



LOS METADATOS

en la recuperación de documentos digitales



Por. Vicent Giménez Chornet
Universidad Politécnica de Valencia (España)
vigicho@har.upv.es

1. METADATOS: ¿A QUÉ NOS REFERIMOS?

En los últimos años se ha desarrollado una abundante bibliografía sobre el concepto de metadatos, incidiendo en diferentes aspectos relacionados con las Tecnologías de la Información y la Comunicación. Básicamente podemos encontrar dos ámbitos especiales que tratan del uso de los metadatos, donde se emplean terminologías y metodologías diferentes, aunque todos ellos tienen una misma finalidad primordial: insertar datos e información para recuperar documentos y garantizar su gestión.

Un ámbito trata los metadatos en el entorno de publicaciones web (un recurso más para poder alcanzar la web semántica mediante la indización¹²⁶), y otro ámbito utiliza los metadatos en el entorno de sistemas de gestión de documentos, que no siempre acaban publicándose en Internet. En el primer caso, Eva Méndez denomina *metadatos en sentido estricto*, a aquellos que tratan de la descripción de la información en Internet, y en el segundo caso los denomina del *todo metadatos*, para los resultados de la actividad catalográfica o descripción bibliográfica¹²⁷. Pero también hay otro ámbito, menos desarrollado, que trata de

¹²⁶Moreiro González, José Antonio. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural*. Gijón: Trea, 2004, p. 93-103.

¹²⁷Méndez Rodríguez, Eva. *Metadatos y recuperación de la información: Estándares, problemas y aplicabilidad en bibliotecas digitales*. Gijón: Trea, 2002, p. 40.

la inserción de metadatos en los mismos archivos digitales (JPEG, TIFF, DOC, MP3, etc.), para su utilización tanto en las publicaciones web como en los repositorios documentales de cualquier organización que tenga implantado un sistema de gestión documental. En consecuencia, cuando hablamos de metadatos debemos hacer un esfuerzo en delimitar cual es el entorno de su uso, y con ello clarificar conceptos y metodologías.

En este sentido, una de las cuestiones más conflictivas es concretar la propia definición de metadatos. Se ha utilizado, y se utiliza, como definición de metadatos la expresión “datos sobre datos”¹²⁸. No cabe duda de que esta definición ha tenido un gran éxito y se ha registrado en muchas publicaciones, pero ¿es exactamente así?, ¿pueden haber datos sobre datos?. Otra definición utilizada es completamente opuesta: “información sobre información”¹²⁹. Y en otros casos se utilizan estas definiciones como sinónimos¹³⁰. En general prevalece un principio, los metadatos se utilizan para ser procesados por los ordenadores¹³¹, aunque esta afirmación tampoco la sostienen todos¹³². Seiner, en el 2000, ya puso de relieve la vinculación de los ordenadores en el procesamiento de los metadatos, que los define como “*la información, documentada mediante las herramientas de las Tecnologías de la Información, que mejora el conocimiento empresarial y técnico, de los datos y los datos relacionados con los procesos*”¹³³.

Esta complejidad terminológica respecto al significado de metadatos ha promovido que los autores cuando aborden este tema partan por establecer una definición que delimite su uso en el discurso de la obra. Pero pocos autores concretan las diferencias entre datos, metadatos e información.

128 Gilliland-Swetland, Anne J. *La definición de los metadatos. Introducción a los metadatos, vías de información digital*. New York: Getty Information Institute, 1999, p. 1-9.

129 Bray, Tim: *RDF and Metadata* <<http://www.xml.com/pub/a/98/06/rdf.html>> [Consulta: 4-09-2008]

130 “Los metadatos son información sobre la información (o datos sobre datos) y como tal son, en realidad, una antigua fórmula.”, en Codina, Lluís; Rovira, Cristòfol, *La Web Semántica* <<http://eprints.rclis.org/archive/00008637/>> [Consulta: 4-09-2008]

131 “Metadata is machine understandable information for the web”, *Metadata and Resource Description* <<http://www.w3.org/Metadata/>> [Consulta: 4-09-2008]

132 Anne J. Gilliland-Swetland opina que los metadatos no tienen que ser necesariamente digitales y que los especialistas de patrimonio cultural los han estado creando desde que administran colecciones. Op. cit.

133 Seiner, Robert S. *Questions Metadata Can Answer. The Data Administration Newsletter*. 1 de enero de 2000. <<http://www.tdan.com/view-articles/4841/>> [Consulta: 5-09-2008]

Una definición más reciente intenta establecer los diferentes ámbitos que puede abarcar el concepto de metadatos¹³⁴:

Los metadatos son los datos que describen el contenido, el formato o los atributos de un registro de datos o de un recurso de información. Puede ser utilizado para describir los recursos altamente estructurados o documentos sin estructura en la información como los documentos de texto. Los metadatos pueden aplicarse a la descripción de: recursos electrónicos; información digital (inclusive imágenes digitales), y a la documentación impresa como libros, diarios e informes. Los metadatos pueden estar insertados dentro del recurso de información (como ocurre frecuentemente con los recursos web) o pueden estar separados en una base de datos.

La *Guía de Buenas Prácticas y Comentarios para la Gestión de la Información y la Documentación en la Era Electrónica*¹³⁵, publicada por The Sedona Conference en el 2004, define los metadatos:

Los metadatos (datos acerca de los datos) incluye todo el contexto, procesamiento y uso de información necesaria para identificar y certificar el alcance, la autenticidad y la integridad del activo o de la información electrónica o documentos de archivo. Los metadatos pueden provenir de diversas fuentes. Se pueden crear automáticamente por un ordenador, proporcionados por un usuario, o extraídos a través de una relación en otro documento. Los metadatos son creados, modificados y disponibles en muchos momentos durante la vida de la documentación y la información electrónica.

Algunos metadatos, como el registro de fechas y tamaños, pueden ser fácilmente visualizados por los usuarios; otros metadatos pueden

134 "Metadata is data that describes the content, format or attributes of a data record or information resource. It can be used to describe highly structured resources or unstructured information such as text documents. Metadata can be applied to description of: electronic resources; digital data (including digital images), and to printed documents such as books, journals and reports. Metadata can be embedded within the information resource (as is often the case with web resources) or it can be held separately in a database." Haynes, David, **Metadata for information management and retrieval**. London: Facet Publishing, 2004, p. 8.

135 THE SEDONA GUIDELINES: **Best Practice Guidelines & Commentary for Managing Information & Records in the Electronic Age**. The Sedona Conference, 2004 <<http://www.thesedonaconference.org/content/miscFiles/RetGuide200409.pdf>> [Consulta: 11-09-2008]

estar ocultos o incrustados y no disponibles para usuarios de los ordenadores que no son técnicamente expertos. Los metadatos generalmente no se reproducen en forma completa cuando un documento se imprime.

Diferentes estándares han incorporado también definiciones sobre metadatos. La norma UNE-ISO 23081-1:2006 *Metadatos para la gestión de documentos. Parte 1: Principios*¹³⁶:

“En el contexto de la gestión de documentos, los metadatos se definen como datos que describen el contexto, contenido y estructura de los documentos, así como su gestión a lo largo del tiempo (ISO 15489-1:2001, 3.12). Como tales, los metadatos son información estructurada o semiestructurada que posibilita la creación, registro, clasificación, acceso, conservación y disposición de los documentos a lo largo del tiempo y dentro de un mismo dominio [*competencia] o entre dominios diferentes. Cada uno de estos dominios, representa un área del discurso intelectual y de la actividad social o de la organización desarrollado por un grupo propio o limitado de individuos que comparten ciertos valores y conocimiento. Los metadatos para la gestión de documentos pueden usarse para identificar, autenticar y contextualizar tanto los documentos como los agentes, procesos y sistemas que los crean, gestionan, mantienen y utilizan, así como las políticas que los rigen”.

Según esta norma los metadatos son tanto *datos que describen* como *información estructurada*, generando tal vez una confusión terminológica entre lo que se entiende por dato y lo que se entiende por información. Según la *Guía de la información electrónica. Cómo tratar los datos legibles por máquina y la documentación electrónica*¹³⁷:

¹³⁶UNE-ISO 23081-1:2006 *Metadatos para la gestión de documentos. Parte 1: Principios*. Madrid: AENOR, 2008, p. 6.

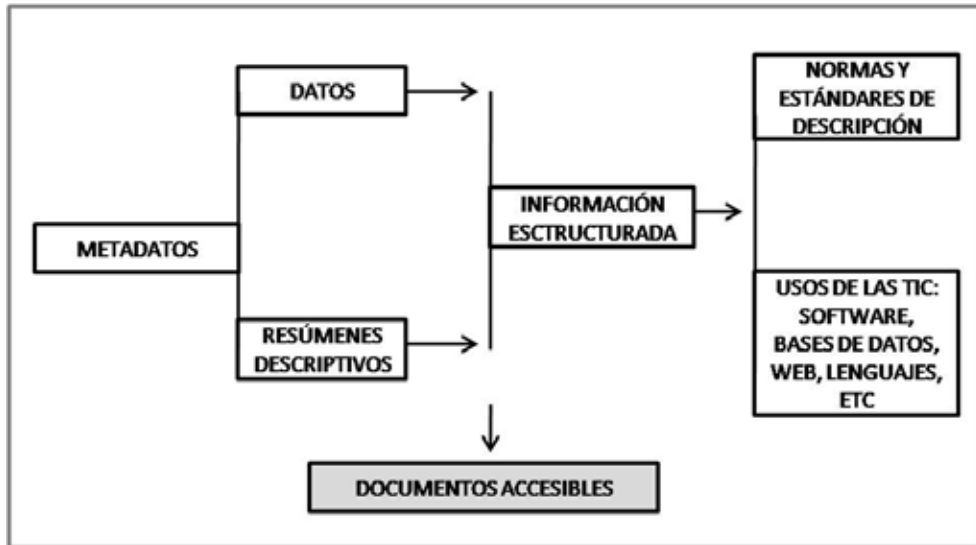
¹³⁷*Guía de la información electrónica. Cómo tratar los datos legibles por máquina y la documentación electrónica*. DLM-FORUM. Luxemburgo: Oficina de las publicaciones oficiales de las Comunidades Europeas, 1997. Traducción a partir de las ediciones originales en lenguas francesa e inglesa. Por José María Fernández Hevia, 2000. Documento accesible en línea: <http://www.cornu.eu.org/files/guidelines_ES.pdf> [Consulta: 1-09-2008]

INFORMACIÓN	DATO
<p>Una información es una indicación o un evento llevado al conocimiento de una persona o de un grupo. Es posible crearla, mantenerla, conservarla y transmitirla. La información está en la base de la organización de todo proceso de trabajo. Este concepto se ha vuelto tan importante, que la sociedad de la información es considerada en el día de hoy como la etapa siguiente a la de la "sociedad industrial".</p>	<p>Un dato es una unidad elemental de información.</p> <p>En un documento, por ejemplo, se agrupan numerosos datos para presentar una argumentación o rendir cuentas de una acción. Hasta hace muy poco, la mayor parte de los datos eran conservados y transmitidos mediante documentos en soporte papel (o a veces, para los más antiguos, en otros soportes, como la piedra tallada).</p> <p>A veces encontramos datos en forma de listas en soporte papel, como ocurre con las guías de teléfonos. Ya no se trata de presentar una argumentación, sino de suministrar materia prima para una acción futura (localizar el número de teléfono del señor García, por ejemplo). En estos casos, es crucial la clasificación de los datos para una fácil localización de la información requerida (por ejemplo, en una guía de teléfono los apellidos se clasifican por núcleos de población y después por orden alfabético).</p>

Distintas normas de descripción normalizada adoptan campos que pueden contener datos (por ejemplo fechas, nombres de personas o entidades, palabras claves, etc.) y campos que contienen resúmenes descriptivos (los podemos considerar como pequeños documentos) fruto de una combinación de datos para presentar una argumentación descriptiva del documento, su contexto, sus requisitos de conservación u otros aspectos, para hacerlos accesibles en cualquier momento e incluso desde cualquier lugar.

En definitiva los datos son una cadena de símbolos elementales (dígitos o letras) que disponen de un valor¹³⁸.

¹³⁸ "A datum is a string of elementary symbols, such as digits or letters. It is the value of an attribute. It need not have meaning to everyone, but it must be clear of what attribute a datum is a value." Charles T. Meadow, Bert R. Boyce, Donald H. Kraft, Carol L Barry, *Text Information Retrieval Systems*. London: Emerald Group Publishing, 2007, p. 38.



Compartimos la definición de la UNE-ISO 23081-1:2006 sobre que los metadatos se definen como datos que describen el contexto, contenido y estructura de los documentos, así como su gestión a lo largo del tiempo. Sin embargo, en relación con esta definición, hay que precisar que el concepto de datos, como unidad elemental de información, es insuficiente, dado que en algunos casos la información registrada en los campos controlados por los metadatos puede contener breves resúmenes descriptivos.

Por otra parte, también compartimos que los metadatos son información estructurada o semiestructurada que posibilita la creación, registro, clasificación, acceso, conservación y disposición de los documentos a lo largo del tiempo, pero no compartimos que esta información esté relacionada con un dominio o competencia, en el sentido de un área del discurso intelectual y de la actividad social o de la organización, ya que consideramos que los metadatos se usan bajo dos principios que pueden conjugar más de un "dominio":

- a) describir uno o más ámbitos de un sistema de gestión de la información (el contexto, el contenido, los criterios de preservación y conservación, etc.)
- b) aplicar una metodología basada en la conjunción de criterios de descripción, generalmente estándares como formato MARC, ISAD (G), RDF, Dublin Core, etc., y de aplicaciones informáticas (bases de datos

como Oracle <<http://www.oracle.com>>, Basis <<http://www.basis.com/>>, etc., editores web, editores de imagen, etc.)

Así, por ejemplo, si queremos describir el contenido, principio a), utilizando el estándar RDF con un editor web, principio b), podemos incorporar metadatos a páginas web cuyo contenido puede ser de cualquier "*área del discurso intelectual y de la actividad social o de la organización...*" cuya información se registra en un medio, la web.

2. CLASIFICACIÓN Y FUNCIONES DE LOS METADATOS

Hace ya una década Anne J. Gilliland-Swetland clasificó los diferentes tipos de metadatos en relación a sus funciones:

- administrativas, relacionadas con el uso de metadatos para la gestión y administración de los recursos de información (adquisición de la información, derechos de reproducción, documentación sobre requerimientos legales de acceso, localización de la información, criterios de digitalización y pistas de auditoría),
- descriptivas, en las que se identifica el recurso de información (catalogación, ayudas de búsqueda, índices especializados, hiperenlaces entre recursos, anotaciones de los usuarios),
- preservación, sobre metadatos relacionados en la conservación de recursos de información (documentación sobre los requisitos físicos para la gestión de la conservación de los recursos, documentación sobre la toma de acciones para la conservación física y versiones digitales de los recursos, como la migración o actualización de datos),
- técnicos, sobre los metadatos relacionados con el funcionamiento del sistema o el comportamiento de los metadatos (documentación sobre el hardware y el software; información sobre la digitalización como formatos, tipo de comprensión, ajustes; seguimiento del tiempo de respuesta del sistema; autenticación y seguridad de los datos, como claves cifradas, contraseñas, etc.),
- uso, sobre metadatos relacionados con el nivel y clase de uso de los recursos de información (registro de exposiciones, seguimiento de usos y usuarios, reutilización del contenido e información sobre múltiples versiones).

Para los archivos digitales generados como consecuencia de la actividad de las organizaciones o personas físicas proponemos otra clasificación de las funciones de los metadatos. Partimos de una premisa: los metadatos contienen información y datos, desde diferentes categorías, mientras que la recuperación de la información se basa en el sistema y arquitectura que se implementa para recuperar esa información o recuperara los datos introducidos en los campos de los metadatos, toda ello relacionada con la gestión documental. Sin metadatos, sin cumplimentar los campos de los metadatos, no puede haber recuperación de la información. El sistema de recuperación de la información (SRI) es básicamente un modelo, más o menos eficaz, que tiene como punto de partida la forma de interrogar la información contenida en los campos de metadatos y que mediante un proceso se podrá recuperar la información más pertinente, preferiblemente estructurada. En este sentido podemos clasificar los metadatos en relación a las siguientes funciones:

Clasificación de los metadatos para los documentos digitales de archivo integrados en un Sistema de Gestión Documental		
Categoría	Definición	Ejemplos
Descripción del contenido	Metadatos de identificación, contenido y indización del documento. Análisis documental.	<ul style="list-style-type: none"> • Catalogación • Indización, tesauros. • Notas
Descripción del Contexto	Metadatos sobre el órgano productor	<ul style="list-style-type: none"> • Catálogo de autoridades • Descripción de funciones institucionales • Cuadros de clasificación y relaciones jerárquicas
Acceso	Metadatos para la localización y el uso documental	<ul style="list-style-type: none"> • Localización de la información • Criterios de acceso: límites legales a la información • Derechos de autor • Derechos de reproducción

Preservación	Metadatos para garantizar la conservación de documentos físicos y electrónicos e información del medio	<ul style="list-style-type: none"> • Requisitos para la conservación física y electrónica de los documentos • Criterios de migración • Información sobre instalaciones • Información sobre hardware y software
Descripción física	Metadatos para describir las propiedades y características de la documentación	<ul style="list-style-type: none"> • Descripción física de los documentos • Información sobre las propiedades físicas de los documentos digitales (formatos, comprensión, etc.) • Información de los parámetros de la captura digital • Información sobre restauración: cambios en el documento digital o en el documento original digitalizado
Autenticidad	Metadatos probatorios y garantes de los creadores documentales	<ul style="list-style-type: none"> • Firmas digitales • Claves seguras • Encriptación
Interoperabilidad	Metadatos para el intercambio de datos e información entre sistemas	<ul style="list-style-type: none"> • Identificación de documentos/registros • Identificación de software • Identificación de usuarios y permisos
Trazabilidad / Pista de Auditoría	Metadatos sobre la información del movimiento y uso de los documentos	<ul style="list-style-type: none"> • Registro de consultas: quien, que, donde y cuando. • Seguimiento de usuarios • Seguimiento de modificaciones, eliminaciones o incorporaciones. • Cambios de custodia y reutilización.

Valoración documental	Metadatos sobre la valoración documental y el expurgo	<ul style="list-style-type: none"> • Tablas de valoración documental • Informes de comisiones de valoración • Datos de eliminación documental
-----------------------	---	--

La recuperación de la información en los documentos electrónicos de archivo está estrechamente relacionada (en la categoría de metadatos a la que se va a interrogar) con los campos que se hayan implementado y cumplimentado.

3. METADATOS PARA LOS ARCHIVOS EN LA ERA DIGITAL

Actualmente se pueden clasificar en dos grandes grupos los archivos que disponen de documentos digitales: aquellos archivos que nacen digitalmente y los que disponen de documentos digitales como consecuencia de la digitalización de los documentos en papel o documentos en otros soportes. En cualquiera de estos dos casos hay que tener presente que cualquier aspecto de las categorías de metadatos les incumbe y que deben ser registrados en el sistema de gestión documental. En estos momentos es muy habitual que los archivos nacidos digitalmente convivan con los archivos generados en soportes analógicos. En estos casos, los archivos que implementan un proceso de digitalización tienen como objetivo poder disponer de un sistema único para el proceso de acceso y recuperación de la información de toda la documentación generada por la organización.

Hay que tener presente tres aspectos que van a ser condicionantes a la hora de la recuperación de la información en los archivos digitales.

- A) El lenguaje utilizado: al margen de las categorías de los metadatos, existen dos tipologías de lenguaje para la descripción del contenido en cualquiera de los innumerables campos, por una parte el lenguaje documental o y por otra el lenguaje natural. Cada uno de estos lenguajes presenta una serie de dificultades y unos aspectos ventajosos. Lo resumimos en el siguiente cuadro:

Tipo de lenguaje	Dificultades	Ventajas
Lenguaje Documental	Trabajo previo de elaboración que puede requerir diversos periodos de tiempo, como las Tablas de Valoración Documental, o los Tesoros para la indización.	Rápida introducción de la información, generalmente realizando un clic con un enlace contra tablas.
Lenguaje Natural	Redacción del contenido, saber discernir entre lo elemental y lo secundario, saber utilizar el vocabulario idóneo.	No existe trabajo previo. Eficaz en la descripción del contenido.

B) Formato e interoperabilidad: en un sistema de gestión de documentos electrónicos el uso de diferentes formatos de archivo electrónico está dificultando, si no imposibilitando, la interoperabilidad o el intercambio de información y datos dentro del mismo sistema, haciendo inviable la recuperación de la información. Actualmente en España ya existen archivos de carácter tradicional (sus fondos eran 100% en papel) que han recibido expedientes generados por el organismo de forma electrónica¹³⁹.

C) Los metadatos pueden estar insertados en el mismo documento digital (contenedores multimedia como AVI, archivos de audio como MP3 y algunos formatos de archivo de imagen como JPEG), o pueden ser externos al documento digital, contenidos en una base de datos donde se designa un campo para indicar la ubicación del documento digital.

En cualquier diseño de una arquitectura que pretenda implantar un sistema de recuperación de la información, independientemente de los elementos y procesos del sistema, el éxito de la recuperación de la información estará intrínsecamente vinculado al lenguaje utilizado, a la interoperabilidad de los formatos digitales y a la ubicación de los metadatos.

¹³⁹Es el caso de la Sindicatura de Comptes de la Generalitat Valenciana. El organismo utiliza un software propietario de TeamMate, y aplica ficheros ACL. Salmon, Alejandro y Rausell, Silvino: "Principales aplicaciones de ACL en la Sindicatura de Comptes de la Comunitat Valenciana", *II FORO TECNOLÓGICO DE LOS OCEX*. Pamplona 2008. En línea: <[http://www.sindicom.gva.es/web/wdweb.nsf/documento/ftpamplona/\\$file/SCCV_Foro2008_ACL.pdf](http://www.sindicom.gva.es/web/wdweb.nsf/documento/ftpamplona/$file/SCCV_Foro2008_ACL.pdf)> [Consulta: 06-02-2009].

3.1 Metadatos insertados en los archivos digitales

Existen diferentes estándares para los formatos de archivos digitales y, independientemente de su configuración, opciones diferentes para la inserción de metadatos relacionados con la descripción del contenido y del contexto.

Se pueden diferenciar cuatro grandes grupos de archivos digitales: los archivos de imagen, los archivos sonoros, los archivos audiovisuales y los archivos hipertexto o web. Cualquiera de los cuatro tipos de archivos los puede generar cualquier organismo en el ejercicio de su actividad.

3.1.1. Archivos de imagen.

El tipo de archivo digital que más abunda en la administración es el archivo de imagen. Es, al fin y al cabo, el traslado del tradicional soporte del papel al soporte digital en la administración electrónica, y es la opción más adoptada en los archivos históricos cuando implantan un proyecto de digitalización.

Actualmente existen tres formatos de archivo de imagen ampliamente adoptados, todos ellos garantizados como estándares: el JPEG¹⁴⁰, el TIFF¹⁴¹ y el PDF¹⁴².

Los archivos JPEG y TIFF permiten incorporar tags o etiquetas. Estos metadatos pueden estructurarse mediante los estándares EXIF, IPTC, Dublin Core o XMP¹⁴³.

EXIF (Exchangeable image file format) ha sido creado por la Japan Electronic

140 ISO/IEC 10918-4:1999 Information technology -- Digital compression and coding of continuous-tone still images: Registration of JPEG profiles, SPIFF profiles, SPIFF tags, SPIFF colour spaces, APPn markers, SPIFF compression types and Registration Authorities

141 ISO 12234-2:2001 Electronic still-picture imaging -- Removable memory -- Part 2: TIFF/EP image data format

142 ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)

143 Giménez Chornet, Vicent; Sellés Carot, Alicia; Roque Izquierdo, Graciela; Puchades Asensi, Yolanda; Monleón Escribano, Daniel: "Recuperación de información descriptiva en imágenes digitales mediante metadatos EXIF. Su utilidad para el archivo del Reino de Valencia"; E-información : integración y rentabilidad en un entorno digital : 10ª Jornadas Españolas de Documentación, Santiago de Compostela. Madrid: FESABID, 2007, p. 25-32.

Industry Development Association (JEIDA) como estándar para facilitar el intercambio de datos de los archivos digitales de imagen, principalmente datos de carácter técnico, con la finalidad de que sean formatos interoperables entre cámaras digitales y aplicaciones¹⁴⁴. Las etiquetas EXIF las podemos encontrar en los formatos de imagen JPG, TIFF, PNG, MIFF y HDP, incluso en formatos TIFF basados en imágenes RAW, y además, en los formatos de videos AVI y MOV. La mayor parte de las etiquetas son de carácter técnico que deben quedar registradas en el momento de la captura¹⁴⁵. Entre todas las etiquetas, disponemos de unas cuantas para la descripción documental:

Etiquetas EXIF para la descripción	
Etiqueta	Nombre de la etiqueta
0x9c9b	XPTitle
0x9c9c	XPComment
0x9c9d	XPAuthor
0x9c9e	XPKeywords
0x9c9f	XPSubject
0x010e	ImageDescription
0x8298	Copyright

A pesar de la estandarización de las etiquetas, el software de edición de los archivos digitales determina la visualización de las mismas. Unas veces el software no recoge todas las posibilidades de las etiquetas y otras veces opta por un nombre para designar lo que puede ser la etiqueta de Título o la de Descripción de la Imagen, de forma que la visualización de las mismas etiquetas por otro software puede provocar que no se visualicen o se visualicen bajo otro campo. Algunos ejemplos los podemos comprobar con

¹⁴⁴ JEIDA se fusionó con EIAJ, y ahora forma el grupo JEITIA (Japan Electronics and Information Technology Industries Association), más información sobre el estándar EXIF en la página de JEITA: <http://www.jeita.or.jp/cgi-bin/standard_e/list.cgi?cateid=1&subcateid=4> [Consulta: 6-02-2009], y en la página no oficial: <<http://www.exif.org/>> [Consulta: 6-02-2009], donde se pueden encontrar las especificaciones EXIF.

¹⁴⁵ EXIF Tags, <<http://www.sno.phy.queensu.ca/~phil/exiftool/TagNames/EXIF.html>> [Consulta: 6-02-2009]

aplicaciones muy populares. En el Archivo del Reino de Valencia¹⁴⁶ se utilizó la aplicación ACDSee¹⁴⁷, una de las pioneras en visualizar los metadatos EXIF, para insertar metadatos de la ISAD (G) en los campos que habíamos comprobado más interoperables de la aplicación:

Correlación de campos ISAD (G) en etiquetas EXIF	
EXIF (ACDSee)	ISAD (G)
Copyright	Código de referencia
Descripción de la imagen	Fechas; Alcance y contenido; Descriptores: geográfico, onomástico, materia, e instituciones y entidades:

Visualización de metadatos EXIF con ACDSee



¹⁴⁶Hemos explicado la experiencia en: "Recovery of descriptive information in images from digital libraries by means of EXIF metadata", *Library Hi Tech*, vol. 26, nº 2, 2008, pp. 302-315.

¹⁴⁷Se utilizó la versión ACDSee 8, <<http://www.acdsee.com/>> [Consulta: 6-02-2009]

Visualización de metadatos EXIF con XnView



Visualización de metadatos EXIF con Photoshop CS3



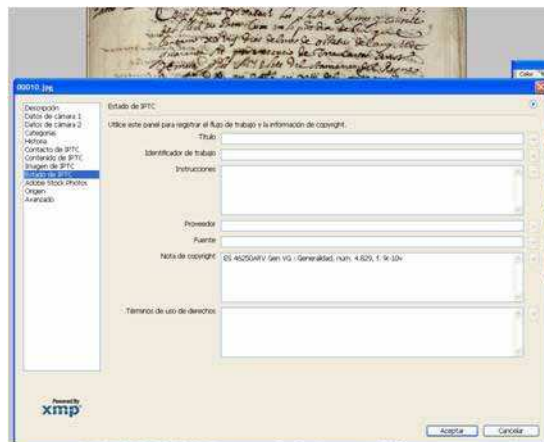
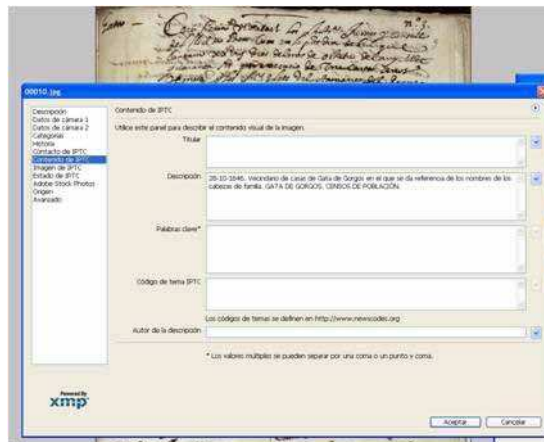
A pesar de la eficacia de la permanencia de los metadatos insertados en el archivo digital, evidenciamos que la visualización de los mismos depende directamente de las posibilidades que ofrece el software.

IPTC, iniciales que originalmente corresponden a un organismo, el International Press Telecommunications Council , nacido en 1965 como grupo de entidades de agencias de noticias, interesado especialmente en el desarrollo y publicación de normas para el intercambio de datos en el ámbito periodístico. En el 2004

iniciaron un estándar de metadatos para los archivos digitales de imagen. En la actualidad esta disponible una especificación en revisión del 2008 . Hoy en día hay un buen número de empresas de software que ha incorporado los metadatos IPTC a sus aplicaciones.

Los campos IPTC destinados a la descripción son más numerosos que los EXIF y posibilitan la lectura de la información introducida en los metadatos EXIF.

Visualización de metadatos IPTC con *PhotoShop CS2*



El objetivo de los metadatos IPTC para los archivos de imagen es describir las fotografías con la finalidad de poderlas recuperar en las organizaciones dedicadas a la actividad periodística. Entre las ventajas de este grupo de metadatos está la creación de un campo para palabras clave, además de los campos que describen el contexto, es decir, el productor o creador del documento: creador, cargo del creador, dirección, ciudad, provincia, código postal, país, etc.

XMP (Extensible Metadata Platform) y **PDF**: la especificación XMP¹⁴⁸ nace con la finalidad de que diferentes aplicaciones puedan trabajar eficazmente con los metadatos, estandarizando la definición, la creación y el procesamiento de los metadatos. El formato PDF/A utiliza la especificación XMP.

El formato PDF/A es el tipo de archivo electrónico recomendado para la conservación a largo plazo¹⁴⁹.

Tabla cruzada entre los campos del PDF y del XMP ¹⁵⁰		
PDF		XMP
<i>Entrada</i>	<i>Características</i>	<i>Entrada</i>
Título (Title)	Cadena de texto	dc:title
Autor (Author)	Cadena de texto	dc:creador
Tema (subject)	Cadena de texto	dc:description ["x-default"]
Palabras clave (Keywords)	Cadena de texto	df:keywords
Creador (Cretaoor)	Cadena de texto	xmp:CreatorTool
Productor (Producer)	Cadena de texto	pdf:Producer
Fecha de creación (CreationDate)	Fecha (date)	xmp:CreateDate
Fecha de modificación (ModDate)	Fecha (Date)	xmp:ModifyDate

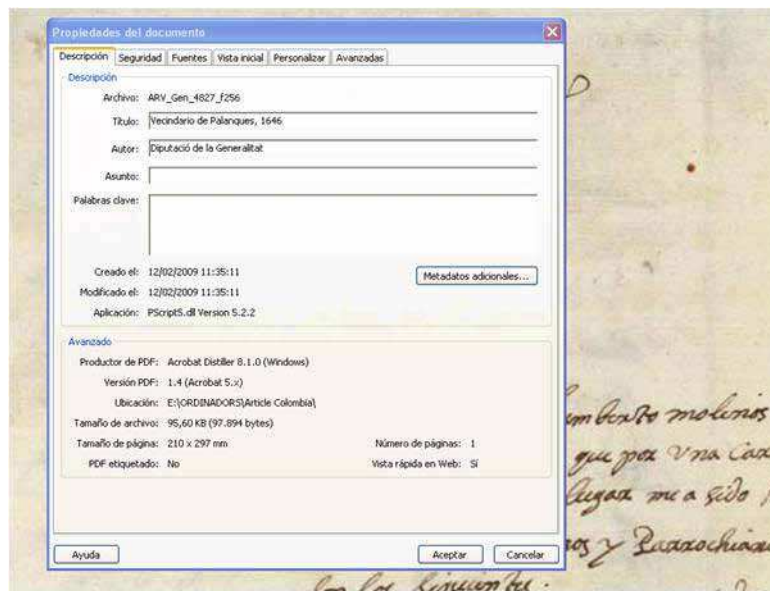
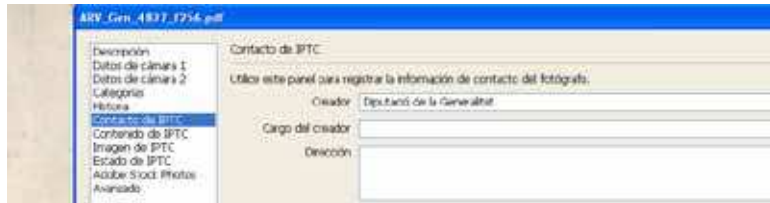
En la actualidad la aplicación para crear pdf, el Adobe Acrobat 8, permite incorporar los metadatos XMP bajo la estructura de los metadatos IPTC.

¹⁴⁸XMP Specification, Adobe, enero 2004, <<http://www.aiim.org//documents/standards/PDF-A/xmpecification.pdf>> [Consulta: 10-02-2009]

¹⁴⁹UNE-ISO 19005-1 Gestión de documentos. Formato de fichero de documento electrónico para la conservación a largo plazo. Parte 1: Uso del PDF 1.4 (PDF/A-1). Madrid: AENOR, 2008

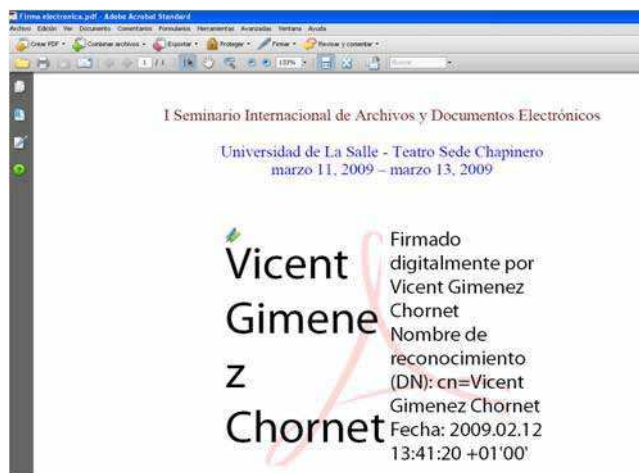
¹⁵⁰Fuente: UNE-ISO 19005-1, p. 19.

Visualización de metadatos en documento PDF, con estructura IPTC



En el entorno de la administración electrónica el formato pdf permite la inserción de metadatos relacionados con la firma digital, dando las máximas garantías de autenticidad e integridad en el documento electrónico.

Visualización de un documento PDF con firma digital



Dublin Core¹⁵¹. Los metadatos Dublin Core parten de una iniciativa de 1995 con el objeto de establecer un conjunto limitado de descriptores para conseguir un estándar para la descripción de recursos de información en distintos dominios informativos. Actualmente es reconocida como estándar por la ISO¹⁵², y como norma NISO¹⁵³. Ha sido un estándar bastante difundido para metadatos en la web¹⁵⁴, pero que también se ha adoptado como estructura de metadatos para ser insertados en archivos de imagen mediante la utilización de algunas aplicaciones, como PhotoShop.

151 Dublin Core Metadata Initiative, <<http://dublincore.org/>> [Consulta: 13-02-2009]

152 UNE-ISO 15836:2007 Información y Documentación. Conjunto de elementos de metadatos Dublin Core. Madrid: AENOR, 2007

153 ANSI NISO Z39.85-2007 The Dublin Core Metadata Element Set. <http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=9b7bffcd2daeca6198b4ee5a848f9beec2f600e5> [Consulta: 13-02-2009]

154 Encoding Dublin Core Metadata in HTML (1999), <<http://www.ietf.org/rfc/rfc2731.txt>> [Consulta: 13-02-2009]

Elementos Dublin Core incorporados por PhotoShop CS2	
Dublin Core	Adobe PhotoShop CS2
Título	
Creador	SI
Materia y palabras clave	
Descripción	SI
Editor	
Colaborador	
Fecha	
Tipo de recurso	
Formato	SI
Identificador de recurso	
Fuente	
Idioma	
Relación	
Cobertura	
Derechos	SI

El inconveniente de los metadatos Dublin Core para los archivos de imagen es el escaso número de aplicaciones informáticas que los ha incorporado y que, por tanto, facultan su uso.

3.1.2. Archivos de audio.

Por lo que respecta a los archivos de audio se ha creado un estándar de facto para la inserción de metadatos, el ID3¹⁵⁵. Actualmente soporta el formato mp3 y existen un buen número de aplicaciones que permiten la edición de los metadatos ID3¹⁵⁶. Existen 39 campos para poder describir el archivo de audio, especialmente etiquetado para los CD de audio de autores musicales: título del álbum, BMP (beats por minuto), compositor, tipo de contenido, copyright, título, subtítulo, fecha de grabación, etc.¹⁵⁷, además de poder incluir 9 campos destinados a un enlace URL,

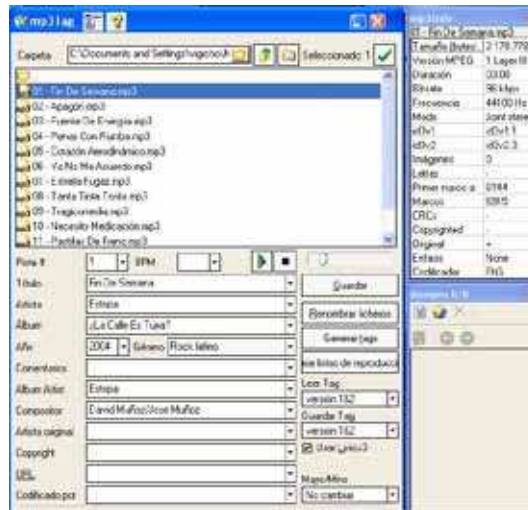
¹⁵⁵ ID3 The Audience is informed, <<http://www.id3.org/>> [Consulta: 16-02-2009]

¹⁵⁶ Una lista del software que incorpora el estándar ID3 lo encontramos en ID3, Implementations, <<http://www.id3.org/Implementations>> [Consulta: 16-02-2009]

¹⁵⁷ ID3, Frames, <<http://www.id3.org/Frames>> [Consulta: 16-02-2009]

para dar información sobre la web oficial del autor, información comercial, web oficial del editor, aviso legal de los derechos de autor, etc.

Metadatos ID3, Visualización con mp3Tag



3.1.3. Archivo audiovisual.

Existen diversos estándares para la inserción de metadatos en contenedores de video (video, audio, y otros datos), tanto de carácter técnico para hacerlos interoperables, como el MXF (Material eXchange Format)¹⁵⁸, perteneciente al conjunto de estándares de SMPTE¹⁵⁹, en el entorno de la televisión, como de carácter descriptivo, el MPEG-7.

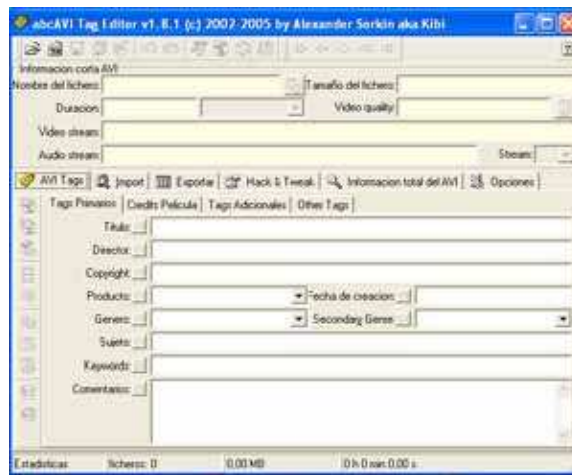
¹⁵⁸Material Exchange Format : Basic User Metadata Implementation, Recommendation R 121, European Broadcasting Union, 2007, <http://www.ebu.ch/CMSimages/en/tec_text_r121-2007_tcm6-50026.pdf> [Consulta:17-02-2009]

¹⁵⁹Society of Motion Picture and Television Engineers, <<http://www.smpete.org/home>> [Consulta: 17-02-2009]. Puede consultarse también Proposed SMPTE Engineering Guideline for Televisión, Material Exchange Format (MXF), MXF Descriptive Metadata, EG 42, <<http://rhea.tele.ucl.ac.be:8081/Plone/Members/egoray/thesaurus-dictionnaire-metadata/eg42-mxf-smpete.pdf>>, [Consulta:17-02-2009]

Uno de los formatos con más ambición para la descripción de contenidos en audiovisual es el MPEG-7, que permite la combinación de la descripción semántica del contenido como la descripción técnica. Actualmente el formato MPEG7 dispone de la norma ISO/IEC 15938 *Multimedia content description interface*¹⁶⁰, pero dispone de escasas aplicaciones informáticas para explotar el potencial de metadatos descriptivos.

Con menos ambiciones, pero con gran eficacia, algunas aplicaciones permiten insertar metadatos descriptivos en los contenedores AVI, como los campos de título, director, copyright, género, palabras clave, etc.

Metadatos en AVI con *abcAVI Tag Editor*



3.2 Metadatos externos a los archivos digitales.

Los metadatos externos a los archivos digitales están estrechamente relacionados con las bases de datos. No consideremos metadatos la información o datos que se registran en fichas en papel o en libros, aunque se hayan utilizado estándares en la catalogación o descripción de documentos.

El grupo de metadatos utilizados para la recuperación de la información va a depender del ámbito documental en que se desarrolle la actividad profesional.

¹⁶⁰ Puede consultarse el estándar en: <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>, [Consulta:17-02-2009]

Cada actividad ha generado uno o más estándares para normalizar la descripción documental, con la finalidad de hacer interoperable el sistema de gestión documental y recuperable la información y los documentos registrados en ha facilitado enormemente no sólo la recuperación de la información sino la interoperabilidad y la recuperación entre distintas bases de datos, sin importar su lejanía física. Ello ha sido posible gracias a la utilización de dos estándares, la descripción mediante el formato MARC¹⁶¹ y la utilización del protocolo Z39.50¹⁶². El formato MARC ha permitido estructurar los metadatos descriptivos de los documentos impresos y el protocolo Z39.50 ha permitido el intercambio de la información. Ahora bien, la recuperación de la información va a depender de cómo las aplicaciones informáticas busquen, capturen y visualicen los metadatos registrados en las bases de datos. Un ejemplo de robot que mediante el protocolo Z39.50 realiza búsquedas bibliográficas en más de 2000 bibliotecas es *BookWhere*¹⁶³.

Búsquedas bibliográficas con *BookWhere* (Protocolo Z39.50 y formato MARC)

dicho sistema.

En el ámbito de las bibliotecas (una de las actividades profesionales que inició más prematuramente la normalización), las normas en descripción bibliográfica ISBD¹⁶⁴ y el formato MARC disponen de los campos de metadatos que más se han implementado. Ello

¹⁶¹ Normas MARC, <<http://www.loc.gov/marc/marcspa.html>> [Consulta: 17-02-2009]

¹⁶² ANSI/NISO Z39.50, <<http://www.cni.org/pub/NISO/docs/Z39.50-1992/>> [Consulta: 17-02-2009]

¹⁶³ *BookWhere Academic*, <<http://www.webclarity.info/support/download/ba.html>> [Consulta: 17-02-2009]

¹⁶⁴ El grupo de normas ISBD, todas ellas destinadas a los diferentes materiales que se pueden localizar en las bibliotecas, se pueden consultar en *Family of superseded ISBDs*, <<http://www.ifla.org/VI/3/nd1/isbdlist.htm>> [Consulta: 17-02-2009]

Author	Title	Date	Database	Rating	Notes
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	Biblioteca Nacional de Portugal	58	
Doménech Chornet, Vicent	Las orden de enclaustrados de los reinos de España	1923	Biblioteca Pública de Navarra	2	
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	New York Public Library	49	
Doménech Chornet, Vicent	Compendio de historia	1775	Memoria de la Biblioteca	57	
Doménech Chornet, Vicent	El inicio de los ferrocarriles y tranvías de vía estrecha	1999	New York Public Library	54	
Doménech Chornet, Vicent	En la historia del asfalto	2002	Campana Public Library	67	
Doménech Chornet, Vicent	El inicio de los ferrocarriles y tranvías de vía estrecha	1999	Universitat de Calicut, MELVLS	64	
Doménech Chornet, Vicent	Compendio de historia	2002	Universitat de Calicut, MELVLS	49	
Doménech Chornet, Vicent	Aguja del començament de l'encastellament de Cal	2000	Biblioteca de la Ciudad de León	68	
Doménech Chornet, Vicent	El inicio de los ferrocarriles y tranvías de vía estrecha	1999	University of Michigan, Ann Arbor	69	
Doménech Chornet, Vicent	El inicio de los ferrocarriles y tranvías de vía estrecha	1999	University of Toronto	64	
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	University of Toronto	69	
Andrés Bello, Andrés	Aguja del començament de l'encastellament de Cal	2000	Biblioteca Nacional Española, M.	64	
Doménech Chornet, Vicent	Compendio de historia	2002	Biblioteca Nacional Española, M.	58	
Doménech Chornet, Vicent	Compendio de historia	1884	Biblioteca Nacional Española, M.	69	
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	Library of Congress	69	
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	Library of Congress	64	
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	New University	69	
Doménech Chornet, Vicent	Desenvolupament capitalista en el sistema feudal	1994	Biblioteca Nacional Española, M.	49	
Doménech Chornet, Vicent	El inicio de los ferrocarriles y tranvías de vía estrecha	1999	Biblioteca Nacional Española, M.	69	

Más recientemente en la Library of Congress se trabaja en el proyecto de un esquema de metadatos para la descripción de autoridades relacionado con el formato MARC, el MADS (Metadata Authority Description Schema)¹⁶⁵. También son metadatos los puntos de acceso, normalizados generalmente mediante la utilización de lenguajes documentales como la *Lista de Encabezamientos de materia para las Bibliotecas Públicas*¹⁶⁶, o las clasificaciones bibliográficas como la CDU¹⁶⁷.

En el ámbito de los archivos, la introducción de los estándares experimenta un notable retraso, lo que dificulta enormemente la recuperación de la información. Actualmente existen dos destacados estándares para la descripción archivística: la EAD y la ISAD (G). La EAD¹⁶⁸ (Encoded Archival Description) es una DTD (Document Type Definition) del lenguaje de marcas XML, impulsada por un grupo de trabajo de la Sociedad de Archiveros Americanos y la Biblioteca del Congreso

165 Metadata Authority Description Schema, <<http://www.loc.gov/standards/mads/>> [Consulta: 17-02-2009]

166 Lista de Encabezamientos de materia para las Bibliotecas Públicas, <<http://www.mcu.es/bibliotecas/MC/LEMBP/index.html>> [Consulta: 17-02-2009]

167 Un extracto de la Clasificación Decimal Universal: <<http://www.mcu.es/libro/docs/TablaCDU.pdf>> [Consulta: 17-02-2009]

168 EAD, <<http://www.loc.gov/ead/>> [Consulta: 17-02-2009]

(desarrolladores del formato MARC)¹⁶⁹, que permite la descripción estructurada de los documentos de archivo y la descripción de las unidades archivísticas a diferente nivel. Para la descripción de las autoridades se ha desarrollado otra DTD, por un grupo de trabajo creado *ad hoc*, la EAC (Encoded Archival Context)¹⁷⁰. Frente a las ventajas de la información estructurada y de la visualización desde la web, incorporando plantillas XSLT, que posibilitan convertir un documento XML en HTML, las EAD y EAC tienen el inconveniente, como lenguaje de marcas o etiquetado, de requerir una base de datos como motor para el almacenamiento y la recuperación de la información, ya sea Oracle, MySQL, etc. El lenguaje de etiquetado es especialmente excelente para la migración de la información, de unas bases de datos a otras.

La ISAD (G), Norma Internacional General de Descripción Archivística, ha sido elaborada por una comisión del Consejo Internacional de Archivos. La norma consta de 26 campos distribuidos en 7 áreas y permite realizar una descripción multinivel. Hay suficientes campos para descripción de las unidades archivísticas, controlando la identificación de la unidad, la descripción, la valoración documental y los aspectos relacionados con dicha unidad. Igualmente existe una norma complementaria para la descripción de autoridades (organismos, familias y personas), la ISAAR-CPF, que sirve para analizar estructuradamente un campo de la ISAD (G), el "Nombre del Productor". Tanto la norma ISAD (G) como la ISAAR-CPF¹⁷¹ definen campos de metadatos pero no el formato. Se pueden utilizar en cualquier base de datos, pero su potencialidad dependerá de las funcionalidades de la base de datos y de la arquitectura que se haya diseñado en el sistema de gestión y el sistema de recuperación de la información.

Los campos de la ISAD (G) y de la EAD tienen una gran similitud, por lo que entre dos sistemas que utilicen estas normas puede haber intercambio de información.

Equivalencias entre los campos de la ISAD (G) y la EAD

¹⁶⁹Encoded Archival Description. Document Type Definition. Version 2002, Chicago : Society of American Archivists, 2002, <<http://xml.coverpages.org/EADv20-dtd.txt>> [Consulta: 17-02-2009]

¹⁷⁰Encoded Archival Context. Document Type Definition. Version Beta (Charlottesville, London, Stockholm), 2003, <<http://xml.coverpages.org/EAC-Beta200408-DTD.txt>> [Consulta: 17-02-2009], proyecto liderado por la Universidad de Virginia, <<http://www.iath.virginia.edu/eac/>> [Consulta: 17-02-2009]

¹⁷¹Las normas ISAD (G) e ISAAR-CPF se pueden localizar en el Consejo Internacional de Archivos, <<http://www.ica.org>> [Web sin servicio temporal, Consulta 17-02-2009]

A.1. ISAD(G) to EAD	
ISAD(G)	EAD
3.1.1 Reference code(s)	<eadid> with COUNTRYCODE and MAINAGENCYCODE attributes <unitid> with COUNTRYCODE and REPOSITORYCODE attributes
3.1.2 Title	<unittitle>
3.1.3 Dates	<unitdate>
3.1.4 Level of description	<archdesc> and <c> LEVEL attribute
3.1.5 Extent and medium of the unit	<physdesc> and subelements <extent>, <dimensions>, <genreform>, <physfacet>
3.2.1 Name of creator	<origination>
3.2.2 Administrative/Biographical history	<bioghist>
3.2.3 Archival history	<custodhist>
3.2.4 Immediate source of acquisition	<acqinfo>
3.3.1 Scope and content	<scopecontent>
3.3.2 Appraisal, destruction and scheduling	<appraisal>
3.3.3 Accruals	<accruals>
3.3.4 System of arrangement	<arrangement>
3.4.1 Conditions governing access	<accessrestrict>
3.4.2 Conditions governing reproduction	<userrestrict>
3.4.3 Language/scripts of material	<langmaterial>
3.4.4 Physical characteristics and technical requirements	<phystech>
3.4.5 Finding aids	<otherfindaid>
3.5.1 Existence and location of originals	<originalsloc>
3.5.2 Existence and location of copies	<altformavail>
3.5.3 Related units of description	<relatedmaterial><separatedmaterial>
3.5.4 Publication note	<bibliography>
3.6.1 Note	<odd><note>
3.7.1 Archivist's note	<processinfo>
3.7.2 Rules or conventions	<descrules>
3.7.3 Date(s) of descriptions	<processinfo><p><date>

Fuente: Biblioteca del Congreso: <http://www.loc.gov/ead/tglib/appendix_a.html#a1>

Otro grupo de metadatos muy utilizados en los archivos son los índices. Lo que en el entorno en papel eran las fichas índices, donde se extraía un vocablo relevante como punto de acceso, en las bases de datos se utiliza el lenguaje natural o el lenguaje documental para indexar, especialmente tesauros¹⁷² como lenguaje que controle las palabras clave.

4. PRINCIPIOS Y ARQUITECTURA EN LA RECUPERACIÓN DE LA INFORMACIÓN

La utilización de los metadatos en la recuperación de la información será eficiente si se cumplen los siguientes principios:

1. La redacción de los campos de metadatos identifican correctamente la unidad de descripción.

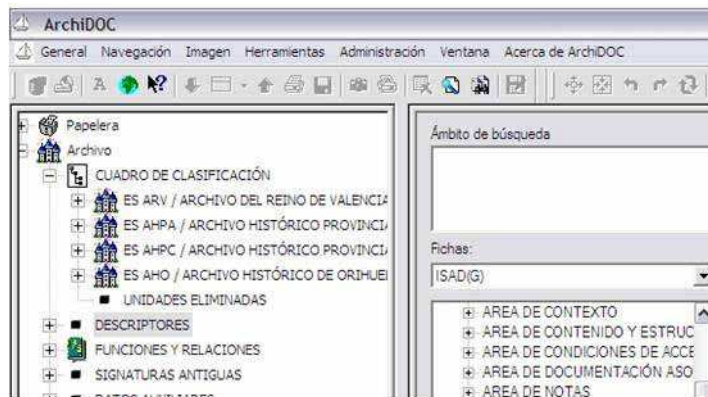
¹⁷²Se han publicado los tesauros que se utilizan en la intranet del Archivo del Reino de Valencia: Giménez Chornet, Vicent; Viciano, Pau; Villalmanzo, Jesús; Escrig, Mercedes, Tesauros del Archivo del Reino de Valencia/Tesaurus de l'Arxiu del Regne de València. Valencia: Generalitat Valenciana, 2006, CD-ROM.

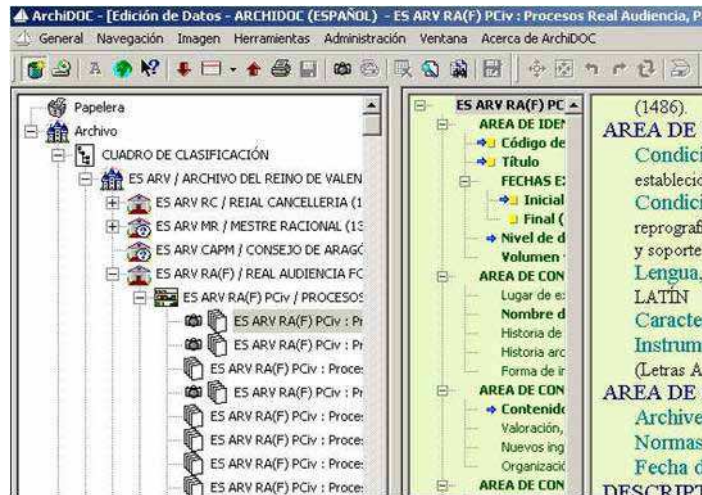
2. La utilización del lenguaje natural es correcta: pertinente y cumple las reglas ortográficas.
3. La utilización del lenguaje documental está integrada en la base de datos y en el sistema de gestión documental.
4. La arquitectura y diseño de la base de datos explota las funcionalidades de los metadatos y la relación entre los campos de metadatos.
5. La arquitectura y diseño del sistema de recuperación de la información explota las funcionalidades de la arquitectura de la base de datos.

La diferencia esencial entre la recuperación de la información en bibliotecas o centros de documentación y en archivos es que en las bibliotecas no hay niveles de descripción y se describen todos los documentos, unos tras otros, y en archivos hay niveles jerárquicos de descripción y, posiblemente, no se describen todas las unidades.

Una buena arquitectura en recuperación de la información debe tener en cuenta la navegación jerárquica en los niveles de descripción, desde los Fondos o Archivos (en casos de un sistema o red de Archivos), los Subfondos, Series, etc.

Navegación Multinivel en un Sistema de Archivos



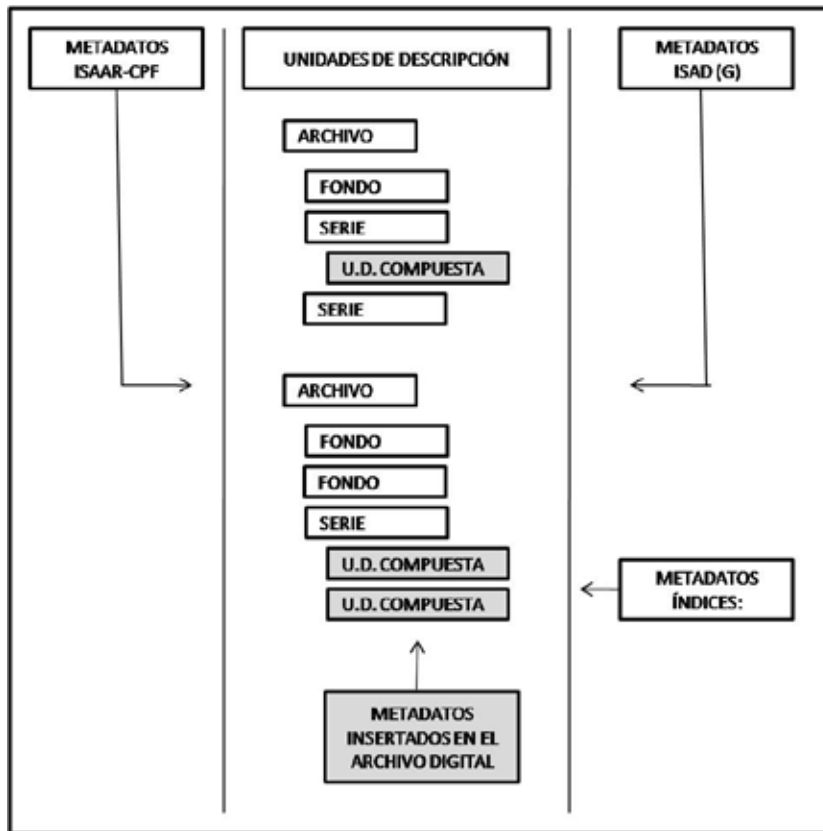


Fuente: Archivo del Reino de Valencia, Intranet.

La recuperación de la información en archivos requiere que los metadatos estén correctamente estructurados y vinculados, siendo el sistema un reflejo de una organización que parte de una clasificación orgánica funcional. La aplicación informática debe permitir registrar los metadatos de contexto (del organismo productor), los metadatos de las unidades, los metadatos de indización y los metadatos insertados en los archivos digitales.

El diseño de los sistemas de recuperación de la información debe explotar esta arquitectura, de forma que las búsquedas avanzadas permitan interrogar qué documentos hay en determinadas unidades de descripción (como podría ser en las series de diferentes archivos), qué hay sobre tal materia indizada, además de interrogar en aquellos parámetros incluidos en la ISAD (G), como podría ser qué hay en tal fecha, de tal productor, o dónde consta tal palabra en un campo determinado como el de "Alcance y Contenido". Los requerimientos de un sistema de recuperación de la información son un reto para los informáticos encargados de la elaboración de software para archivos.

Arquitectura de Metadatos para la Recuperación de la Información



En el diseño de los sistemas de gestión de documentos electrónicos, desde el inicio, es especialmente relevante el establecimiento de los requerimientos para los metadatos. La práctica habitual es la de acumular en los “servidores” o repositorios de documentos electrónicos los documentos generados por los organismos en la administración electrónica, de la misma forma que se acumulan los documentos en los archivos tradicionales de papel. La ausencia de aplicación de criterios archivísticos o documentales en estos sistemas de gestión de documentos electrónicos ocasionará en un futuro próximo una deficiente recuperación de la información. Además, es ya un hecho que cuando se trasladan

los archivos digitales de la administración electrónica al archivo intermedio o histórico, la ausencia de metadatos y de aplicación de las técnicas de organización archivística limita enormemente la búsqueda documental.

Los metadatos que no se describen no se recuperan, y el diseño del software determina la eficacia de la recuperación de la información.

