de este perfil con otros

[40] Este perfil se basa en el creado previamente por María Luis Martínez-Conde y Julio Cordal Elviro (de la Subdirección de Coordinación bibliotecaria **del Ministerio de Educación, Cultura y Deporte)** en colaboración con Libnova (http://www.libnova.com/index.php).

En las notas del perfil de la NLA se lee:

"este perfil describe las reglas y requisitos para utilizar METS como formato de intercambio como soporte, preservación y acceso a la colección y al contenido de los repositorios digitales australianos. Se trata de un perfil genérico, no específico para ningún sistema o implementación concreta. Los repositorios deberán desarrollar y registrar subperfiles que detallen requisitos específicos para una determinada implementación.

2. **Qué esquemas (*schemas*[41]) (PREMIS, MOD, MIX) se utilizan y dónde se localizan**

Según el documento de componentes de un perfil METS[42], un perfil debe de registrar de manera explícita los esquemas de metadatos que utiliza y dónde se encuentra ese esquema ej: PREMIS, METS…

En el pantallazo de abajo se muestran los esquemas de extensión que actualmente se utilizan en el perfil de METS-PREMIS de la BNE. En concreto éstos son:

- MARC21 para la descripción bibliográfica. Esquema incluido en la sección dmdSec dentro de un elemento <mdWrap>
- METSrights (desarrollado por la Universidad de Stanford) para consignar el estatus de los derechos de propiedad intelectual y lo que se puede ó no hacer con un determinado archivo. Este esquema se ubica en la sección amdSec

[41] Siguiendo a Eva Méndez, llamamos la atención en este punto que no debe confundirse en el mundo de los metadatos la palabra "scheme", con "schema", ambos traducibles por esquema. Méndez distingue cada una de ellas con las siguientes definiciones:

Scheme: hace referencia a un fichero de tesauro controlado, una lista de valores posibles que puede tener una metaetiqueta en concreto.

Schema: son los elementos y reglas que constituyen un modelo de metadatos…Así por ejemplo, Dublin Core es un schema de metadatos y para completar la información del elemento DC.Subject (materia) de este schema se puede utilizar el scheme de la Clasificación Decimal de Dewey.

[42] En este documento se recoge una breve relación de todos los componentes que debe de tener un perfil de METS. Está disponible en: http://www.loc.gov/standards/mets/profile_docs/components.html

- MIX para consignar las características técnicas de las imágenes, y se ubica en la sección amdSec, en el subelemento techMD, bajo una etiqueta mdWrap en el que se ubica la unidad semántica de PREMIS objectCharacteristics, dentro de la cual a su vez se utiliza MIX como extensión.
- PREMIS para consignar metadatos de preservación; dentro de la sección amdSec, en el subelemento techMD, bajo una etiqueta mdWrap. Siguiendo las directrices de la LoC.

```
        <extension_schema ID="MARC">
                <name>Marc21</name>
                <URI>http://www.loc.gov/MARC21/slim</URI>
                <context> mets/dmdSec/mdWrap @MDTYPE="MARC" </context>
                <note>Registro MARC21 en codificación XML. Preferentemente eliminando etiquetas 856
para validación</note>
        </extension_schema>

        <extension_schema ID="METSrights">
                <name>METSrights</name>
                <URI>http://cosimo.stanford.edu/sdr/metsrights/</URI>
                <context> mets/amdSec </context>
        </extension_schema>


        <extension_schema ID="MIX">
                <name>NISO Metadata for Images in XML (NISO MIX)</name>
                <URI>http://www.loc.gov/mix/v20</URI>
                <context>mets/amdSec/techMD/MdWrap @MDTYPE="PREMIS"/xmlData/object
                        xsi:type="file"/objectCharacteristics/objectCharacteristicsExtension</context>
                <note> Se incluye también el objectCharacteristicsExtension con la información MIX,
                        pero únicamente para ficheros que no sean PDF.
                </note>
        </extension_schema>


        <extension_schema ID="PREMIS">
                <name>PREMIS</name>
                <URI>http://www.loc.gov/standards/premis</URI>
                <context>mets/amdSec/techMD/mdWrap @MDTYPE="PREMIS"</context>
                <note> En base a las recomendaciones sobre interacción entre PREMIS y METS publicadas
por la LOC,
                        se opta en la presente implementación por utilizar PREMIS en su variante
descriptiva del objeto
                        con fines de preservación, descartando cualquier otra relación
potencialmente repetida o
                        redundante y colocando dicha referencia o relación dentro del esquema METS.
                </note>
        </extension_schema>
```

Ilustración 31: Pantallazo del perfil METS-PREMIS de la BNE en el que se consignan todos los esquemas de metadatos utilizados

3. Qué vocabularios controlados se utilizan para las unidades semánticas de PREMIS y dónde se ubican

El diccionario de metadatos PREMIS recomienda situar los vocabularios controlados en un espacio ó servicio común a partir del cual el vocabulario puede reutilizarse por múltiples repositorios. Si no se puede acudir a un único vocabulario compartido se pueden documentar los vocabularios para cada unidad semántica de PREMIS dentro del propio perfil de METS, ó se puede referir a documentos locales a los que se hace referencia (esto es lo que hace la

Universidad de San Diego tal cual se recoge en el documento de checklist que estamos utilizando de guía para este epígrafe).

Señalar que, aunque a día de hoy no existe un único espacio donde registrar y compartir vocabularios controlados para unidades semánticas de PREMIS y elementos de METS, resulta de gran interés conocer la página web de la LoC que tiene precisamente el objetivo de englobar todos los vocabularios relacionados con la preservación[43].

En el siguiente pantallazo se ven los vocabularios utilizados en el ejemplo de la BNE.

```
<controlled_vocabularies>

	<vocabulary>
		<name>MARC Country Codes</name>
		<maintenance_agency>Library of Congress</maintenance_agency>
		<URI>http://www.loc.gov/marc/</URI>
		<context>
			<p>Prestar atención al tratamiento de los espacios en blanco de los códigos
de control MARC. Deben respetarse pues contienen información sensible.</p>
		</context>
	</vocabulary>

	<vocabulary>
		<name>Elementos TYPE para los elementos DIV dentro de structMap</name>
		<maintenance_agency>BNE</maintenance_agency>
		<values>
			<value>Imágenes Máser</value>
			<value>Derivado - Imágenes cortadas</value>
			<value>Derivado - Imágenes editadas</value>
			<value>Derivado - Archivos PDF</value>
		</values>
		<context>
			<p>mets/structMap/div/@TYPE</p>
		</context>
		<description>
			<p>Usaremos cada uno de los tipos indicados para cada uno de los tipos de
archivos generados en el proceso.</p>
		</description>
	</vocabulary>

	<vocabulary>
		<name>Atributo TYPE para los elementos structMap</name>
		<maintenance_agency>Universidad de California, Berkeley</maintenance_agency>
		<values>
			<value>logical</value>
		</values>
		<context>
			<p>mets/structMap/@TYPE</p>
		</context>
		<description>
			<p>Sólo está permitido usar el TYPE logical, pues está siempre relacionado
con la organización lógica de los objetos</p>
		</description>
	</vocabulary>

</controlled_vocabularies>
```

Ilustración 32: Pantallazo en el que se recoge la declaración de vocabularios utilizados en el perfil METS-PREMIS de la BNE

[43] Disponible en: http://id.loc.gov/vocabulary/preservation.html

4. **¿Se incluye la información de PREMIS dentro del documento METS, o se hace referencia al mismo?**

Cuando se redactó esta lista de comprobación, todos los ejemplos que implementaban la combinación de METS y PREMIS, incluían PREMIS utilizando el elemento de METS mdWrap. Como ya hemos visto, se podría utilizar el elemento mdRef para hacer referencia a un documento externo en el que registrar la información PREMIS.

El perfil de la BNE no es una excepción y se incluye la información PREMIS en el propio documento METS, gracias al elemento mdWrap.

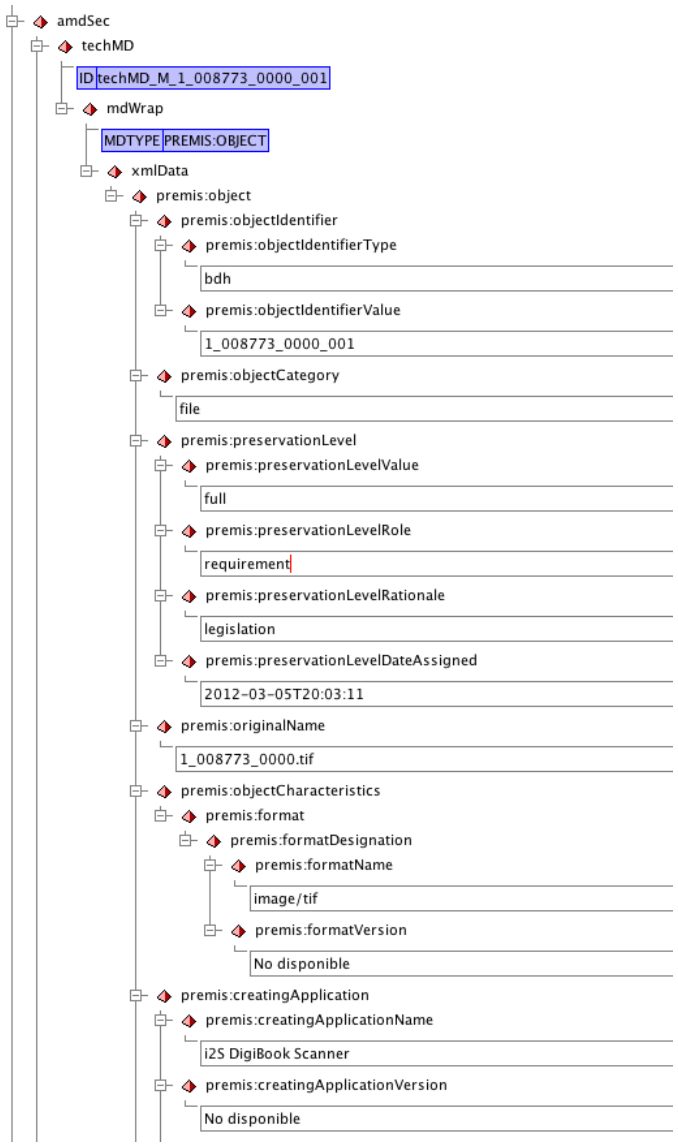

Ilustración 33: Ejemplo de la utilización del elemento mdWrap de METS utilizado en el perfil METS-PREMIS de la BNE para incluir los metadatos de preservación PREMIS

5. **La información PREMIS: ¿se agrupa o se distribuye en varios lugares del documento METS?**

Dentro del perfil METS debemos de consignar cómo distribuimos los metadatos PREMIS. En las directrices que hemos visto anteriormente se recuerda que son varios los lugares en los que se puede ubicar PREMIS en METS; aunque también pueden agruparse como un paquete único en digiProvMD con el elemento PREMIS como contenedor.

En el caso de la BNE se sigue el primer caso, ubicándose la información de premis en varios elementos METS:

- Premis:event bajo digiProvMD
- Premis:object bajo techMD
- Premis: event bajo digiProvMD
- Premis:agent bajo digiProvMD (no bajo rights MD ya que no se utiliza las unidades semánticas de PREMIS referidas a derechos, para ello se utiliza el schema METSRights

6. **La información de PREMIS ¿se sitúa en elementos amdSec separados ó subelementos de amdSec?**

Los metadatos PREMIS pueden codificarse en secciones amdSecs independientes o en un amdSec con distintos subelementos (techMD, digiprovMD, rightsMD…)

En el caso de la BNE nos hemos inclinado por la primera opción y se registra la información PREMIS en una única sección amdSec, dentro de la cual se codifican los subelementos METS aplicables a cada imagen.

Es importante tener en cuenta que en este caso cuando se hace referencia a la admSec, se está haciendo referencia al amdSec y a todos sus hijos.

En las siguientes imágenes se recogen pantallazos que recorren la amdSec de un ejemplo del perfil de BNE; en ellos se ve cómo cada subelemento de METS es aplicable a una imagen en concreto. Esto además se intenta reflejar con el identificador asignado en cada caso. Así, por ejemplo, en la primera

imagen podéis ver la sección amdSec y los metadatos técnicos de la primera imagen de master con su correspondiente ID (<techMD ID="techMD_M_1_008773_001"> […]</techMD>) y que está codificado mediante PREMIS dentro de un elemento mdWrap.

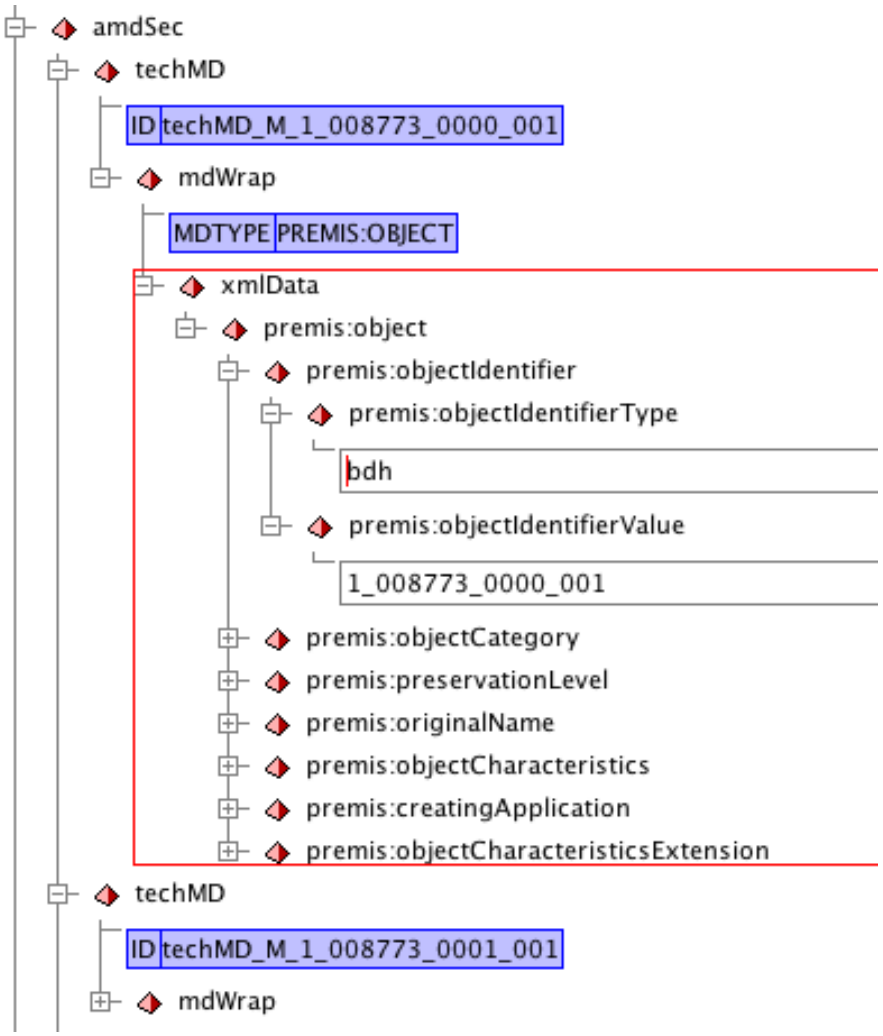

Ilustración 34: Ejemplo de la utilización amdSec en el perfil METS-PREMIS de la BNE

7. **Los metadatos técnicos ¿se registran en secciones techMD independientes o con el componente semántico objectCharacteristicExtension?**

Esta decisión depende de cada implementación; en las directrices se dejan elegir entre una y otra localización, o incluso ambas. En el caso de la BNE se utilizan ambas: secciones techMD para cada imagen, pero además se utiliza el

componente semántico de PREMIS objectCharacteristicExtensión para incluir metadatos técnicos de cada imagen conforme al schema MIX.

```
<amdSec>
    <techMD ID="techMD_M_File1">
        <!-- (Tipo de objeto)_(Tipo de archivo) (Master, Cortado, Editado, Pdf)_(Nombre real del fichero sin extensión)_001 -->
        <mdWrap MDTYPE="PREMIS:OBJECT">
            <xmlData>
                <premis:object xsi:type="premis:representation">
                    <premis:objectIdentifier>
                        <premis:objectIdentifierType>bdh</premis:objectIdentifierType>
                        <premis:objectIdentifierValue>(Identificador único)_(Nombre del
                            fichero)</premis:objectIdentifierValue>
                        <!-- Composición: (Identificador único de la obra)_(Nombre real del fichero) -->
                    </premis:objectIdentifier>
                    <premis:objectCategory>file</premis:objectCategory>
                    <premis:preservationLevel>
                        <premis:preservationLevelValue>full</premis:preservationLevelValue>
                        <premis:preservationLevelRole>requirement</premis:preservationLevelRole>
                        <premis:preservationLevelRationale>legislation</premis:preservationLevelRationale>
                        <premis:preservationLevelDateAssigned>(fecha de
                            generación)</premis:preservationLevelDateAssigned>
                    </premis:preservationLevel>
                    <premis:originalName>(nombre del fichero original)</premis:originalName>
                    <premis:objectCharacteristics>
                        <premis:format>
                            <premis:formatDesignation>
                                <premis:formatName>image/tiff</premis:formatName>
                                <premis:formatVersion>6.0</premis:formatVersion>
                            </premis:formatDesignation>
                        </premis:format>
                    </premis:objectCharacteristics>
                    <premis:creatingApplication>
                        <premis:creatingApplicationName>(Software creador del
                            fichero)</premis:creatingApplicationName>
                        <premis:creatingApplicationVersion>(versión del
                            software)</premis:creatingApplicationVersion>
                        <premis:dateCreatedByApplication>(hora y fecha de creación del
                            fichero</premis:dateCreatedByApplication>
                    </premis:creatingApplication>
                    <premis:objectCharacteristicsExtension>
                        <mix:mix>
                            <mix:BasicDigitalObjectInformation>
                                <mix:Compression>
                                    <mix:compressionScheme>0</mix:compressionScheme>
                                </mix:Compression>
                            </mix:BasicDigitalObjectInformation>
                            <mix:BasicImageInformation>
                                <mix:BasicImageCharacteristics>
                                    <mix:imageWidth>(ancho de la imagen)</mix:imageWidth>
                                    <mix:imageHeight>(alto de la imagen)</mix:imageHeight>
                                    <mix:PhotometricInterpretation>
                                        <mix:colorSpace>RGB</mix:colorSpace>
                                    </mix:PhotometricInterpretation>
                                </mix:BasicImageCharacteristics>
                            </mix:BasicImageInformation>
                            <mix:ImageCaptureMetadata>
                                <mix:ScannerCapture>
                                    <mix:scannerManufacturer>(Fabricante del
                                        Scanner)</mix:scannerManufacturer>
                                    <mix:ScannerModel>
                                        <mix:scannerModelName>(Modelo del
                                            scanner)</mix:scannerModelName>
                                    </mix:ScannerModel>
                                </mix:ScannerCapture>
                            </mix:ImageCaptureMetadata>
                            <mix:ImageAssessmentMetadata>
                                <mix:ImageColorEncoding>
                                    <mix:BitsPerSample>
                                        <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>
                                    </mix:BitsPerSample>
                                    <mix:samplesPerPixel>1</mix:samplesPerPixel>
                                </mix:ImageColorEncoding>
                            </mix:ImageAssessmentMetadata>
                        </mix:mix>
                    </premis:objectCharacteristicsExtension>
                </premis:object>
            </xmlData>
        </mdWrap>
    </techMD>
```

Ilustración 35: Fragmento del perfil de METS-PREMIS de la BNE en el que se ve la sección techMD a utilizar por imagen, y cómo además se utiliza el esquema MIX

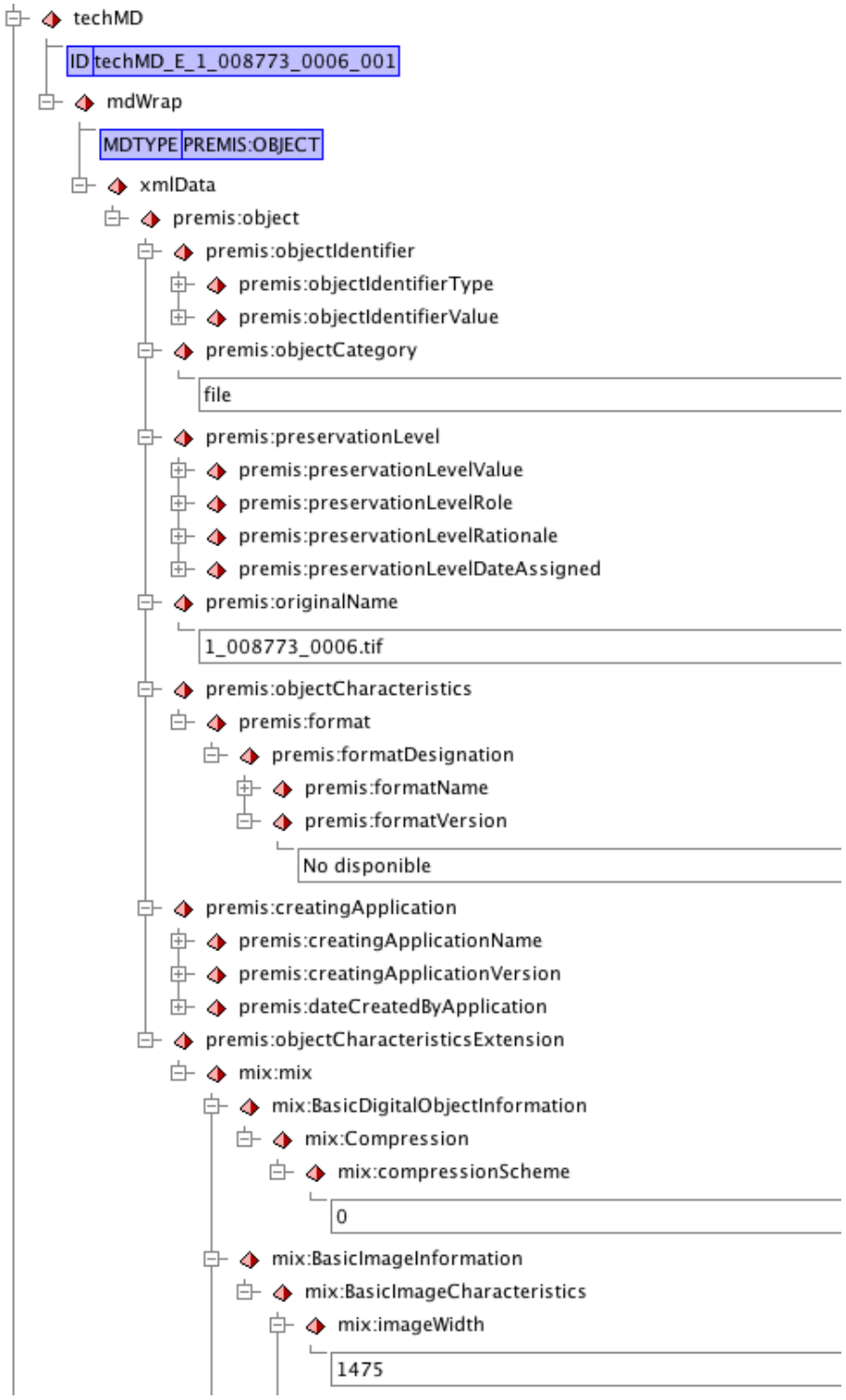

Ilustración 36: fragmento de un ejemplo de fichero METS-PREMIS de la BNE en el que se ve la sección techMD a utilizar por imagen, y cómo además se utiliza el esquema MIX para cada una de ellas

8. ¿Qué unidades semánticas PREMIS exige o recomienda el perfil?

La checklist recomienda que, dentro de la sección de requisitos estructurales de un perfil METS, el repositorio debería registrar las unidades semánticas de PREMIS que son necesarias o recomendadas, además de los valores y/o condiciones de uso necesarios para cada una de ellas.

Esto se puede ver claramente en el perfil que la Universidad de California preparó para tesis y disertaciones, del cual recogemos un pantallazo:

```
<requirement ID="amdSec1">
<p>A METS record conforming to this profile must have an <amdSec> section. The <amdSec> section must contain at least one
techMD/mdWrap/xmlData section and two rightsMD/mdWrap/xmlData sections.</p>
</requirement>
<requirement ID="techMD1">
<p>The required <techMD> must contain elements from the PREMIS schema.</p>
</requirement>
<requirement ID="techMD2">
<p>The required techMD/mdWrap/xmlData must describe the primary (pdf) digital file.</p>
</requirement>
<requirement ID="techMD3">
<p>The required techMD/mdWrap/xmlData must contain a /object/objectIdentifierType and /objectIdentifierValue element</p>
</requirement>
<requirement ID="techMD4">
<p>The required techMD/mdWrap/xmlData must contain a /object/preservationLevel element. The value should be either "Full" or "Bit-level".</p>
</requirement>
...
<requirement ID="copyrightMD1">
The required rightsMD/mdWrap/xmlData must contain elements from the PREMIS Rights schema.</p>
</requirement>
<requirement ID="copyrightMD2">
<p>A conforming METS record must include a rightsStatement/rightsBasis element with the value expressed as "Copyright."</p>
</requirement>
<requirement ID="copyrightMD3">
<p>A conforming METS record must include a rightsStatement/copyrightInformation/copyrightStatus element with the value expressed as "Under
copyright" if the work is still in copyright, or expressed as "In the public domain" if copyright for the work has expired or been gifted to the public
domain.</p>
</requirement>
...
<requirement ID="licenserightsMD1">
<p>A second required rightsMD/mdWrap/xmlData must contain elements from the PREMIS Rights schema.</p>
</requirement>
<requirement ID="licenserightsMD2">
<p>A conforming METS record must include a rightsStatement/rightsBasis element with the value expressed as "License."</p>
</requirement>
...
```

Ilustración 37: Ejemplo de modelo METS-PREMIS de la Universidad de California en el que se especifican las unidades semánticas de PREMIS que se utilizan.

9. Las relaciones entre los objetos se expresan ¿utilizando elementos div de METS, las relationships de PREMIS ó ambas?

Las directrices de la LoC se inclinan por recomendar la utilización de elementos div para expresar relaciones entre los ficheros. Sin embargo, si el perfil tiene funciones de preservación o se expresan relaciones de tipo derivativo (por ejemplo, una imagen B que deriva de una imagen A), entonces también se

recomienda el uso de las relationships de PREMIS.

En el caso del perfil de la BNE, tal como se puede ver en el pantallazo de abajo, únicamente se utilizan los elementos div METS.

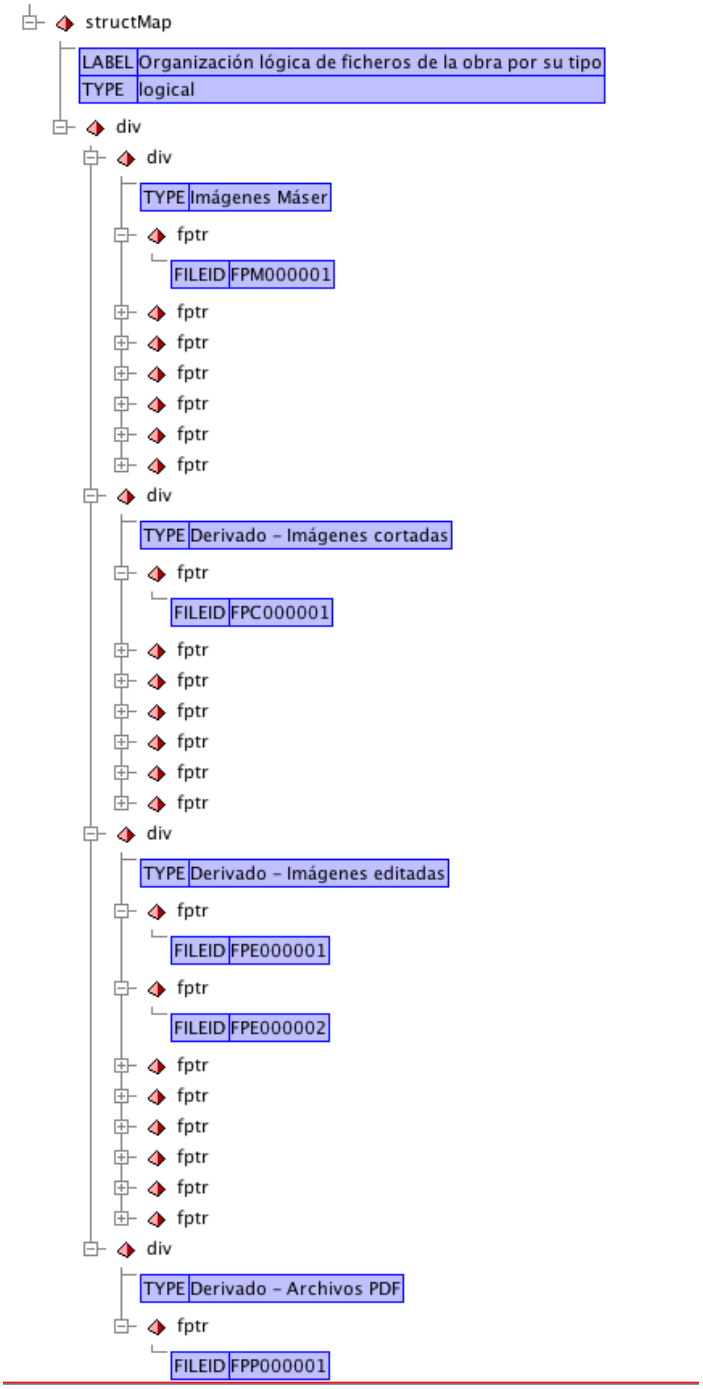

Ilustración 38: Uso de elementos div en el mapa estructura del perfil METS-PREMIS de la BNE

10. ¿A qué nivel de objeto llega la información PREMIS?

La información de PREMIS debería conectarse al objeto u objetos a los que se refiere el METS en su sección de mapa estructural.

En el caso del perfil de la BNE, efectivamente la información PREMIS se refiere a cada uno de los objetos recogidos en el mapa estructura (las imágenes máster, las máster editadas, las máster cortadas y el archivo de difusión).

11. ¿Cómo se utilizan los identificadores de enlace de METS, los IDRefs y los identificadores de PREMIS?

La gestión de identificadores en archivos complejos de este tipo es muy delicada…El estudio de los que se utilizan, de la información que se recoge debe ser minucioso y evitar que ninguno de los elementos quede huérfano.

En concreto las directrices afirman que cuando se incluye información PREMIS en un documento METS los implementadores debe utilizar los ID/IDRefs para conectar los archivos en la sección fileSec con la información PREMIS que se refiere a dichos archivos.

La imagen que mostramos a continuación tienen como fin mostraros los identificadores utilizados para el primer master (FPM000001) y cómo se relacionan todos los metadatos con el archivo adecuado.

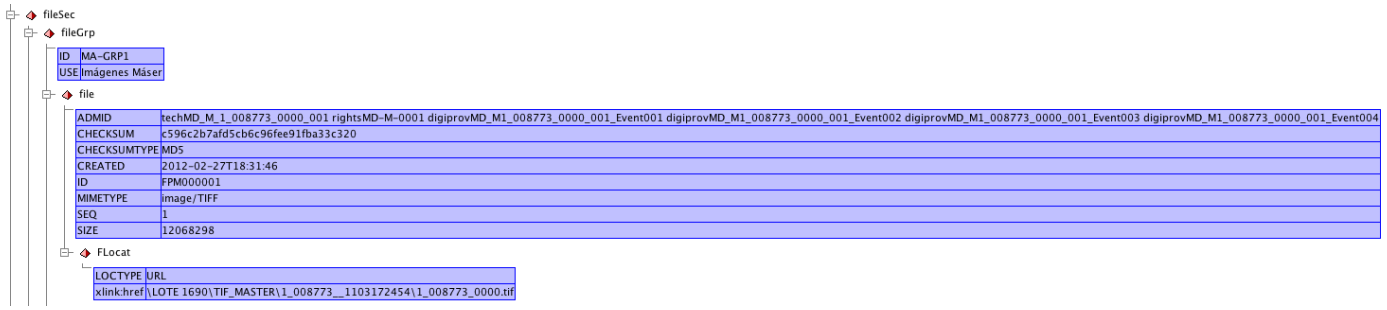

Ilustración 39: Sección FileGrp del perfil METS-PREMIS de la BNE

12. ¿Cómo se manejan las redundancias entre METS y PREMIS?

Como ya vimos al hablar de las directrices de la LoC, se recomienda dejar reflejado en nuestro perfil el manejo de las redundancias. En concreto dicha guía enumera alguna de estas redundancias:

	PREMIS	METS
SIZE	in *size* under *objectCharacteristics*	an attribute of *file* in the *fileGrp*
CHECKSUM and CHECKSUMTYPE	in *fixity* under *objectCharacteristics*	attributes of *file*
MIMETYPE	in *format* under *objectCharacteristics*	an attribute of *file*

Ilustración 40: Tabla de ejemplo que recomienda utilizar la LoC en sus directrices sobre la combinación de METS-PREMIS. En ella se debe explicar el manejo de las redundancias que pueden darse al utilizar conjuntamente sendos modelos de metadatos

En el caso del perfil de la BNE se ha evitado al máximo esta redundancia, de hecho así se recoge en la sección en la que se recogen los distintos schemas de metadatos utilizados:

```
<extension_schema ID="PREMIS">
        <name>PREMIS</name>
        <URI>http://www.loc.gov/standards/premis</URI>
        <context>mets/amdSec/techMD/mdWrap @MDTYPE="PREMIS"</context>
        <note> En base a las recomendaciones sobre interacción entre PREMIS y METS publicadas por la LOC,
                se opta en la presente implementación por utilizar PREMIS en su variante descriptiva del objeto
                con fines de preservación, descartando cualquier otra relación potencialmente repetida o
                redundante y colocando dicha referencia o relación dentro del esquema METS.
        </note>
</extension_schema>
```

Ilustración 41: Declaración del manejo de la redundancia entre METS y PREMIS en el perfil de la BNE

13. ¿Qué aplicaciones o herramientas se utilizan? en la creación, transformación ó preservación de los metadatos PREMIS y METS?

En el documento de componentes de un perfil de METS se señala que se deberá registrar todas las herramientas utilizadas en la creación, transformación o preservación de los metadatos PREMIS y METS.

A continuación un ejemplo del perfil de BNE en el que se recoge la

herramienta de validación JHOVE que os mencionamos en el epígrafe sobre control de formatos.

```
<tool>
        <name>JHOVE</name>
        <agency>JSTOR y la Biblioteca de la universidad de Harvard.</agency>
        <URI>http://hul.harvard.edu/jhove/</URI>
        <description>
                <p>JHOVE es una aplicación capaz de identificar formatos, validar y extraer
                        características de los objetos digitales.</p>
        </description>
        <note>
                <p>Los mets acordes a este perfil se deben validar con JHOVE.</p>
        </note>
</tool>
```

Ilustración 42: declaración del perfil METS-PREMIS de la BNE en el que se declara la utilización de JHOVE como herramienta de identificación, validación y extracción de características de los objetos digitales

1.6. Conclusiones sobre el uso combinado de METS-PREMIS

A la hora de implementar un modelo de datos que combine METS y PREMIS (y en realidad cualquier con esquema que vayamos a empezar a utilizar en nuestros perfiles), se recomienda el uso de perfiles que especifiquen las decisiones tomadas (por ejemplo, qué elementos utilizar en caso de redundancia entre PREMIS y METS…)

Como habéis visto, son muchas las decisiones que deben de tomarse a la hora de incluir PREMIS en METS. Tanto en el camino del diseño de nuestro perfil, como en la validación de un documento final (que bien hayamos producido nosotros ó que nos facilite un proveedor). Y como también hemos visto, hay varias herramientas que nos pueden resultar de gran utilidad; entre ellos destacaríamos:

- Ver y analizar ejemplos de otras instituciones[44]
- Seguir las directrices propuestas por Library of Congress (LoC)
- Utilizar la checklist[45] de que dispone la LoC; un documento muy claro y breve que guía sobre los puntos clave sobre los que tomar decisiones. Ilustrándolo con ejemplos muy claros. En las páginas anteriores hemos recorrido uno a uno cada uno de sus puntos, y hemos mostrado las decisiones tomadas en el caso del perfil de la BNE.
- utilizar herramientas de validación como la desarrollada por el Florida Center para la Library of Congress[46] en 2009.

[44] Algunos ejemplos pueden consultarse en la página de la Library of Congress: http://www.loc.gov/standards/premis/premis-mets.html

[45] Disponible en: http://www.loc.gov/standards/premis/premis_mets_checklist.pdf

[46] Disponible en: http://pim.fcla.edu/validate…En esta página pueden accederse además a otras herramientas que ayudan en: la conversión (a partir de un documento PREMIS-XML se puede crear un documento METS con los elementos PREMIS embebidos ó a partir de un METS-XML se crea un documento PREMIS); la descripción (se crea un PREMIS directamente a partir de un archivo)

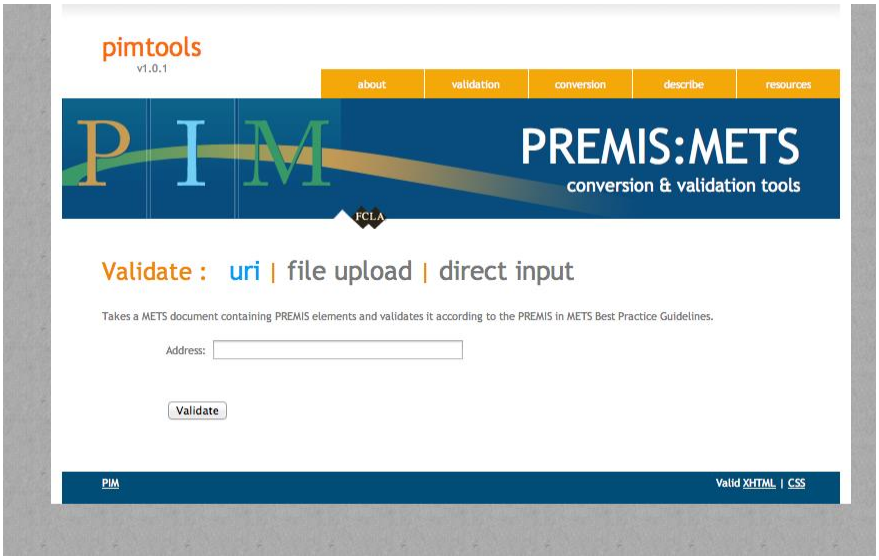

Ilustración 43: Interfaz de la herramienta de validación de un documento METS-PREMIS

En cualquier caso, después de todas estas directrices y ejemplos recomendamos no perder de vista lo siguiente:

- en preservación digital no hay una única solución; a pesar de que podamos seguir lo que van haciendo otros, es inevitable la fase de análisis para adecuar los perfiles y modelos de datos a nuestros casos particulares
- una decisión tomada en un determinado momento pudo tener justificación, sin embargo, nada nos garantiza que sea la última a tomar, ni la más correcta y/o idónea…Todo está abierto al cambio y la revisión. Eso sí, será fundamental siempre registrar nuestras decisiones, y para ello los perfiles, las políticas de preservación y demás documentos son herramientas fundamentales.

Como recoge el informe sobre metadatos de preservación de Lavoie y Garner en su edición de 2013 ahora, más que nunca, falta acumular y consolidar las mejores prácticas; e incluso analizar los costes y beneficios derivados de utilizar los metadatos de preservación e incluso demostrar que ayudan en los flujos de trabajo y la toma de decisiones.

BIBLIOGRAFÍA

CAPLAN, Priscilla. (2009) *Understanding PREMIS*. The Library of Congress. Disponible en: http://www.loc.gov/standards/premis/understanding-premis.pdf. También disponible en español la traducción de MªLuisa Martínez-Conde en: http://www.mcu.es/bibliotecas/docs/MC/PREMIS/Contenido.pdf

JONES; Beagrie. (2001). Preservation Management of Digital Materials: The handbook. Digital Preservation Coalition. Disponible en: http://www.dpconline.org/pages/handbook/docs/DPCHandbook.pdf.

LAVOIE, B & GARTNER, R. (2013). *Preservation Metadata Technology Watch Report*. Digital Preservation Coalition. Disponible en: http://dx.doi.org/10.7207/twr13-03

TERMENS, M. (2013). *Preservación digital*. Barcelona: UOC.

VV.AA. (2008). *Diverse exploding digital universo*. IDC. Disponible en: http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

VV.AA. (2008). *Guidelines for using PREMIS with METS for Exchange*. LoC. Disponible en: http://www.loc.gov/standards/premis/guidelines-premismets.pdf

VV.AA. (2008). *PREMIS Data Dictionary for Preservation metadata.* Version 2.1. LoC. Disponible en: http://loc.gov/standards/premis/v2/premis-2-2.pdf

VV.AA. (2012). *Reference Model for an Open Archival Information System (oais). Recommended practice. Magenta book.* Washington: Consultative Committee for Space Data Systems (CCSDS). Disponible en: http://public.ccsds.org/publications/archive/650x0m2.pdf

VERMAATEN, Sally. (2010). *A checklist for Documenting PREMIS-METS Decisions in a METS profile*. OCLC. Disponible en: http://www.loc.gov/standards/premis/premis_mets_checklist.pdf

ÍNDICES

Índice de ilustraciones