Un recurso para la localización e indexación de contenidos

POR **DAVID FERNÁNDEZ QUIJADA**

La gestión eficaz de grandes volúmenes de información resulta vital para el progreso de las industrias culturales. El desarrollo de sistemas multimedia conectados a través de redes y la proliferación de contenidos hacen necesaria la implementación de bases de datos multimedia que optimicen esta gestión.

El modelo de Sociedad de la Información que estamos intentando construir en las sociedades desarrolladas «subraya el carácter central del conocimiento teórico como eje alrededor del cual se organizarán la nueva tecnología, el crecimiento económico y la estratificación de la sociedad» (Bell, 2001). La información como elemento central, como materia prima para el funcionamiento del operativo industrial y de servicios, precisa, no obstante, de una base tecnológica que permita su rápida circulación en algunos sectores críticos incluso en tiempo real.

Resulta ingente la cantidad de información que hoy en día está disponible en las redes electrónicas mundiales. No existe un cálculo fiable de la cantidad ni del tipo de contenidos que albergan las grandes redes de comunicación y, paradigmáticamente, Internet. La explosión de contenidos vivida en unos pocos años ha traído consigo las ventajas de una mayor cantidad y variedad de información y de fuentes, pero también ha provocado, fundamentalmente, un problema de selección y localización de contenidos.

La manera fundamental de abordar esta hiperinflación de contenidos es la indexación y el tratamiento automatizado de datos. Estas dos características se reúnen en un elemento cuya importancia es creciente y determinante en los actuales sistemas multimedia: la base de datos. Una base de datos es simplemente un gran índice referencial. Su materia prima son los metadatos, datos que hacen referencia a otros datos.

Existen multitud de bases de datos en todo tipo de sistemas de información. A pesar de su heterogeneidad, hay tres maneras fundamentales de crear y mantener una base de datos (Bricklin, 2001):

a) Manual organizada: es el tipo que utilizan directorios como Yahoo! y que incluye el factor humano. La metodología es tan simple como una persona o un grupo de personas introduciendo manualmente los datos en la base a partir de las fuentes de información disponibles.

b) Automática organizada: es la clase que emplean Google o Altavista mediante robots, denominados arañas en la jerga tecnológica debido a la metodología que emplean, consistente en rastrear la Web explotando las capacidades del hipervínculo para ir descubriendo nuevos enlaces y, de esta manera, tejer una red con la información disponible.

c) Manual voluntaria: los propios usuarios proporcionan los datos y los introducen a partir de su interés por el tema y la voluntad de que se conozca de él o, en muchas ocasiones, para poder crear sus propios bancos de datos. Es el modelo que utiliza la mayoría de las bases de datos multimedia a las que se hará referencia en este artículo.

Una base de datos multimedia universal

Al implementar un sistema de base de datos en un entorno multimedia, se entiende que el objetivo final debería ser crear una base de datos única y universal, un estándar que permitiera un acceso gratuito o previo pago a la información y con una base tecnológica multiplataforma que le hiciera accesible desde cualquier tipo de red y cualquier tipo de dispositivo. Al estilo del sistema ISBN usado en el mundo editorial, pero directamente asignable por vía electrónica y con valor global.

El ISBN proporciona desde 1970 un modelo de identificación único para cada obra editorial publicada. Está reconocido internacionalmente por 159 países y normalizado con el código ISO 2108, lo que le ha permitido convertirse en el sistema de referencia global para el intercambio de libros y para la automatización de su gestión.

Por lo que se refiere específicamente a las bases de datos multimedia, se han desarrollado algunos modelos exitosos, explotados por empresas que han conseguido generar valor a partir de otro valor ajeno, ya que si las descripciones de los contenidos son de calidad pueden adquirir valor en sí mismas (Villegas / García, 2002).

Bases de datos multimedia

Las bases de datos multimedia están viviendo un rápido desarrollo en los últimos años, paralelo a la introducción de Internet y otras herramientas telemáticas en el seno de la

industria. En este sentido, podemos distinguir dos tipos de bases de datos multimedia fundamentales: las referenciales y las descriptivas.

Las bases de datos referenciales son bancos de datos sobre material como películas, series de televisión o música. En la mayoría de los casos la información que se almacena hace referencia a cuestiones descriptivas (autor, título, duración, productor, etc.) o técnicas (formato, duración, etc.).

En cambio, al hablar de bases de datos descriptivas avanzamos un paso más, ya que se trata de sistemas de análisis de contenido que, más allá de los datos técnicos o generales que contiene la mayoría de las bases de datos referenciales, aportan información específica sobre el contenido, indicando, por ejemplo, dónde se sitúan los cambios de plano en una película o la trascripción de un diálogo determinado. Estos bancos de datos no resultan tan habituales y de hecho se encuentran en un estado de desarrollo embrionario, ya que el análisis de la imagen y del sonido no se halla tan automatizado como el del texto.

Existe, sin embargo, un número importante de bases de datos referenciales que actualmente se emplea tanto en entornos cerrados (por ejemplo, las bases de datos que gestionan las plataformas de televisión) como en redes abiertas del tipo Internet, que permite una consulta en muchos casos gratuita y libre por parte de los usuarios. Nos vamos a centrar en algunos de los modelos más desarrollados de este último paradigma.

Modelos de bases de datos multimedia

Uno de los ejemplos más populares de bases de datos multimedia entre los usuarios de Internet es seguramente el de CDDB (www.gracenote.com), siglas que corresponden a Compact Disc Data Base. Cada vez que se introduce un disco compacto en el reproductor de un ordenador personal, CDDB proporciona información en línea sobre la obra discográfica. Esta aplicación se puede encontrar en reproductores como Winamp, Musicmatch o Real Jukebox.

Creada en 1995, CDDB es el ejemplo más completo, el que está haciendo una mayor explotación de su información y el que presenta una orientación comercial más evidente. Su servicio ofrece trece datos distintos para cada disco como título del álbum, artista, sello discográfico, año, género, etc. Para cada pista del CD también ofrece catorce datos adicionales.

Según sus estadísticas más recientes, contiene entradas de cerca de cuatro millones de discos compactos y más de 48 millones de canciones (1). El sistema es multiplataforma y soporta caracteres en diversos alfabetos, ampliando su rango de cobertura a nivel global. Su valor diferencial reside en la conexión con los dispositivos finales, físicos o virtuales, en los que se reproduce el contenido multimedia, ya que, aparte de los reproductores instalados en los ordenadores, su servicio también funciona en diversos dispositivos electrónicos de audio comercializados por marcas como Samsung, Philips, Pioneer o Sony y en la tienda *on line*

iTunes de Apple. Al contrario de los usuarios individuales, estas compañías pagan derechos de licencia para utilizar la información contenida en CDDB.

Este sistema suple la falta de información de la mayoría de discos, cubriendo una carencia cada vez más frustrante para el usuario por la multiplicación de pantallas en todo tipo de dispositivos mediáticos que muestran la fatal leyenda «Disco 1, pista 1». La manera en la que trabaja habitualmente CDDB es el análisis del número de canciones y de la duración de éstas en cada CD, un sistema conocido como TOC (*Table of Contents*). Así, por pura estadística, es remota la posibilidad de que coincidan dos discos de iguales características, por lo que se utiliza como método de identificación.

La recolección de los datos que alimentan este sistema se hace a través de los propios usuarios, que son los que voluntariamente editan los metadatos y los envían a los servidores centrales del sistema para actualizar la información disponible sobre sus colecciones de música. Se trata, según la tipología establecida al inicio del artículo, de una base manual voluntaria.

Proporcionar únicamente datos de obras discográficas editadas significa dejar fuera de su cobertura las canciones individuales que se pueden obtener legal e ilegalmente en la Red, así como la música importada desde otros soportes como el vinilo, el casete o el *minidisc*. Por esta razón, la última implementación de Gracenote, nombre de la empresa propietaria de CDDB, ha sido MusicID, un sistema de reconocimiento de la onda digital de cada canción, una huella única y universal que permite afinar el análisis y proporcionar información de cualquier canción sea cual sea su procedencia y sin posibilidad de error. Esta evolución supone un primer paso hacia el análisis e indexación de los contenidos multimedia, aunque de momento se queda en la forma del objeto, en su superficie.

El sistema de huella digital es el mismo que utiliza desde hace tiempo MoodLogic (www.moodlogic.com). La tecnología de MoodLogic se puede encontrar en dispositivos como ordenadores, asistentes personales digitales, discos duros portátiles, reproductores de CD, MP3 y *minidisc* y, como gran novedad, en los grabadores digitales con disco duro de TiVo. Trabaja únicamente con los archivos musicales, aunque ya ha dado el paso de colarse en la misma caja que la televisión gracias a TiVo.

El desarrollo de una base de datos sobre los programas de televisión parece una empresa mucho más compleja, aunque hay proyectos que trabajan desde hace años en clasificar un material muy apreciado por las televisiones: los filmes. Internet Movie Database (IMDB / www.imdb.com), una extensísima y detallada base de datos sobre películas, nació del grupo de noticias Usenet rec.arts.movies en 1990 gracias a un grupo de aficionados a las películas que compartían su mutuo interés por el séptimo arte. IMDB fue una de las primeras páginas web, y en 1996 dejó de lado sus idealistas inicios para convertirse en empresa y acabar en las manos del gigante *on line* Amazon. La información almacenada en su base de datos es de libre uso para fines no comerciales, aunque existe un servicio *premium* de pago. La fuente de información sigue siendo básicamente la de los aficionados que obtienen los datos de los créditos de las películas y las series de televisión, además de incorporar contenidos originales de los propios usuarios, como resúmenes de los argumentos. IMDB ha ampliado

sus horizontes seminales y actualmente tiene indexados unos 500.000 títulos (2) entre películas, *tv-movies*, series de televisión, miniseries y videojuegos, y a casi dos millones de profesionales que intervienen en ellos.

Una de las últimas aportaciones en este sector ha venido de la mano de Google, empresa que está desarrollando Google Video (http://video.google.com), un sistema, de momento experimental, para el análisis de los contenidos de las principales cadenas de televisión estadounidenses. Yahoo!, uno de sus competidores directos, ya había puesto en marcha anteriormente un servicio similar destinado a la localización de vídeos en Internet (http://video.search.yahoo.com/search/video).

Un problema que todavía queda por resolver es el tamaño de las bases de datos. La producción mediática crece a gran velocidad, lo que aumenta el nivel de dificultad de los sistemas, ya que «los sistemas de información más fáciles de implementar son los que trabajan con colecciones pequeñas y homogéneas destinadas a una comunidad de usuarios también reducida», mientras que «una biblioteca digital global es el escenario más difícil: un vasto y disperso conjunto de colecciones destinada a una comunidad de usuarios amplia y heterogénea» (Borgman, 2000). De hecho, Al Gore ya lanzó la idea de la biblioteca digital global en su famoso discurso del 21 de marzo de 1994 ante la Asamblea de la Unión Internacional de Telecomunicaciones en Buenos Aires.

A la búsqueda de un estándar

A pesar de que CDDB se ha implementado en multitud de aplicaciones, no tiene el valor de estándar normalizado. De hecho, no existe un estándar reconocido, ya que la competencia entre los diferentes proyectos ha sido la norma que ha regido los movimientos de los diferentes actores implicados en el sector. En este sentido, se aprecia una lucha por convertirse en estándar de facto, por su cuota de mercado, entre el ya mencionado sistema CDDB y el sistema ID3 (www.id3.org).

ID3 es un sistema distribuido bajo licencia GPL y, por tanto, de libre uso y distribución. Es compatible con CDDB y ha intentado desarrollar todo un sistema normalizado de especificaciones. El programa fue una idea original de Eric Kemp para incorporar metainformación en los primeros archivos del formato MP3, allá por el año 1996. Aprovechando los bits libres del sistema, Kemp desarrolló una aplicación denominada ID3 que permitía incorporar información sobre el título de la canción, el artista, el álbum, el año, el género e incluso realizar observaciones y comentarios. Este sistema evolucionó con el tiempo y actualmente, en su versión 2.4, se ha optimizado para los archivos destinados al consumo en *streaming*. También permite incorporar una mayor cantidad de metadatos, así como otro tipo de archivos (por ejemplo, una fotografía del autor) y, en general, ha ampliado sus funcionalidades, aunque sigue focalizado en los archivos de audio. No obstante, este conjunto de metadatos se puede encontrar en los archivos audiovisuales de la norma MPEG.

Tanto CDDB como ID3 aspiran a convertirse en un código de barras digital, el código genético de la mercancía cultural electrónica, un identificador universal de archivos sonoros que

requiere, eso sí, una inmensa base de datos.

A pesar de todos estos propósitos, la comunidad científica espera que el auténtico estándar normalizado aparezca con el lanzamiento comercial de la norma MPEG-7, estándar definido dentro del marco de la ISO. La importancia de esta norma radica en la aplicación de una capa semántica, destinada a la descripción e indexación del contenido multimedia, junto a las normas técnicas ya desarrolladas en estándares anteriores como MPEG-1, MPEG-2 o MPEG-4. No en vano, este estándar es denominado formalmente Interfaz de Descripción de Contenido Multimedia, toda una declaración de las funciones que pretende desarrollar.

El estándar MPEG-7 está basado en el lenguaje de programación XML, al igual que otro estándar de metadatos, el RDF (www.w3c.org/rdf). Estas siglas corresponden a la formulación inglesa del Marco de Descripción del Recurso (Resource Description Framework). El objetivo del RDF, aprobado como recomendación por el World Wide Web Consortium, es producir un lenguaje para el intercambio legible por máquinas de cualquier tipo de metadatos sobre descripciones de recursos en el entorno web. En realidad, el sistema de referencia RDF se inscribe en el marco más general de la web semántica (www.w3c.org/2001/sw), un proyecto que persigue un sistema común de metadatos que permita el intercambio de recursos a través de la Red.

En cualquier caso, resulta necesario que los sistemas que se desarrollen cumplan una serie de requisitos mínimos que permitan la búsqueda, localización e intercambio automatizado de los archivos multimedia. Estos requisitos son la compatibilidad sintáctica y semántica (Villegas / García, 2002), es decir, que se utilicen formatos comunes y que los resultados sean consistentes, con significados comunes.

Nuevos retos: Internet y televisión digital

En el incipiente mundo de la televisión digital terrestre, la metainformación sobre los propios productos disponibles resulta esencial, como lo demuestra la importancia que adquiere la guía electrónica de programación (EPG) como elemento de ordenación y acceso de los contenidos. De hecho, la EPG se convierte en un espacio abonado para la aparición de cuellos de botella, elemento de poder para las empresas que obtengan el control de este recurso situado en un lugar estratégico de la cadena de valor.

En un entorno plenamente convergente, todo apunta a que la televisión digital terrestre ha de acabar conectándose a otras redes como Internet gracias al desarrollo de las conexiones a través del televisor o de los servicios asociados de la televisión interactiva. Al menos deberá compartir una misma plataforma tecnológica, como en el caso de las bases de datos.

El ordenador está unido a la Red, y todo lo que por allí pasa (cada vez más productos mediáticos, como el CD, el DVD e incluso la recepción de televisión y radio, aparte del propio contenido multimedia de Internet) debería poder indexarse. La señal de televisión ya incorpora datos con información sobre los programas, por lo que resulta técnicamente viable.

A la vez, la televisión tenderá, con la tecnología digital y los nuevos *set-top boxes*, a conectarse a la gran red global a través de la propia red de difusión, buscando la configuración del Gran Almacén Universal Virtual (Prado, 1997). De hecho, el estándar de televisión digital DVB incorpora una banda de metadatos sensible de ser leída por las EPG. Los discos del formato DVD, por su parte, también contienen información extra, al igual que los productos de la televisión digital. No obstante, parece lógico pensar que la posibilidad de actualizar los datos a través de una red ya sea Internet o la propia red de difusión televisiva se acabará imponiendo, que pueden acabar por fusionarse o, como mínimo, interconectarse.

Para el caso específico de Internet, la norma europea DVB ha desarrollado la especificación DVB-HTML para MHP, basada en el mismo lenguaje de marcas que utiliza la World Wide Web. Ésta permite a los *set-top boxes* presentar aplicaciones de televisión interactiva basadas en HTML, ya que el lenguaje DVB-HTML no es más que una adaptación de HTML para el entorno televisivo.

HTML incorpora la posibilidad de introducir metadatos. La etiqueta define este tipo de información pero su uso no está extendido. Además, los robots de los buscadores que rastrean la Web hacen una explotación mínima de este tipo de datos (Dornfest y Brickley, 2001). Uno de los principales problemas de estos robots de búsqueda es la ausencia de semántica. El paso de HTML a XML como sintaxis de la Web ha de permitir superar este vacío y hacer una mayor explotación de los metadatos.

Conclusiones

El complejo universo de la convergencia multimedia también presenta en las bases de datos una de sus manifestaciones, seguramente más alejada del foco de atención primordial de las grandes compañías. Ello no deja de resultar paradójico dada la potencialidad de estas aplicaciones para constituirse en *gatekeeper* de una importante serie de recursos mediáticos y, entre ellos, el fundamental: los contenidos.

El eje sobre el que se vehicula el desarrollo de las bases de datos es el desarrollo de arquitecturas de metadatos que deberían permitir una indexación ajustada. Esta aproximación las convierte en una herramienta fundamental en la búsqueda y catalogación de material multimedia en las redes digitales. Un paso necesario ante el alud de productos y contenidos disponibles hoy en día. El interés que debería mover a las empresas a implicarse es el suyo propio: la maximización de las posibilidades de difusión de su producto. Además, permitiría la institucionalización de la indexación y la identificación inequívoca de la autoría, con las repercusiones que ello tendría en la gestión de los derechos de propiedad intelectual, un campo que ha impulsado notablemente el desarrollo de estas bases de datos electrónicas.

El paso de los sistemas de tipo referencial a otros de tipo descriptivo, estadio en el que justamente se ha entrado con la última generación de sistemas implementados, ha de permitir dar un salto cualitativo en el manejo y gestión de la metainformación. En algunos casos, la referencialidad es el valor necesario y no se requieren sistemas más sofisticados.

Pero el desarrollo de bases de datos descriptivas, aún incipiente, debe ofrecer oportunidades para nuevas intermediaciones de los actores implicados en las industrias culturales, una vía de crecimiento del negocio a través del desarrollo de funciones hasta ahora inexistentes y cuyo núcleo central es la gestión de la información.

Bibliografía

BELL, D.: *El advenimiento de la sociedad post-industrial*, Alianza, Madrid, 2001.

BERNERS-LEE, T.: *Tejiendo la Red*, Siglo Veintiuno, Madrid, 2000.

BORGMAN, C. L.: *From Gutenberg to the Global Information Infrastructure*, The MIT Press, Cambridge, 2000.

BRICKLIN, D.: «The Cornucopia of the Commons», en ORAM, A. (ed.): *Peer-to-peer*, O□Reilly & Associates, Sebastopol (California), 2001.

COPELAND, M. V.: «The Magic Behind the Music», *Business 2.0*, marzo de 2004, pág. 40. Disponible en la Web: www.business2.com/b2/web/articles/0,17863,592809,00.html

DÍAZ ORTUÑO, P. M.: «Problemática y tendencias en la arquitectura de metadatos web», *Anales de Documentación*, núm. 6, Universidad de Murcia, Murcia, 2003; págs. 35-58. Disponible en la Web: www.um.es/fccd/anales/ad06/ad0603.pdf

DORNFEST, R. y BRICKLEY, D.: «Metadata», en ORAM, A. (ed.): *Peer-to-peer*, O□Reilly & Associates, Sebastopol (California), 2001.

PRADO, E.: «Nuevas tecnologías e interactividad: gran almacén universal virtual», *Diálogos de la Comunicación*, núm. 48, Felafacs, Lima, 1997; págs. 89-95.

VILLEGAS, P. y GARCÍA, D.: «Sistemas de descripción de contenidos multimedia», *Comunicaciones de Telefónica I+D*, núm. 24, enero, Telefónica Investigación y Desarrollo, Madrid, 2002; págs. 133-144. Disponible en la Web: www.tid.es/presencia/publicaciones/comsid/esp/24/art7.pdf.