IFLA WLIC 2013
SINGAPORE

Submitted on: June 18, 2013

La construcción de un flujo de trabajo de metadatos sostenible para recursos audiovisuales: el repositorio de preservación digital Medusa de la Biblioteca de la Universidad de Illinois

Spanish translation of the original paper:
"Building a Sustainable Metadata Workflow for Audio-visual Resources: University of Illinois Library's Medusa Digital Preservation Repository"
***Translated by**: Isabel Lozano Martínez, José Carlos Teruel Carrera, Biblioteca Nacional de España, Madrid, Spain*

Ryan Edge
Preservation Unit, University Library, University of Illinois at Urbana-Champaign,
Urbana, Illinois, United States.
E-mail address: edge2@illinois.edu

Myung-Ja Han
Content Access Management, University of Illinois at Urbana-Champaign,
Urbana, Illinois, United States.
E-mail address: mhan3@illinois.edu

Resumen:

Cuando la Biblioteca de la Universidad de Illinois inició el desarrollo de su sistema de repositorio de preservación digital –Medusa- la biblioteca encontró que había muchos recursos audiovisuales que todavía no habían sido catalogados y por tanto eran inaccesibles para los usuarios. De cara a que los recursos estuvieran disponibles para los usuarios finales y para cumplir con el modelo de referencia Open Archival Information System (OAIS), la biblioteca desarrolló un conjunto de medidas que incluyera metadatos descriptivos. El grupo de trabajo del programa de preservación digital del modelo OAIS para archivos multimedia, examinó cuidadosamente las recomendaciones y documentos de mejores prácticas disponibles en ese momento para la catalogación de recursos audiovisuales para elaborar los niveles de cumplimiento de la Sumission Information Packages (SIPs) para audio y metadatos descriptivos. El documento presenta al detalle el proceso de decisión para el estándar de metadatos descriptivos de la Biblioteca y el conjunto de elementos incluidos dentro de la SIPs para audio con su denominación de formatos de archivo, especificaciones de archivo y estructuras de directorio. El documento también discute la comparación, elemento por

elemento, entre PBCore y MODS, y la creación del flujo de trabajo de metadatos descriptivos basado en XML que ya que se aplicó en un proyecto piloto reciente.

Palabras clave: Metadatos, Audiovisual, Preservación, Repositorio de preservación digital, MODS

Introducción

El repositorio de preservación digital Medusa[1] de la Biblioteca de la Universidad de Illinois en Urbana-Champaign proporciona un entorno de almacenamiento que asegura el acceso y uso a largo plazo del contenido digital seleccionado por los gestores y productores de contenido. Como Medusa ha pasado de la planificación a la implementación, los departamentos de la biblioteca han empezado a revisar los flujos de trabajo y los perfiles para proyectos digitales. Hasta este punto, gran parte del proyecto de estructura de directorios y creación de metadatos ha sido dirigida ad hoc, con poca consistencia o documentación. Para proyectos internos de digitalización audiovisual y de cambio de formato del proveedor a gran escala, se necesita actuar urgentemente.

Este informe aporta documentación sobre la planificación e implementación de estas directrices, incluyendo desafíos y soluciones que surgieron en el proceso de recogida y transformación de la información desorganizada de la colección en discretos registros de metadatos descriptivos. También se incluye un análisis del considerado empaquetado del objeto de audio, especificaciones de archivo, convenciones en la denominación y estructura del directorio. El informe discutirá también el proyecto piloto que se benefició de la implementación de la creación de un flujo de trabajo de metadatos descriptivos basados en XML apoyado en estas directrices.

Antecedentes

Como la Biblioteca de la Universidad avanza hacia la implementación de flujos de trabajo del proyecto digital y de directrices de metadatos para el desarrollo del contenido del repositorio, los proyectos de digitalización del pasado reciente deben ser tomados en consideración en este marco de trabajo. Como principal parte interesada en el desarrollo del repositorio de preservación digital Medusa y un significativo cambio de formato del contenido digital y de su propia producción, la Unidad de Preservación ha empezado a explorar sus propias necesidades y desafíos en la definición de las prácticas de empaquetado y de las estructuras funcionales del proyecto de directorio en el entorno Medusa.

En el otoño de 2012, Preservación reunió al Grupo de trabajo del modelo OAIS[2] para archivos multimedia para investigar métodos definidos de organización y descripción del flujo de trabajo para el proyecto de cambio de formato audiovisual, principalmente para ser usado por el Programa de Preservación de Medios[3] y sus proveedores de terceros. Como consecuencia, se recogió el esquema básico de los requerimientos para metadatos descriptivos, especificaciones de archivo, y los perfiles para audio reformateado de la

[1] https://wiki.duraspace.org/display/hydra/Medusa

[2] Grupo de trabajo*:* Josh Harris (Media Preservation Coordinator), Annette Morris (Preservation Reformatting Coordinator), Tracy Popp (Digital Preservationist), Kyle Rimkus (Preservation Librarian), Ryan Edge (Graduate Assistant), Gary Maixner (Graduate Assistant), and Thomas Padilla (Research Assistant).

[3] http://www.library.illinois.edu/prescons/services/media_preservation/media_preservation.html

Submission Information Package (SIP). El grupo de trabajo también comparó PBCore[4] y Metadata Object Description Schema (MODS)[5] con la información necesaria según los niveles de descripción, decidiendo en última estancia exigir MODS y su estándar de metadatos descriptivos.

En la comunidad de preservación existe un amplio uso y familiaridad con el formato de audio WAV. La calidad de "Preservation Master" tiene un perfil de tasa de muestra WAV de 24-bit y 96 KHz, lo que debe considerarse como el estándar para la conversión de analógico a digital sin pérdidas (International Association of Sound and Audiovisual Archives Technical Committee, 2009). El audio es un flujo de bits relativamente lineal y simple, tanto sus metadatos asociados como su preservación son menos exigentes que los de vídeo.

Ya que pueden contener múltiples flujos de audio y texto, además de imágenes en movimiento, con innumerables variaciones de envoltura y de formato códec, el vídeo puede ser exponencialmente complejo. Por esta razón, el grupo de trabajo se ha centrado en primer lugar en el audio digital y en su investigación preliminar y proyecto piloto, abordado en este informe.

El proyecto piloto: La colección de Discos de Transcripción WILL

Como el grupo de trabajo ha tenido en cuenta las especificaciones de archivo y el esquema de metadatos descriptivos para objetos de audio, la escalabilidad y sostenibilidad fueron las principales preocupaciones. Se hizo evidente que, en efecto, una aplicación de investigación directa sería inmensamente beneficiosa en la política de configuración e información. Por lo tanto, se realizó un proyecto piloto para medir el éxito y la completa aplicabilidad en el marco de estudio de las colecciones de audio reformateadas a digital en el pasado y en el futuro. Afortunadamente, esta colección objeto de estudio era una necesidad considerable y urgente. Desde 2011, Media Preservation ha supervisado su primer gran volumen a reformatear: la digitalización de cerca de 6.000 grabaciones de la radio pública local alojadas en más de 3.000 discos de transcripción eléctricos[6] en el Archivo de la Universidad de Illinois. Dado que el sustituto digital de la colección de Discos de Transcripción WILL está en constante crecimiento en el repositorio Medusa –uno de los más valorados- se decidió que recibiría una atención completa e inmediata.

Las grabaciones de WILL están documentadas de forma variada e inconsistente, son ricas en información, y están cargadas de muchos de los potenciales desafíos y enigmas intelectuales que el procesamiento y descripción retrospectivos puedan presentar. La propia información del programa se extrajo de múltiples fuentes a lo largo de muchos años, y fue muy complicada y en gran parte inconsistente, con numerosas omisiones y redundancias. Pero es indicativo de los miles de obstáculos prácticos a los que se enfrentan las bibliotecas en la complicada extracción y alto volumen de colecciones multimedia basadas en la duración temporal, particularmente aquellos indizados por dudosas relaciones, y recopilados por multitud de estudiantes de apoyo a través de los años. Las grabaciones sonoras se realizaron entre 1938 y 1970, con un número de supervisores diverso a lo largo del tiempo. Sólo en la

[4] http://pbcore.org/

[5] http://www.loc.gov/standards/mods/

[6] Un disco de transcripción es un registro fonográfico especial para, o grabado de, emisiones de radio. Generalmente en lo más alto de la lista de cambio de formato en las instituciones, son extremadamente frágiles y a menudo únicos. Para más información respecto a las prestaciones, manejo y conservación de las listas de transcripción, véase: http://www.theaudioarchive.com/TAA_Resources_Disc_Transcription.htm

última década ha habido un registro autorizado de sus contenidos. Esos registros autorizados existieron únicamente en una hoja de cálculo Microsoft Excel, con una gran parte de unidades de datos compartiendo celdas, absolutamente carentes de uniformidad sintáctica: un mar de caracteres, cadenas y valores numéricos en una celda, con puntos y comas, periodos, y espacios utilizados indistintamente. Los discos fueron inventariados por número de disco/ítem, aunque a menudo la distinción entre cara A y B estuviera sin etiqueta u oculta. A falta de mejores registros, la hoja de cálculo del Archivo de la Universidad tuvo que servir como fuente origen de metadatos y base para el empaquetado de objetos de audio y registros MODS. Surgieron complicaciones eventuales de esta precaria dependencia, que discutiremos más adelante en la sección *Límites a la extracción de metadatos.*

Repositorio de Preservación Digital

Principios de diseño:

Usando el paquete de información OASIS como modelo, el grupo de trabajo esbozó un paquete adicional de principios de diseño. Estos objetivos centrales en gran parte existen para subrayar la modularidad y cierto grado de metadatos en todos los niveles de la estructura del directorio. Actuando con sentido común, con enfoque pragmático, se decidió que las convenciones sobre los nombres de los archivos de audio deberían transmitir una cantidad razonable de información técnica y descriptiva ya que era también necesario un identificador único dentro del repositorio, adoptando convenciones del identificador de los repositorios de origen, siempre y cuando estos existan o todavía se consideren relevantes.

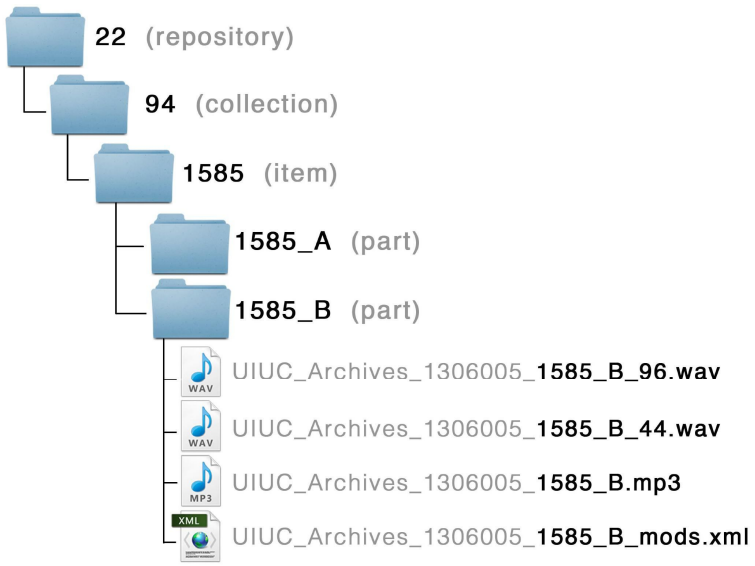
Estructuras de directorios:

En el caso de las colecciones de audio del Archivo de la Universidad, hay una convención de nomenclatura de archivos que Preservación se esfuerza en respetar mientras también discretamente modifica la funcionalidad de preservación. Por ejemplo, "UIUC_Archives_[series]_[item]_[part]" es la plantilla general implementada por el Archivo. “UIUC_Archives” indica el repositorio de origen, seguido de un desglose de series archivísticas (o colecciones) de ítem a parte de ítem, si fuera aplicable. Media Preservation -y proveedores- adjuntarán una etiqueta a este nombre de archivo base para indicar además la calidad del audio “_96.wav” or “_44.wav” for “Preservation” and “Mezzanine”, archivos WAV, respectivamente. Para más información sobre la práctica en la denominación de archivos de Media Preservation en relación a la calidad y estructura de empaquetado, continuar con la sección *Clasificaciones Audio SIP.*.

Las estructuras del proyecto de directorio Media Preservation en el servicio del Archivo intentan seguir esta sintaxis jerárquica, aunque la semántica del directorio interno del repositorio Medusa es divergente y menos intuitiva. Un sustituto de audio digital que lleva el nombre de archivo "UIUC_Archives_1306005_1060_A_96.wav"[7] se localiza como sigue: el Repositorio 22 mantiene todas las colecciones del Archivo de la Universidad, que contiene la Colección 94, un directorio de sustitutos de audio digital derivados de discos ranurados. Dentro de 94 hay paquetes para cada ítem, o disco de audio, en esta colección particular de sonido grabado. El Ítem 60, por ejemplo, tiene dos subcarpetas"1060_A" and "1060_B", uno para cada objeto de audio digital. Ver Figura 1 a continuación como explicación de esta organización.

[7]-digit ID 1306005 se refiere a una clasificación Archon, a menudo con ceros añadidos de acuerdo al patrón ##/##/##. 1306005, por ejemplo, se derivó del grupo de registros 13/6/5. Ver los Archivos de la Universidad (via Archon)): http://archives.library.illinois.edu/archon/index.php?p=collections/controlcard&id=1994

Figura 1. Estructura del proyecto del directorio de audio (Colección de Discos de Transcripción WILL)



Clasificaciones de audio SIP:
Los requisitos explícitos para clasificaciones mínimas, medias y completas de empaquetado de objetos de audio, fueron ideados no solo para Preservación sino también para proveedores actuales y futuros, con potencial para extender como una plantilla para proyectos de medios para todas las bibliotecas universitarias. Las especificaciones fundamentales requieren una carpeta para cada objeto y un identificador único como su nombre de archivo. Este identificador se amplía como nombre de archivo base para todos los otros ítems contenidos dentro del directorio, incluyendo el correspondiente archivo de metadatos descriptivos y archivos de audio derivados.

En la línea de base mínima, se requieren uno o más archivos de audio en formato WAV. Para auténticos ficheros WAV "Preservation Master" (96kHz/24-bit), la combinación del sufijo del nombre de archivo y la extensión debe ser "_96.wav" que indica una velocidad de muestreo de 96 kHz. Para ficheros WAV de calidad de producción "Mezzanine" (44.1kHz/16-bit), la combinación del sufijo del nombre de archivo y la extensión debe ser "_44.wav" que indica una velocidad de muestreo de 44.1 kHz. Además dentro de cada paquete de objetos debe haber al menos un fichero de audio MP3 acompañando a los citados archivos WAV. Esta copia MP3 estará indicada con la extensión de archivo ".mp3".

El perfil de nivel medio SIP para audio incluye metadatos descriptivos MODS XML además de las especificaciones mínimas antes mencionadas. Los archivos de metadatos también llevan el indicador de base de los ficheros de audio anejos (masters y derivados), aunque con el sufijo "_mods.xml" y la extensión de archivo adjuntado al nombre de archivo principal.

El cumplimiento con los requerimientos de nivel completo SIP ve la adopción de un uso apropiado de los campos de metadatos MODS requeridos (examinados en la siguiente sección), la inclusión absoluta de un archivo de audio WAV Preservation Master, y un

archivo con documentación fotográfica (ej.: imagen JPEG), con sufijo "_pd.jpeg" y la extensión de archivo adjuntado al nombre de archivo principal.

Tabla 1. Perfiles de objeto de audio SIP

Atributos	Mínimo	Medio	Completo
Archivos(s)	1 o más arhivos(s) de audio	1 o más archive(s) de audio, archivo(s) de metadatos XML	1 o más archivo(s) audio, archive(s) de audio XML documentación de archivo(s) fotográficos opcional
Formato(s) de archivo	WAV, MP3	WAV, XML, MP3	WAV, XML, MP3, JPEG
Especificaciones de archivo	44.1 kHz velocidad de muestreo, 16-bit	44.1 kHz velocidad de muestreo, 16-bit	96 kHz velocidad de muestreo, 24-bit
Formato de metadatos descriptivos	(*SIP minimo que no require metadatos)*	MODS (version 3.4)	MODS (version 3.4)
Campos de los metadatos requridos	(*no aplicable*)	Repositorio, Creador, Título, Formato, Fecha, Identificador, Tiempo de ejecución	Repositorio, Creador, Título, Formato, Fecha, Identificador, Tiempo de ejecución

Elección de una norma de metadatos

Las demandas de un esquema de metadatos descriptivos y su codificación fueron bastante modestas. La escalabilidad, por encima de la especificidad, estuvo en la máxima operatividad en la fase de planificación. Una fundación de metadatos universal fue el intento que se traduciría a través de proyectos de cambio de formato digital de los medios variables y del contenido. Todos los SIPs de audio que deben incluir metadatos descriptivos se requiere que lleven 7 campos esenciales: Repositorio, Creador, Título, Formato, Fecha, Identificación y Tiempo de ejecución. Como estos elementos son vitales para el descubrimiento y el acceso de los usuarios, casi todo esquema de metadatos descriptivos los mantiene de una manera u otra, desde Dublin Core en adelante. Considerando la naturaleza del cambio de formato audiovisual y de su derivación, sin embargo, estaba claro que BPCORE y MODS eran los competidores principales entre las normas de metadatos disponibles usados en la descripción de los recursos audiovisuales (De Sutter, Notabaert, & Van de Walle, 2006).

Mediante el uso de BPCORE, se ganaría capacidad para expresar todas las ejemplificaciones analógicas y digitales para una determinada entidad. Otro paso más allá con BPCORE permitiría conexiones a los programas generales y series así como también a la colección del Archivo de la Universidad y a posibles sub-colecciones que le pertenecieran, todo comunicado a través de un único registro de metadatos. Sin embargo, el esquema y los vocabularios están concebidos específicamente para los productores de medios, para ser "utilizados como un modelo de datos para la catalogación de medios y de sistemas de gestión de fondos. Como un esquema, permite el intercambio de datos entre las colecciones de medios, los sistemas y las organizaciones". (Corporación para la Difusión Pública, sin fecha).

Mientras BPCORE complementa los flujos de trabajo del cambio de formato de los medios, muchos de sus lenguajes y características están orientados hacia comunidades y recursos de difusión pública, bibliotecas no académicas, instituciones de patrimonio cultural y servicios de preservación de los medios. Ya que BPCORE era increíblemente robusto en la descripción del contenido y de los metadatos técnicos, se determinó que fuera de más alto calibre de lo necesario.

MODS, por otro lado, es un esquema de descripción de metadatos más aerodinámico para los objetos de audio en la biblioteca. No hay necesidad adicional de metadatos técnicos o administrativos dentro del respositorio Medusa debido a su funcionalidad incorporada en los metadatos de preservación: estrategias de implementación (PREMIS). MODS, como metadatos descriptivos, se adapta muy bien, capaz de soportar infomación de procedencia y ofrecer múltiples puntos de acceso. Adecuado para los fines de los proyectos audiovisuales en Medusa, intuitivo para cualquier profesional de biblioteca con competencias básicas de catalogación, el esquema MODS proporciona un lenguaje susceptible para colaborar internamente y con los proveedores. Además la comparación elemento por elemento entre MODS y BPCORE demostró que MODS posee un rico conjunto de semántica que puede ser rediseñado con cualquier otra norma de metadatos en el futuro. (ver Apéndice I)

Tabla 2. Campos mínimos requeridos de los metadatos

Nombre de campo	Nota	Elemento MOD
Repositorio	Departamento que mantiene el recurso.	<origenInfo> <editor>
Creador	Persona primaria/organización asociada con el recurso.	<nombre> (función)
Título	Titulo del recurso.	<tituloInfo> <titulo>
Formato	Físico/ archivo formato del recurso	<fisicaDescripcion> <internetMedioTipo>
Fecha	Fecha de publicación – fecha cuando el programa se emitió originalmente.	<origenInfo> <fechaEmision>
Identificador	Único identificador del recurso.	<relacionadoItem> <identificador>
Tiempo de ejecución	Duración del recurso o duración de toda la serie.	<relacionadoItem> <fisicaDescripcion> <medida>

Tras un examen más minucioso: Preocupaciones adicionales de los metadatos

Los metadatos funcionan como una hoja de ruta para facilitar el uso de los recursos de información: rico en información, bien estructurado y granular adecuadamente los metadatos permite a los profesionales de las bibliotecas gestionar y organizar las colecciones eficazmente (Cole & Han, 2013). La práctica de creación y gestión de metadatos de los recursos audiovisuales ha sido un desafío para muchas bibliotecas ya que los recursos en formatos audiovisuales requieren la captura de tipos de información muy diferentes de los requeridos para los objetos textuales, en particular información técnica y normalmente

constan de muchos ítems (partes) que demandan una norma de metadatos adecuada para apoyar la estructura jerárquica (O'Brien, 2012). Siguiendo la decisión de usar MODS como norma de metadatos descriptivos para todos los recursos audiovisuales, surgió la pregunta de qué información debería ser capturada con estos metadatos. La respuesta llegó con bastante facilidad: el repositorio de preservación digital Medusa se basa en PREMIS, que tiene la capacidad de manejar muchos tipos de información diferentes, ya sea utilizando la semántica de PREMIS o enlazando con los metadatos creados en diferentes normas. La información técnica esbozada en las mejores prácticas[8] de la biblioteca será extraída automáticamente por un software, como JHOVE[9], durante el proceso de envío. De este modo los metadatos MODS contienen estrictamente aspectos descriptivos, incluyendo el identificador del recurso que funciona como un punto coincidente entre el recurso y los metadatos. Sin embargo, la granularidad de los metadatos y el nivel de descripción sigue siendo un asunto para ser examinado más a fondo.

Debido a que el repositorio de preservación digital Medusa funcionará en la preservación a nivel de bit[10], el plan inicial para la creación de metadatos estaba también basado en el mismo principio: la creación de un registro MODS por cada disco físico. Sin embargo, durante el transcurso del proyecto piloto, el grupo se dio cuenta de que había un problema en el logro de este objetivo. Cuando se crean metadatos para cada objeto, la identificación y captura de las relaciones entre las caras del disco, los títulos y los nombres de los programas eran los componentes más críticos de los metadatos. Pero las jerarquías de las relaciones varían programa por programa y título por título. En la mayoría de los casos, un programa se componía de muchas emisiones, y un título puede ser difundido a través de múltiples discos de audio (y caras). En casos poco frecuentes, una de las caras del disco podía contener más de un título de emisión, o una parte. Al final se adoptó otro enfoque: los metadatos MODS deberían ser creados al nivel de título como parte del SIP. Ya que MODS funciona bien para describir objetos compuestos, la información asociada con algunos recursos relacionados puede ser añadida en el nivel de título de los metadatos (Dulock & Cronin, 2009). Los metadatos tienen múltiples elementos <relacionadoItem> para todos los discos asociados y la información del programa padre; la información del ítem (cara del disco) se añade con un atributo “componente” y la información del programa se añade con un atributo “servidor”. La información detallada asociada con cada disco es también capturada en sub-elementos permitidos en el elemento <relacionadoItem> como <títuloInfo>, <nombre>, <parte> <extensión> y como se muestra en la figura 2 debajo.

[8] http://www.library.illinois.edu/dcc/bestpractices/chapter_10_technicalmetadata.html#10.2.3DigitalAudioFiles

[9] http://jhove.sourceforge.net/

[10] Nivel básico de servicios de preservación digital y métodos, la preservación a nivel de bit generalmente dirigirá el almacenamiento seguro y supervisado de los archivos digitales. Para más información en los niveles de preservación digital, ver: http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

Figura 2. Información específica del disco capturada en el elemento <relacionadoItem> del nivel de título de los metadatos MODS

```
<relacionadoItem tipo="componente" ID="disco2ID">
        <tituloInfo>
          <titulo>titulo2</titulo>
        </tituloInfo>
        <nombre tipo="personal">
          <nombreParte>nombre1</nombreParte>
          <funcion>
       <funcionTermino tipo="texto" autoridad="marcrelator">funcion</funcionTermino>
          </funcion>
        </nombre>
        <fisicaDescripcion>
          <extension>13 min.</extension>
        </físicaDescripción>
    <identificador>disco2ID</identificador>
  </relacionadoItem>
```

Proyecto flujo de trabajo

De los datos a los metadatos:

Aceptando que una hoja de cálculo era la única ventana de entrada a la colección, el uso de Extensible Stylesheet Language Transformations (XSLT) para crear metadatos MODS fue otro punto de referencia que lograr simultáneamente a la primera etapa de normalización del nombre del archivo y de los esfuerzos de empaquetado. Con la finalidad de lograr esto, los datos de la hoja de cálculo tenían que ser destilados solo a valores de absoluta necesidad. Elementos adicionales más allá de esos siete fundamentales de potencial valor para los investigadores también fueron incluidos; datos no útiles se dejaron atrás en el proceso de migración. Este nivel de atención exigía un equilibrio entre el tiempo del editor humano (interpretación, formato, división**)** y los procesos de normalización de scripting mejorados**.** Por ejemplo, después de la manipulación manual de formatos de nombres incompatibles en una estricta sintaxis Apellido, Primer nombre personal, el posterior scripting Ruby permitió el análisis sintáctico en el valor de cadena de los campos de Nombre y la asignación de funciones. Esto hizo la eventual transformación de XSLT aun más perfecta al partir los valores de la cadena y asignar tipos de parte del nombre, Ej.: <nombreParte tipo="given"> Timothy </nombreParte> <nombreParte tipo="familia"> Trimble </nombreParte>.

Debido a la secuencia fracturada de partes de emisión y de series, así como a las múltiples entidades almacenadas en una cara del disco, se decidió que los metadatos MODS serían entrenados en un nivel de título. Los identificadores eran asignados a títulos únicos en vez de a cada parte/cara del disco. Estas asignaciones sistemáticas se lograron por medio del scripting Ruby. Este script analiza el título, el hablante y las celdas del Programa para establecer coincidencias exactas dentro de intervalos de quince filas. En el caso de coincidencias exactas en cada celda, se supone que estas entradas representan partes de una emisión segmentada, por lo tanto vinculándolas por un común identificador. Cuando la transformación de XSLT ocurre, en cada parte secuencial del conjunto se incluye un <relacionadoItem> componente, distinguido por el identificador del disco del archivo, así

como en todos los metadatos descriptivos principales recuperables que lo poseen y comparten con la entidad completa MODS.

Transformación

Las transformaciones exitosas de XSLT exigen información rígida en conformidad con su modelo. Esencialmente, la transformación de los datos comenzó cuando la hoja de cálculo fue reformateada dentro de un registro XML completo. Este archivo entonces funciona como un intermediario. En el formato XML, estos datos pueden entonces ser abiertos en un editor (Ej.: oXygen), donde se asocia con el archivo casero Excel-to-MODS XSLT que rediseña números de columna como identificadores en la creación de elementos MODS, traducción de cada fila como una salida de metadatos MODS en un directorio. Hay un considerable grado de personalización sintáctica y de consideraciones específicas en las grabaciones de emisiones WILL en el trabajo dentro de XSLT. Por ejemplo, los valores <nombre-tipo="personal"> son divididos en partes de apellido primer nombre. Además de los siete elementos requeridos, unos cuantos atributos han sido añadidos para facilitar mejor las únicas propiedades de esta colección local. Por ejemplo, por defecto "Urbana-Champaign, Illinois" los valores de ubicación para producciones locales se colocaron en el campo <origenInfo> <lugar>. Por las demandas mencionadas de MODS para llevar relaciones jerárquicas, la elección de títulos de entidades e identificadores se transmiten como componentes relacionados mientras que la información de series se encamina como el servidor o colección principal.

Limitaciones a la extracción de metadatos

El objetivo del proyecto piloto era demostrar la capacidad de las directrices del empaquetado de los objetos de audio en la extracción del audio digital y la anterior catalogación de datos sin procesar y a veces de fuentes poco refinadas. Aquí aprendimos una lección significativa: ¿cómo puede uno verificar que un registro dado (en este caso una hoja de cálculo) es de hecho la máxima representación de una colección multimedia basada en un tiempo inmenso? En la catalogación retrospectiva de tales colecciones, muchas veces uno no puede estar seguro de que todos los sustitutos digitales contienen el contenido que ellos pretenden contener. Por lo tanto, uno tiene que improvisar.

Las revisiones del control de calidad de los anteriores medios de preservación de la colección fueron informales, mediante un muestreo aleatorio a la llegada de las transferencias digitales desde el proveedor, en lotes. Quizá porque esto no fue un muestreo adecuado, se sugirió erróneamente un acuerdo que fue ofrecido por el proveedor sobre los datos del Archivo de la Universidad. Sin embargo, antes de unirse los registros de metadatos a sus respectivos SIPs, se descubrió una importante discrepancia en los primeros lotes del proveedor, la identificación incorrecta de los identificadores del archivo en los nombres de fichero enviados por el proveedor. Parece que en un número de los primeros discos, había distinciones inconsistentes entre las etiquetas de las caras A y B. Por lo tanto, en unas cuantas decenas de casos, el registro del archivo y la asignación técnica del proveedor de un identificador (Ej.: 1082_A) y el sustituto identificador/nombre de archivo se contraponían. Identificar y revertir todos los desacuerdos A/B en una colección por encima de 6000 archivos de audio estaba fuera de cuestión. La supervisión del flujo de trabajo certifica que estos errores hace tiempo que han sido resueltos, por lo que en una colección de este volumen, unas pocas decenas de errores podrían parecer incluso insignificantes, si se mitigan correctamente.

Por otro lado, basar el empaquetado de miles de registros de metadatos en un documento no confirmado es siempre un riesgo. Aun así la creación de metadatos tuvo que proceder, sin los errores existentes de ajuste a ser inamovible. Con el fin de hacer esta colección disponible a los usuarios, se tomaron decisiones difíciles, algunas quedan fuera de las mejores prácticas señaladas con asteriscos a lo largo del camino. Suscribiendo a Greene y Meissner "Más producto, menos proceso" (2005), el enfoque de eficiencia del tratamiento archivístico de atrasos, se decidió aflojar el directorio de objetos para todas los posibles disparidades en la problemática gama de (discos # 1- 200). En lugar de una unión rígida para cada objeto de audio contenido en el disco, todos los ficheros MODS XML se ponen en el nivel de disco/ejemplar, permitiendo al usuario final el descubrimiento y la interpretación en tales casos de disconformidad. Por ejemplo el directorio del disco "12" contendría metadatos MODS para ambos títulos A/B además de paquetes de objetos "12_A" y "12_B". También se acordó que las notas de resumen de la colección final deberían reconocer este error y la información con respecto a la solución del empaquetado flexible. De nuevo, este ablandamiento de las mejores prácticas no es una recomendación amplia, pero vale la pena reconocerlo como un medio para un fin. Estas son excepciones en cada colección, distintos problemas en cada colección de biblioteca requerirán únicas y a veces soluciones descuidadas.

Lecciones aprendidas

Los metadatos llegan a ser un componente esencial en la gestión del ciclo de vida de la colección de una biblioteca porque permiten el acceso y la preservación de los recursos de la biblioteca. En la ejecución del proyecto de repositorio digital, se aprendieron un número de lecciones sobre la creación de un flujo de trabajo de metadatos escalable y sostenible, que los autores creían que pudiera ser aplicable a proyectos de biblioteca similares.

Valoración e identificación de las necesidades de los metadatos

Con el fin de crear metadatos que puedan ser reutilizados y rediseñados en el futuro, éstos deberían ser creados con una norma con un amplio apoyo comunitario. La valoración e identificación de las necesidades de los metadatos ayudará a escoger la norma que funcione mejor para el proyecto. Para este proyecto de repositorio digital de la Biblioteca de la Universidad de Illinois, los metadatos apoyan el acceso y deberían incluir la información de procedencia del recurso. Basados en esas necesidades, los siete elementos de metadatos fueron necesarios para los metadatos descriptivos. Sin embargo una necesidad de preservación audio-visual y de metadatos técnicos y administrativos fue incluida en la infraestructura del repositorio de PREMIS. Después de comparar las dos normas de metadatos ampliamente utilizados, MODS y PBCORE, el grupo decidió usar MODS. Por las necesidades específicas, esta preferencia se ha basado en gran parte por su rica semántica y su flexibilidad al describir las relaciones entre los recursos.

Construcción de un flujo de trabajo de metadatos sostenible y escalable

Las bibliotecas hoy en día deben dirigir los metadatos creados por y para muchas partes interesadas y en muchos formatos diferentes. En muchos casos, los metadatos se crean en sistemas de bases de datos locales en un formato que no se ajusta a ninguna norma. Para trabajar con metadatos de diferentes tipos y cualidades, debería disponerse de un flujo de trabajo de metadatos sostenible y escalable. La Biblioteca de la Universidad de Illinois es una de las primeras en adoptar las tecnologías de la información en los flujos de trabajo de metadatos, en particular XML y tecnologías relacionadas. Puesto que la mayoría se crean localmente o son proporcionadas por proveedores de metadatos pueden ser fácilmente

exportadas o están en Microsoft Excel, XML, se utilizó para transformar y mejorar los metadatos. Sin embargo para que las tecnologías XML funcionasen mejor, la intervención humana fue crucial, Ej.: la calidad de los metadatos es mejor servida por alguien que conozca la colección y pueda tomar decisiones en la limpieza y normalización de los datos en Excel.

Compartiendo la creación de metadatos y la ejecución de decisiones con las partes interesadas

Como los recursos que una biblioteca recoge provienen de muchas diferentes fuentes, incluyendo unidades de campus, escolares, proveedores y editores, la creación de metadatos y las decisiones de implementación deberían ser compartidas con las partes interesadas. Dependiendo de las fuentes que producen los recursos, las necesidades de los metadatos, el grupo de usuarios principales y la manera en que los recursos son almacenados y accesibles, pueden variar. La unidad de catalogación de la biblioteca no es la única responsable de la creación de metadatos y de proporcionar servicios de acceso. En cambio, la consulta con otros grupos que necesitan decisiones y guías de metadatos y la provisión de tecnologías de metadatos accesibles se han convertido en una responsabilidad de la unidad de catalogación.

Referencias

Cole, T. W. & Han, M.J. (2013). *XML for Catalogers and Metadata Librarians*. Westport, Conn. : Libraries Unlimited.

Consultative Committee for Space Data Systems (CCSDS). (2012). Reference Model for an Open Archival Information System (OAIS), Magenta Book. Available at http://public.ccsds.org/publications/archive/650x0m2.pdf

Corporation for Public Broadcasting. (n.d.). *PBCore: About PBCore*. Retrieved from http://www.pbcore.org/about/

De Sutter, R., Notebaert, S., & Van de Walle, R. (2006). Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries. *Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science* (Volume 4172, 220-231). Berlin Heidelberg: Springer.

Dulock, M. & Cronin, C. (2009). Providing metadata for compound digital objects: Strategic planning for an institution's first use of METS, MODS, and MIX. *Journal of Library Metadata*, 9:3-4, 289-304.

Greene, Mark A., & Meissner, Dennis. (2005). More Product, Less Process:
Revamping Traditional Archival Processing. *The American Archivist*, Volume 68:Fall/Winter 2005, 208-263. Available at
http://archivists.metapress.com/content/c741823776k65863/fulltext.pdf

International Association of Sound and Audiovisual Archives Technical Committee. (2009). *Guidelines on the Production and Preservation of Digital Audio Objects*. Ed. by Kevin Bradley. Second edition 2009. (= Standards, Recommended Practices and Strategies, IASA-TC 04). Retrieved from www.iasa-web.org/tc04/audio-preservation

O'Brien, J. R. (2012). Sound Bytes: Audio Metadata Standards in Slightly More Than Six Seconds. A Report of the ALCTS Metadata Interest Group Program, American Library Association Annual Conference, New Orleans, June 2011. *Technical Services Quarterly*, 29:3, 217-219.