5ta. Jornada sobre Bibliotecas Digitales Universitarias
"El ciclo del conocimiento en el entorno académico"
8 y 9 de Noviembre de 2007

Una experiencia con el estándar XMP de Adobe y el software de Biblioteca Digital Greenstone

D. H. Biset, A. T. Chávez Flores

Centro de Información, Centro Atómico Constituyentes, Comisión Nacional de Energía Atómica

Resumen

Una de las condiciones del Proyecto Piloto de Preservación a Largo Plazo que se viene desarrollando en el Centro de Información del Centro Atómico Constituyentes (CICAC), es que los archivos producidos en el origen del proceso de digitalización, contengan los metadatos de su propia descripción, y que los mismos puedan ser extraídos de manera automática al procesarse los documentos con un software de biblioteca digital. Para ello se estudiaron la estructura y las funcionalidades de los archivos XMP que se encuentran embebidos en los documentos PDF, y se crearon sucesivas colecciones de prueba en Greenstone 2.72. Del análisis de los archivos XML producidos por ese programa, se concluyó que no realizaba una correcta extracción de los metadatos embebidos en los documentos.Uno de los especialistas en esta aplicación, aportó una solución al problema, creando un nuevo plug in que cumple con la función requerida.

Antecedentes

El Centro de Información CAC (CICAC) es una de las Unidades de Información de la Comisión Nacional de Energía Atómica. Cuenta con un fondo documental especializado en temas de los usos pacíficos de la energía nuclear y ciencias relacionadas (materiales, física, química, ensayos no destructivos, etc.).

Como país miembro del Sistema Internacional de Información Nuclear (INIS, International Nuclear Information Systems), debe realizar el envío de la información producida en el país para su incorporación a la base de datos del INIS. A tal fin envía los registros bibliográficos junto con el texto completo de los documentos, en aquellos casos en que los derechos de autor lo permitan. Inicialmente se enviaba una versión impresa a partir de la cual se generaba una microficha. Posteriormente el INIS comenzó a requerir el envío del texto en formato digital. Por tal motivo, a partir de 1999, el personal del CICAC se vio en la necesidad de adquirir experiencia en el escaneo de documentación impresa, dado que para entonces no todos los documentos que se debían enviar estaban disponibles en formato digital.

En la siguiente tabla se indican los formatos y software utilizados.

Período	Formato	Software
1999-2002	TIFF CCITT Group IV	PixTools
2003-presente	PDF	Adobe Acrobat™

En el año 2001 se decide realizar el escaneo de todas las tesis de Maestría en Ciencia y Tecnología del Instituto de Tecnología "Jorge A. Sábato", a fin de brindar acceso a los textos completos a través de la Intranet institucional.

Es a partir de la participación en una reunión internacional sobre la preservación a largo plazo de los documentos nacidos en formato digital[1], que se decide comenzar a trabajar con la normativa disponible a nivel internacional para tal fin. Es así como se procede a seleccionar información de valor histórico para la institución para conformar un Proyecto Piloto de Preservación a Largo Plazo. Debido a que se seleccionaron tres colecciones pequeñas, se tomó la decisión de

[1] "*Long Term Archiving of Digital Documents in Physics*". Lyon, Francia, 5-6 de noviembre de 2001.

realizar todos los procedimientos en el CICAC a fin de ganar experiencia en el tema en lugar de contratar un servicio de terceros.

Las colecciones seleccionadas para el Proyecto Piloto fueron las siguientes:

Nombre de la colección	Cantidad de ejemplares	Cantidad de páginas
Boletín Informativo CNEA	16	659
Informes CNEA	501	9707
Memoria CNEA	12	612

Contexto, consignas y tareas de la investigación

En el marco del mencionado Proyecto se ha investigado y trabajado con diferentes variables: procesos, estándares, normas y aplicaciones informáticas, entre otras. Estos componentes se incorporan e interrelacionan en un flujo de trabajo que se representa en la Fig. 1

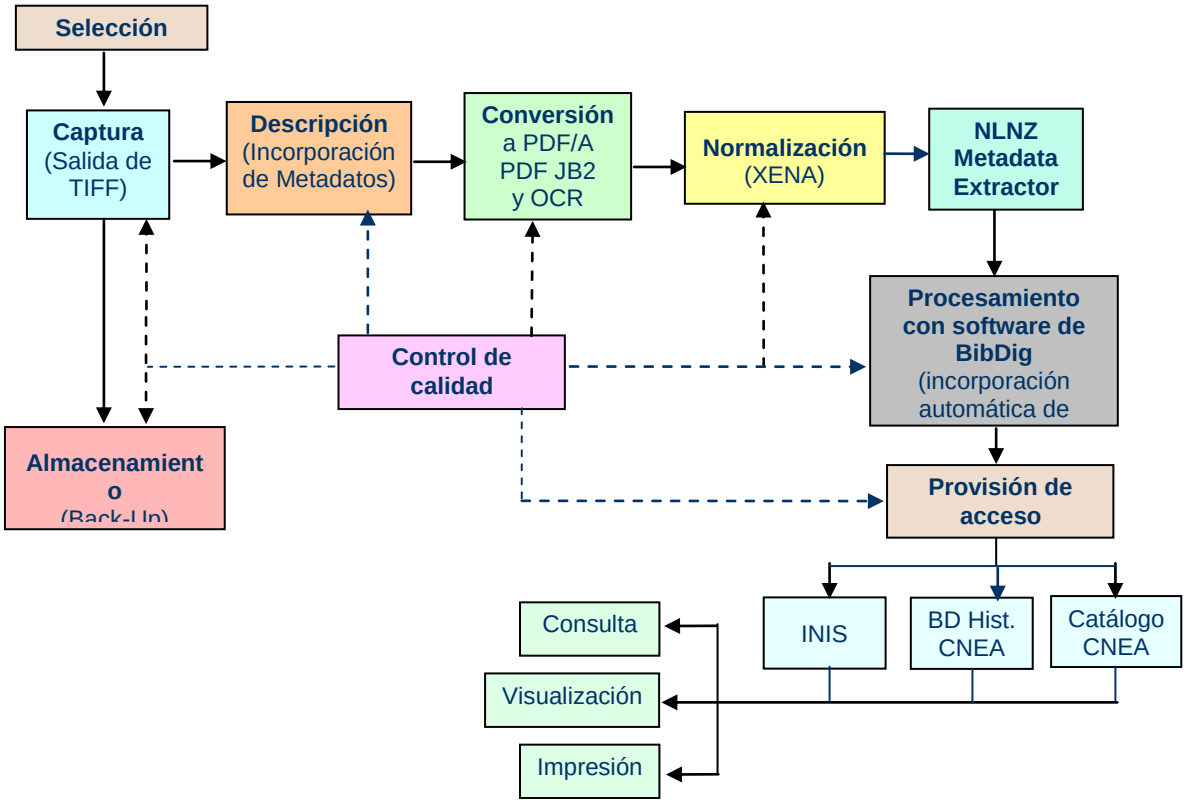

Fig. 1. Flujo de trabajo

Considerando que dos consignas relevantes de este proyecto son:

1. realizar pruebas de inclusión de metadatos en los documentos originados en el escaneo, en el inicio del flujo de trabajo del proceso de digitalización, y
2. que los mismos sean extraídos de manera automática por el software de biblioteca digital que se decida utilizar para poner a disposición de los usuarios la documentación digitalizada,

la experiencia expuesta en el presente trabajo, se centró en el componente "Descripción" del diagrama de la Fig. 1, ya que la misma se realiza sobre el propio documento digital, y no como un registro externo al mismo. Esa información se registra y guarda en un archivo XMP (eXtensible Metadata Platform), que todo documento PDF lleva incorporado en sí mismo.

Durante la realización de las sucesivas pruebas, surgió la necesidad de contar con mayores conocimientos sobre la estructura y el funcionamiento de los metadatos de un archivo PDF, por lo que se realizó una búsqueda y selección de bibliografía sobre PDF, XMP y RDF (Resource Description Framework), normativas que dan marco y conforman los metadatos en un documento elaborado, procesado o modificado con las aplicaciones de Adobe.

El estudio de las especificaciones de estas aplicaciones y normas condujo a realizar diferentes y sucesivas pruebas con un grupo de documentos PDF a fin de:

1. Estudiar las diferencias y relaciones entre "Metadatos de documento y "Propiedades de documento" en Adobe Acrobat (se usó la versión 6.0 de Adobe Acrobat Professional)
2. Estudiar las relaciones entre las interfaces "Descripción" y "Avanzado" en un documento de Adobe Acrobat
3. Estudiar las relaciones existentes entre los esquemas que contiene (por defecto) un archivo XMP: pdf, xap, xmp y dc

Los pasos que se llevaron a cabo fueron los siguientes:

- Carga de metadatos en un documento PDF
- Estudio del archivo fuente XMP y la especificación RDF
- Extracción, modificación y reemplazo del archivo XMP de un documento Adobe Acrobat.
- Ingreso y realización de pruebas con el software de biblioteca digital Greenstone, utilizando las sucesivas modificaciones realizadas en los documentos PDF de prueba
- Extracción y estudio de los archivos XML generados para cada documento procesado en Greenstone
- Comparación y análisis entre los archivos XML de Greenstone y los archivos XMP de Adobe Acrobat.
- Registro de los resultados y redacción de un informe sobre las pruebas de procesamiento de archivos PDFs con Greenstone

XMP

XMP de Adobe es una tecnología de etiquetado que permite embeber los metadatos de un archivo, dentro del mismo archivo. Con XMP las aplicaciones de escritorio y el back-end de los sistemas de publicación consiguen un método común para capturar, compartir y desarrollar el aprovechamiento de estos metadatos, posibilitando un proceso de trabajo más eficiente, la automatización de un flujo de trabajo y la gestión de administración de los derechos de autor, entre otras posibilidades.

XMP permite la integración de metadatos, ofreciendo creación de contenidos para embeber información significativa sobre los proyectos en desarrollo, con estándares basados en la construcción de bloques para el desarrollo de soluciones que optimicen el flujo de trabajo. Descripciones significativas, títulos, palabras claves, y actualización de autores y de información sobre derechos de autor, pueden ser capturadas en un formato comprensible, tanto por un usuario como por una aplicación de software, equipo de hardware y aún por formatos de archivo.

XMP provee un estándar compatible con las normas W3C de etiquetados de archivos con metadatos, y es una tecnología open-source, de libre acceso a los desarrolladores, lo cual significa que la comunidad de usuarios se beneficia con las innovaciones aportadas por desarrolladores de

todo el mundo. Además de las de Adobe, un creciente número de terceras aplicaciones, ya soportan XMP.

El diseño constructivo de XMP fue pensado en función de hacerlo fácilmente extensible, para que pueda albergar esquemas de metadatos existentes, particularmente para la adición de esquemas personalizados. XMP soporta la extensión por adición de esquemas estándar, para la adición de esquemas específicos de aplicación o privados, y para la actualización de nuevas versiones de esquemas. Las especificaciones XMP brindan las directrices que las aplicaciones deben seguir, a fin de que un sistema de software permita leer, modificar, escribir y serializar las extensiones de metadatos dentro del formato XMP, asegurando el acceso a los metadatos por medio del panel de 'Metadatos del documento' o 'Propiedades de documento', en una aplicación de Adobe, durante todas y en cualquiera de las actividades que conforman un proceso o un flujo de trabajo.

Algunos de los esquemas y estándares factibles de ser utilizados en XMP son los siguientes:

- AdsML: estándar de la industria creado para uniformar la negociación, el trámite, la creación, el cobro, el seguimiento, etc., de todas las operaciones involucradas en la publicación de publicidad impresa y online.
- Creative Commons: organización no gubernamental que desarrolla planes para ayudar a reducir las barreras legales de la creatividad por medio de nueva legislación y de las nuevas tecnologías.
- Digital Image Submission Criteria (DISC): estándar que define un grupo de metadatos para ser adjuntados a imágenes digitales para su presentación y publicación en medios de comunicación gráficos.
- Dublin Core Metadata Initiative (DCMI, Ohio, Estados Unidos): desarrolladores y auspiciantes de un modelo de metadatos diseñado específicamente para proporcionar un vocabulario de características "base", capaces de proporcionar la información descriptiva básica sobre cualquier recurso, sin que importe el formato de origen, el área de especialización o el origen cultural.
- International Press Telecommunications Council (IPTC): consorcio que agrupa a las más importantes agencias de noticias y empresas de comunicación. Desarrollan y mantienen estándares técnicos para mejorar el intercambio de noticias que son usadas por las mayores agencias de noticias del mundo. Definió un conjunto de atributos de metadatos que pueden ser aplicados a imágenes, y que tuvo un importante avance en 1994 cuando Adobe implementó una especificación para introducir los metadatos en archivos de imágenes digitales, conocidos como "IPTC headers" (encabezados IPTC).
- Picture Licensing Universal System (PLUS) Coalition: sistema de estándares para la industria, creado y aprobado por una coalición de fotógrafos, ilustradores, publicistas, diseñadores gráficos, anunciantes, representantes de artistas y operadores del mercado del arte.
- Publishing Requirements for Industry Standard Metadata (PRISM): Proyecto para el que se desarrolló un módulo de tablas de contenido para revistas electrónicas. Es un estándar XML de un vocabulario de metadatos para la industria editorial que facilita la agregación y sindicación de contenido digital.
- World Wide Web Consortium (W3C): el consorcio internacional que produce estándares para la World Wide Web.

XMP no es un esquema de metadatos, sino, más bien, un marco de trabajo extensible construido a partir de RDF, que puede ser usado para representar cualquier número de esquemas, algunos de los cuales son estándares, como Dublin Core, otros pueden ser recomendaciones (esquemas para una mejor gestión documental) y algunos pueden ser definidos y usados por particulares o por un segmento específico de una industria para sus necesidades.

RDF

RDF es un marco para describir recursos e intercambiar metadatos. Está construido en base a las siguientes reglas:

- Un **recurso** es cualquier cosa que puede tener un URI, esto incluye todas las páginas web, todos los elementos individuales de cada documento XML y mucho más.
- Una **propiedad** es un recurso que tienen un nombre y que puede usarse como una propiedad, por ejemplo autor o titulo. En muchos casos todo lo que importa en realidad es el nombre, pero una propiedad necesita ser un recurso de forma tal que pueda tener sus propias propiedades.
- Una **sentencia** o **valor de la propiedad** consiste en la combinación de un recurso, una propiedad y un valor. Estas partes son conocidas como el sujeto, predicado y el objeto de la sentencia.

RDF provee un modelo de datos, y una sintaxis para que partes independientes puedan intercambiarse y usarse. Está escrito en XML y es parte de las actividades que sobre la Web semántica desarrolla el World Wide Web Consortium (W3C), siendo una recomendación del mismo.

Por el uso de XML, la información almacenada en un RDF puede ser fácilmente intercambiada entre diferentes tipos de computadoras que usen diferentes sistemas operativos y lenguajes de aplicación.

Diferentes elementos conforman un RDF. Los principales son:

- elemento raíz**,** <rdf:RDF> que define al RDF como un documento XML y hace referencia a un nombre de espacio.
- elemento <rdf:Description> que identifica a un recurso con sus atributos relacionados, conteniendo los elementos que lo describen (autor, título, país, publicación, fechas, precio, etc.) y que son definidos en un *espacio de nombre*, externo a RDF, que no es parte de RDF (http://purl.org/dc/elements/1.1/, por ejemplo, para referirse al sitio donde residen las especificaciones del esquema de metadatos de DC.)

A su vez, el esquema de RDF incorpora otros elementos: <rdf:Bag>, <rdf:Seq> y <rdf:Alt> que se aplican a determinados tipos de metadatos.

- El elemento <rdf:Bag> es usado para describir una lista de valores, que no son susceptibles de encontrarse ordenados, pudiendo contener valores duplicados (aplica en el metadato <dc.subject> de DC, por ejemplo).
- El elemento <rdf:Seq> es usado para describir una lista de valores, que se presentarán de manera ordenada, pudiendo contener valores duplicados (aplica en el metadato <dc.creator> de DC, por ejemplo).
- El elemento <rdf:Alt> se utiliza para describir una lista de valores alternativos, de los cuales se seleccionará uno (aplica en el metadato <dc.title> de DC, por ejemplo).

¿Porqué no usar una simple DTD o una descripción XMLSchema? La respuesta simple es porque los marcos de trabajo que estos ofrecen son insuficientes. Por ejemplo, si se necesitan usar dos esquemas diferentes: uno que represente información básica, tal como palabras claves, y otro una lista de personas que hayan aprobado el documento. La estructura básica de un RDF (en este caso un <rdf:bag>), puede ser usada en ambos casos. Sin RDF, la estructura de los datos y cómo representarlos en XML debería estar descripta en cada esquema, potencialmente de una manera diferente, haciendo mucho más difícil el procesamiento de los metadatos.

Conformación de un archivo XMP

En la Fig. 2 se pueden apreciar los esquemas de metadatos característicos de un archivo XMP, contenidos entre las etiquetas <rdf:Description> y </ rdf:Description>

Estructura del contenido XML en un paquete XMP:

- Instrucciones de procesamiento XML comprendidas en el envoltorio del paquete XMP.
- Elemento más extremo, contiene un elemento **x:xmpmeta**, el cual contiene un elemento simple **rdf:RDF**
- Elemento **rdf:RDF** contiene uno o más elementos **rdf:Description**.
- Elemento **rdf:Description**, hace referencia a un esquema particular. Requiere un atributo **rdf:about** que puede ser usado para identificar el recurso que el XMP describe. Puede estar vacío o ser una URI (Universal Resource Identifier) basada en una UUID (Universally Unique Identifier) abstracta[2].
- Propiedades XMP de tipo simple
- Propiedades XMP de tipo array sin ordenar: una lista de valores en la cual el orden no es significativo
- Propiedades XMP de tipo array ordenada: analista en la cual el orden es importante
- XAP(eXtensible Authoring and Publishing), metadatos internos usado en versiones anteriores, mantenidos por razones de compatibilidad.

Fig. 2. Partes constituyentes de un archivo XMP

[2]Una URI puede clasificarse como un localizador, un nombre, o ambos. El término 'Universal Resource Locator' (URL) designa al subconjunto de URIs que, además de identificar un recurso, proveen una manera de ubicar el recurso describiendo su mecanismo primario de acceso (por ej. su dirección de red). El término 'Uniform Resource Name' (URN) ha sido históricamente usado para referir a URIs bajo el esquema 'urn', que deben permanecer globalmente únicas y persistentes aun cuando el recurso deje de existir o se vuelva no disponible. Los UUID, resuelven los problemas de persistencia, asignación y unicidad, pero resultan inadecuados para la interacción entre repositorios debido a la carencia de un mecanismo de resolución global. Se crea por la combinación de tiempo y dirección de la tarjeta de red Ethernet. Es un número aleatorio si no se dispone de uno fijo en la tarjeta, uuid:045455gh-3fgf-5ghg-565f-4343343cgcvc, por ej. Este identificador es único, lo que permite crear identificadores sin la utilización de sistemas centralizados como los dominios de nombres.

En el siguiente cuadro se aprecian los esquemas de metadatos incorporados en el archivo XMP de la Fig. 2, y el 'Namespace' asociado a cada uno de ellos.

Esquema	Namespace o Espacio de nombre
pdf	http://ns.adobe.com/pdf/1.3/
xap	http://ns.adobe.com/xap/1.0/
xapMM	http://ns.adobe.com/xap/1.0/mm/
Dublin Core	http://purl.org/dc/elements/1.1/

Con este marco referencial respecto de los esquemas de metadatos y las posibilidades que el archivo XMP ofrece, se encaró el procesamiento de documentos, a fin de testear y corroborar la factibilidad de una de las consignas del proyecto: que los metadatos ingresados en las etapas iniciales del flujo de trabajo sean extraídos de manera automática por el software de biblioteca digital que se decida utilizar.

Carga de metadatos en un documento PDF

En la versión 6.0 de Adobe Acrobat Professional con la que se trabajó, hay dos vías para realizar la carga de metadatos. Una es a través de Archivo → Propiedades de documento (Fig. 3)

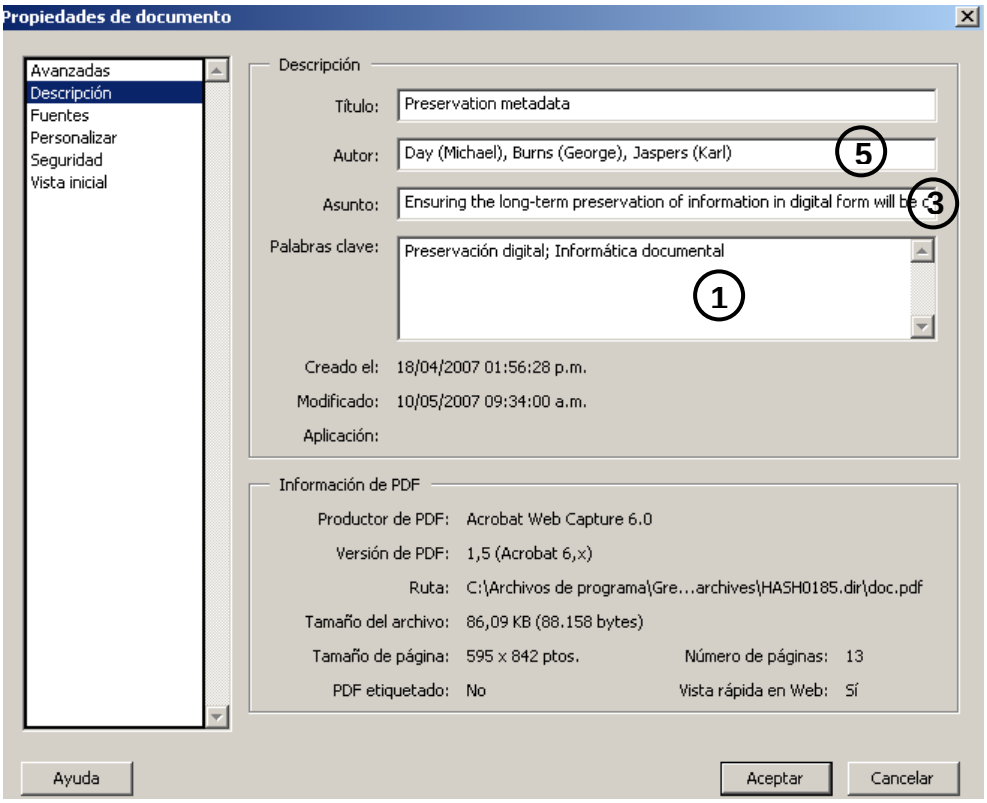

Fig. 3 Panel de Propiedades de documentos

Otra posibilidad es desde Avanzadas → Metadatos de documento (Fig. 4)

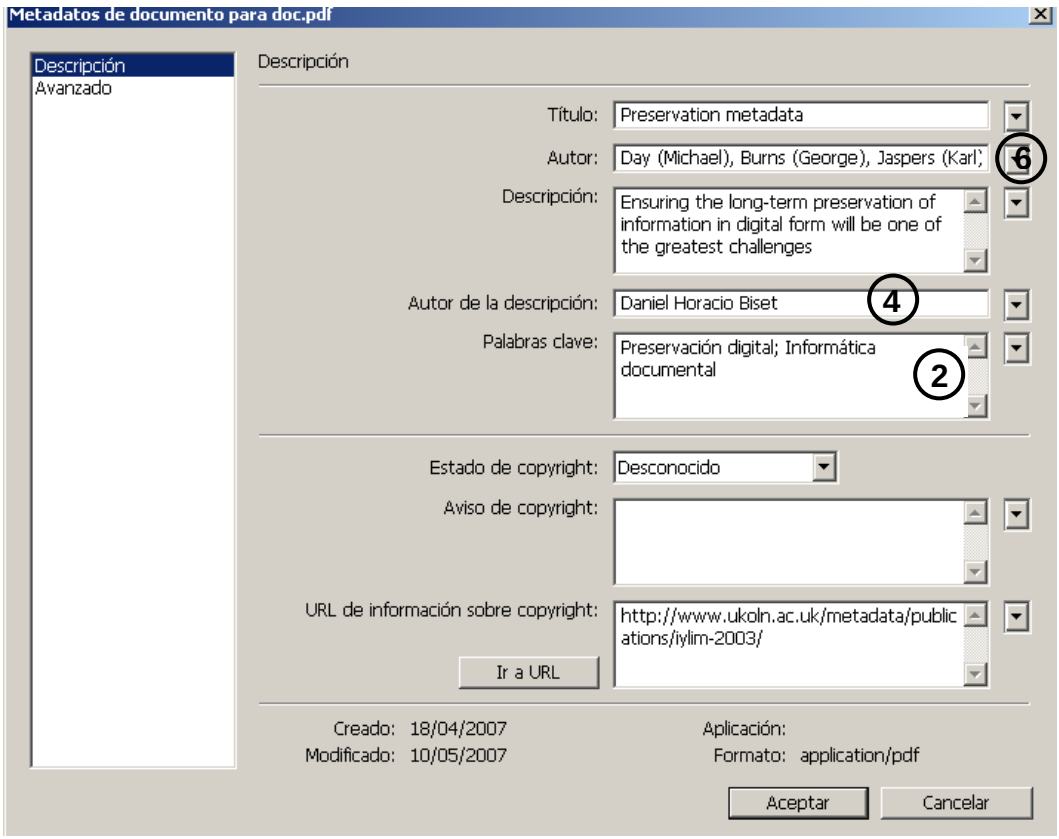

Fig. 4 Panel de Metadatos de documentos

Dado que, por defecto, la especificación XMP opera con diferentes esquemas de metadatos, fue necesario aclarar cuestiones relacionadas con la semántica en la definición y relación entre los nombres de las cajas de texto y las estructuras de metadatos de dichos esquemas.

A título de ejemplo, en el siguiente cuadro se muestran las diferencias entre los términos utilizados por los esquemas de metadatos PDF y DC, y en que caja de texto se ingresa el dato respectivo, con la correspondiente referencia ya sea a la Fig. 3 o a la Fig. 4.

Ref.	PDF	DC	En el Panel	Nombre de la caja
1	Keyword		Archivo → Propiedades de documento	Palabras clave
2		Subject	Avanzadas → Metadatos de documento	Palabras clave
3	Subject	Description	Archivo → Propiedades de documento	Asunto
4			Avanzadas → Metadatos de documento	Descripción
5	Author	Creator	Archivo → Propiedades de documento	Autor*
6			Avanzadas → Metadatos de documento	Autor**

* Si hay varios, desde esta caja se cargan como una sola ocurrencia tanto en el esquema PDF como en el DC
** Si hay varios, desde esta caja se carga solo el primero en el esquema PDF y se abren las ocurrencias necesarias en el DC

Los paneles para la incorporación de metadatos varían de acuerdo a la versión de Adobe Acrobat que se utilice.

El panel 'Avanzado': modos Resumen y Fuente

En la Fig. 4 se puede apreciar el panel 'Descripción'. En la Fig. 5 se aprecia el Panel 'Avanzado', en el modo 'Resumen'.

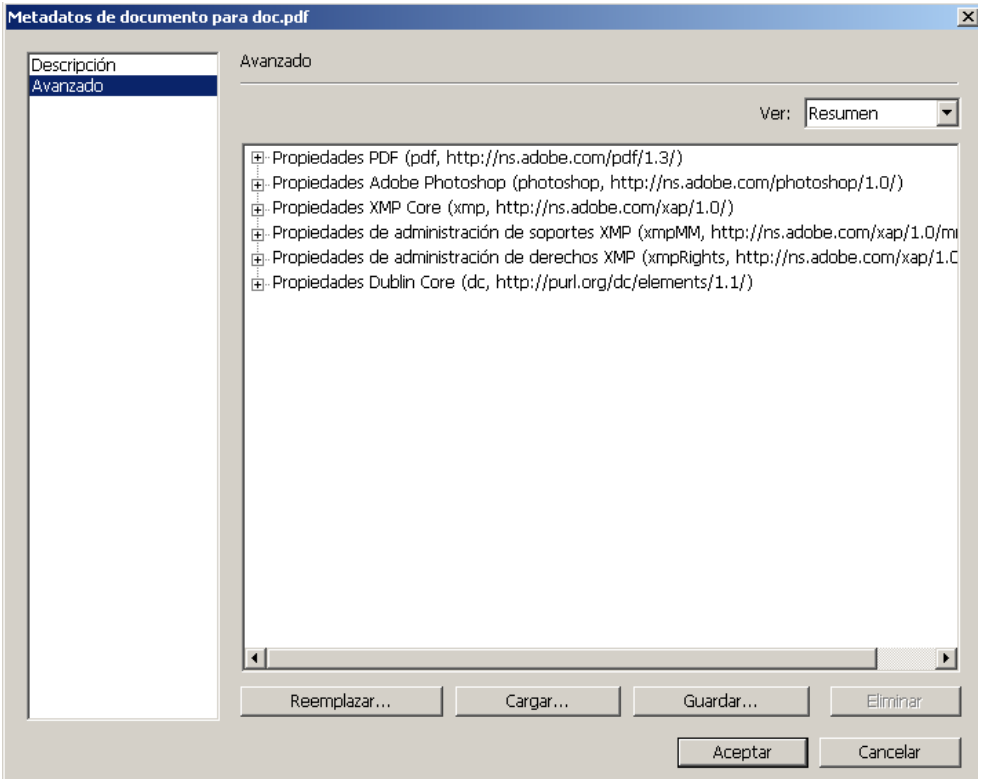

Fig. 5. Panel ‘Avanzado’, modo ’Resumen’

Cada una de las líneas ‘Propiedades…’ corresponde a un conjunto o esquema particular de metadatos, qué, como lo indica el símbolo ⊞, se puede desplegar para visualizar su contenido. Por ejemplo, como se ve en la Fig. 6. Esta es una manera de presentar la información contenida en el archivo XMP que todo documento PDF alberga dentro de sí mismo.

Fig. 6. Metadatos del esquema DC desplegados

Si desde el panel de la Fig. 3 en el desplegable ‘Ver’ se selecciona el modo ‘Fuente’, se accede al archivo XMP del documento activo (Fig. 7)

```
      <?xpacket begin='ï»¿' id='W5M0MpCehiHzreSzNTczkc9d'?><?adobe-xap-filters esc="CR"?>
<x:xmpmeta xmlns:x='adobe:ns:meta/' x:xmptk='XMP toolkit 2.9.1-13, framework 1.6'>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:iX='http://ns.adobe.com/iX/1.0/'>

 <rdf:Description rdf:about='uuid:1d862a03-e87a-414c-a3bc-438844b8b643'
  xmlns:pdf='http://ns.adobe.com/pdf/1.3/'>
 </rdf:Description>

 <rdf:Description rdf:about='uuid:1d862a03-e87a-414c-a3bc-438844b8b643'
  xmlns:xap='http://ns.adobe.com/xap/1.0/'>
  <xap:ModifyDate>2006-10-24T16:47:28-04:00</xap:ModifyDate>
  <xap:CreateDate>2006-10-24T16:47:27-04:00</xap:CreateDate>
  <xap:MetadataDate>2006-10-24T16:47:28-04:00</xap:MetadataDate>
 </rdf:Description>

 <rdf:Description rdf:about='uuid:1d862a03-e87a-414c-a3bc-438844b8b643'
  xmlns:xapMM='http://ns.adobe.com/xap/1.0/mm/'>
  <xapMM:DocumentID>uuid:1aa82404-7080-4651-bfef-1dd39b9b9ed8</xapMM:DocumentID>
 </rdf:Description>

 <rdf:Description rdf:about='uuid:1d862a03-e87a-414c-a3bc-438844b8b643'
  xmlns:dc='http://purl.org/dc/elements/1.1/'>
  <dc:format>application/pdf</dc:format>
  <dc:creator>
   <rdf:Seq>
    <rdf:li>Matthew Beacom</rdf:li>
    <rdf:li>Reed Beaman</rdf:li>
   </rdf:Seq>
  </dc:creator>
  <dc:subject>
   <rdf:Bag>
    <rdf:li>Preservación digital</rdf:li>
    <rdf:li>Archivos digitales</rdf:li>
    <rdf:li>Metadatos</rdf:li>
   </rdf:Bag>
  </dc:subject>
 </rdf:Description>

</rdf:RDF>
</x:xmpmeta>
<?xpacket end='r'?>
```

Fig. 7. Fuente de un archivo XMP

Los metadatos codificados en XMP pueden ser modificados y actualizados en tiempo real durante el curso normal de un flujo de trabajo. En la parte inferior del panel “Metadatos de documento para doc.pdf”, opción “Avanzado”, se encuentran los botones que permiten extraer y guardar el archivo XMP para realizar estas operaciones. (Fig. 8)

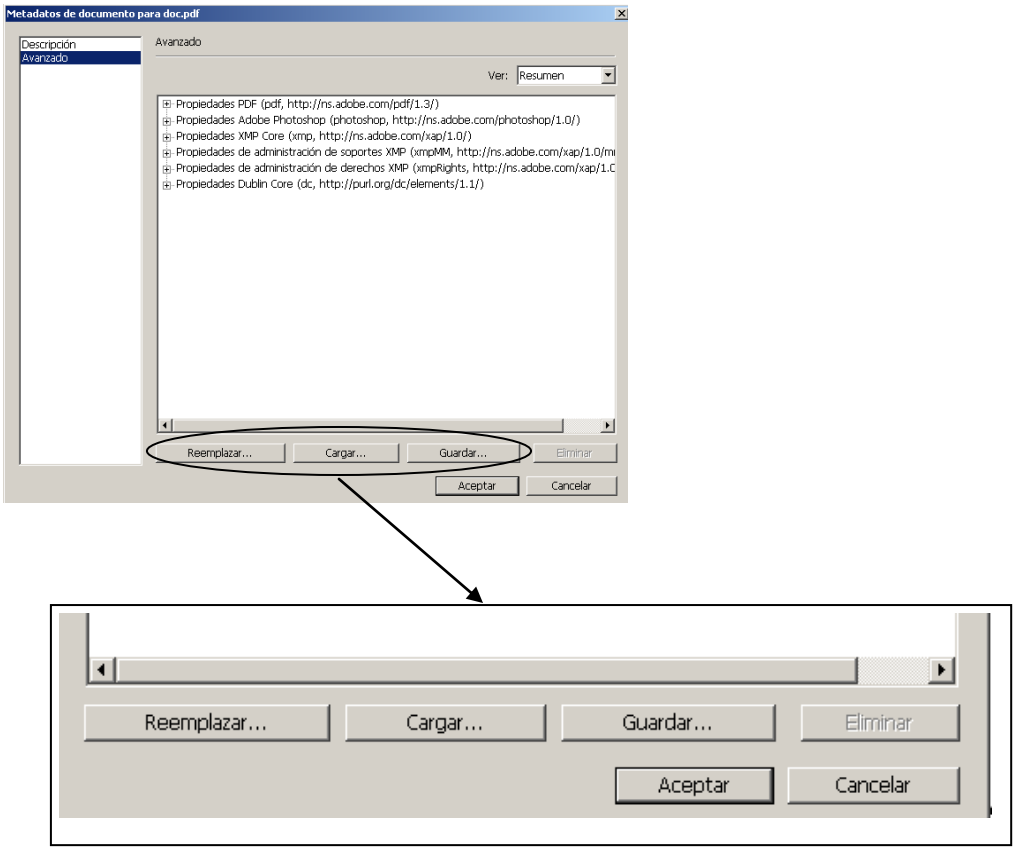

Fig. 8. Botones de edición de archivo XMP

Procesamiento de los metadatos de un archivo XMP en Greenstone

En esta etapa, las pruebas realizadas se enfocaron en lo siguiente: lograr que los metadatos ingresados en la etapa de "Descripción" fueran extraídos en forma automática por el software para biblioteca digital elegido. A tal fin, los documentos en formato PDF, con sus metadatos embebidos utilizando tecnología XMP, fueron procesados con el programa de biblioteca digital Greenstone versión 2.72 [3].

Para las pruebas se configuró el plugin para PDF que utiliza Greenstone[4]. El motivo fue mejorar la performance en la extracción que, por defecto, realiza el software y obtener los metadatos de acuerdo a lo especificado por el esquema DC.

En el archivo de configuración de la colección (colletc.cfg), el plugin para PDFs quedó configurado de la siguiente manera:

```
Plugin PDFPlug -metadata_fields
Title<dc.Title>,Author<dc.Creator>,Subject<dc.Description>,Keywords<dc.Subject>
```

donde los términos en azul son los que Greenstone busca y extrae en el proceso de creación de la colección., y los que se encuentran encerrados entre los signos mayor y menor constituyen el nombre de la etiqueta de metadato que contiene el dato extraído, por ejemplo

<Metadata name="**dc.Creator**"> **Jaspers, Karl**</Metadata>

En este caso, como ya se mencionó, se indicaron etiquetas DC.

[3] http://www.greeenstone.org

[4] Para ello se contó con el aporte del Lic. Diego Spano, del Archivo Nacional de la Memoria.

Con este procedimiento se optimizaron los resultados obtenidos, sin embargo se presentaron algunos problemas por los motivos que se describen a continuación.

Greenstone convierte el documento que se ingresa en un archivo XML desde el cual recupera la información para, por ejemplo, agrupar y configurar los índices de títulos, temas, etc., por medio de los llamados clasificadores (Fig. 9)

Fig. 9. Pantalla de la colección de prueba

En la Fig. 10 se presenta la parte superior de uno de los archivos XML generados por Greenstone y que corresponde a uno de los documentos PDF procesados.

```
<?xml  etadat="1.0" encoding="UTF-8" standalone="no" ¿>
  <!DOCTYPE Archive (View Source for full doctype…)>
- <Archive>
- <Section>
- <Description>
 <Metadata name="gsdldoctype">indexed_doc</Metadata>
 <Metadata name="Language">en</Metadata>
 <Metadata name="Encoding">utf8</Metadata>
 <Metadata name="dc.Creator">Day, Michael; Burns, George; Jaspers, Karl</Metadata>
 <Metadata name="dc.Subject">Preservación digital; Informática documental</Metadata>
 <Metadata name="dc.Description">Ensuring the long-term preservation of information in digital form will be one of the
     greatest challenges</Metadata>
 <Metadata name="Title">Preservation metadata</Metadata>
 <Metadata name="URL">http://C:/Archivos de
     programa/Greenstone/collect/preserva/tmp/preservationmatadataMichaelDay2.html</Metadata>
 <Metadata name="gsdlsourcefilename">import\preservation  etadata- Michael Day2.pdf</Metadata>
 <Metadata name="gsdlconvertedfilename">tmp\preservationmatadataMichaelDay2.html</Metadata>
 <Metadata name="Source">preservation  etadata- Michael Day2.pdf</Metadata>
 <Metadata name="Plugin">PDFPlug</Metadata>
 <Metadata name="FileSize">88158</Metadata>
 <Metadata name="FileFormat">PDF</Metadata>
 <Metadata name="srclink"><a href="_httpprefix_/collect/[collection]/index/assoc/[archivedir]/doc.pdf"></Metadata>
 <Metadata name="srcicon">_iconpdf_</Metadata>
 <Metadata name="/srclink"></a></Metadata>
 <Metadata name="NumPages">14</Metadata>
 <Metadata name="Identifier">HASH01853a29cf2591d884c9aa5a</Metadata>
 <Metadata name="assocfilepath">HASH0185.dir</Metadata>
 <Metadata name="gsdlassocfile">preservationmatadataMichaelDay2-3_1.jpg:image/jpeg:</Metadata>
 <Metadata name="gsdlassocfile">doc.pdf:application/pdf:</Metadata>
   </Description>
 <Content><A name=1></ …(continua el texto del documento)
```

Fig. 10. Archivo XML generado por Greenstone

En el caso de dc.creator debería haber tres ocurrencias y en dc.subject debería haber dos ocurrencias. Sin embargo en ambos casos la extracción que produce el plug in, genera sólo una.

La conclusión a la que se arribó luego de varias pruebas, es que el plugin PDF de Greenstone no extractaba de la manera correcta para nuestras necesidades, la información almacenada en el archivo fuente XMP de un documento PDF, en el caso de metadatos tales como autor y palabras claves.

En vista de lo mencionado, se decidió enviar un mensaje de consulta a la lista de desarrolladores de Greenstone[5]. Dicha consulta fue contestada por el Consultor en Digital Library de Greenstone, Sr. John Thompson, quien ratificó la hipótesis a la que se había arribado: Greenstone no soporta la estructura en que son embebidos los diversos esquemas de metadatos en XMP. A partir de constatar esta situación, dicho consultor elaboró un nuevo plug-in (MetadataXMPPlug)[6] que permite interpretar y extraer los diversos esquemas de metadatos que se pueden encontrar en el archivo XMP de un documento PDF.

Este plugin se debe instalar en la carpeta …\Greenstone\perllib\plugins\, y configurar el archivo collect.cfg de la siguiente manera (se presentan solo las líneas relacionadas con el caso):

```
...
plugin        MetadataXMPPlug
plugin        PDFPlug
...
```

Como se puede observar, ya no se hace necesario definir los campos de metadatos en el plugin PDF.

Se realizaron los cambios en directorios y archivos de Greenstone indicados por el Sr. Thompson, a fin de posibilitar el correcto funcionamiento del nuevo plug in, con el que se volvió a procesar la colección de prueba.

Se pudieron apreciar los cambios en la colección creada, al abrir y examinar los archivos XML generados por Greenstone a partir de la aplicación del MetadataXMPPlug, comprobándose que la estructura de los metadatos extraídos se había realizado de manera correcta y acorde a los datos que residían en los archivos XMP correspondientes a cada documento pdf.

En la siguiente figura se presenta parte de un archivo XML obtenido en Greenstone utilizando el nuevo plug in:

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
  <!DOCTYPE Archive (View Source for full doctype...)>
- <Archive>
- <Section>
- <Description>
  <Metadata name="gsdldoctype">indexed_doc</Metadata>
  <Metadata name="Language">en</Metadata>
  <Metadata name="Encoding">utf8</Metadata>
  <Metadata name="xap.ModifyDate">2007-05-31T10:53:53-03:00</Metadata>
  <Metadata name="dc.Subject">Preservación digital</Metadata>
  <Metadata name="dc.Subject">Archivos digitales</Metadata>
  <Metadata name="dc.Subject">Metadatos</Metadata>
  <Metadata name="dc.Creator">Matthew Beacom</Metadata>
  <Metadata name="dc.Creator">Reed Beaman</Metadata>
  <Metadata name="xap.MetadataDate">2007-05-31T10:53:53-03:00</Metadata>
```

Fig. 11. Archivo XML generado con la utilización del MetadataXMPPlug

Se observa en este caso cómo se genera una etiqueta para cada dc.subject y dc.creator que existe.

Conclusión

El enfoque que se le ha dado en el CICAC a este proyecto de preservación del patrimonio intelectual de la CNEA, centrado en la digitalización de sus Memorias, Boletines e Informes, está basado y tiene un importante componente en la investigación de todos los tópicos relacionados con

[5] greenstone-users@list.scms.waikato.ac.nz
[6] Véase http://www.dlconsulting.com/blog/?p=5 [Consultado: 20 Julio 2007]

cada una de las actividades que forman parte del proceso y del flujo de trabajo, desde la selección del material a digitalizar hasta su disponibilidad en Internet.

Esta investigación, en el caso específico que se comenta en el presente documento, ha conducido a la obtención de un producto de utilidad para toda la comunidad de usuarios de Greenstone, gracias al aporte recibido por parte de un consultor de esta aplicación, quien se abocó a resolver una problemática que se planteo en el proceso investigativo.

En esta línea de trabajo se espera que el CICAC pueda realizar nuevos aportes en el futuro, pues se continuarán desarrollando actividades encaminadas a investigar el comportamiento de otros programas de biblioteca digital en la extracción de metadatos de documentos en formato PDF y XMP.

Bibliografía

Adobe Systems Incorporated. *A manager's introduction to Adobe eXtensible Metadata Platform : the Adobe XML metadata framework".* Disponible en: http://www.adobe.com/products/xmp/pdfs/whitepaper.pdf [Consultado: 12 Jun. 2007].

Adobe Systems Incorporated. *Extensible Metadata Platform (XMP)*. Disponible en: http://www.adobe.com/products/xmp/index.html [Consultado: 12 Jun. 2007].

Adobe Systems Incorporated. *PDF Reference : Adobe portable document format : version 1.6.* Disponible en: http://www.adobe.com/devnet/pdf/pdfs/PDFReference16.pdf [Consultado: 12 Jun. 2007].

Adobe Systems Incorporated. *XMP Specification*. Disponible en: http://www.adobe.com/devnet/xmp/pdfs/xmp_specification.pdf [Consultado: 12 Jun. 2007]

Bainbridge, David ; McKay, Dana; H. Witten, Ian. *Guía del programador : Biblioteca Digital Greenstone*. Disponible en: http://www.greenstone.org/manuals/gsdl2/es/html/Develop_es_index.html [Consultado: 12 Jun. 2007].

Bray, Tim. *What is RDF?*. Disponible en: http://www.xml.com/pub/a/2001/01/24/rdf1.html [Consultado: 12 Jun. 2007].

Myers, Chuck. *Adding intelligence to media : metadata Strategy Adobe XMP – PRIMEX*. Disponible en: http://www.idealliance.org/primex/presentations/04/slides/myers.pdf [Consultado: 12 Jun. 2007]

Roszkiewycz, Ron. *Metadata in context*. Disponible en: http://www.adobe.com/products/xmp/pdfs/seybold_metadata.pdf [Consultado: 12 Jun. 2007]

World Wide Web Consortium. *RDF Primer : W3C Recommendation 10 February 2004*. Disponible en: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ [Consultado: 12 Jun. 2007].

World Wide Web Consortium. *RDF Semantics : W3C Recommendation 10 February 2004*. Disponible en: http://www.w3.org/TR/2004/REC-rdf-mt-20040210/ [Consultado: 12 Jun. 2007].