Un vocabulario controlado para una hemeroteca: posibilidades y características de los Topicsets

José Antonio Moreiro González[1], Guillermina Franco Álvarez[2], David García Martul[3]

[1] Departamento de Biblioteconomía y Documentación. Universidad Carlos III. Getafe. Madrid. jamore@bib.uc3m.es
[2] Departamento de Periodismo y Comunicación Audiovisual. Universidad Carlos III. Getafe. Madrid. gfranco@hum.uc3m.es
[3] Departamento de Biblioteconomía y Documentación. Universidad Carlos III. Getafe. Madrid. dgmartul@bib.uc3m.es

Resumen

La rápida penetración y difusión de las tecnologías de la información en los distintos ámbitos profesionales ha provocado una serie de lagunas teóricas y metodológicas que a la larga impiden su desarrollo ordenado, llegando incluso a ser un obstáculo para las adaptaciones que continuamente están apareciendo. El campo de la comunicación, por sus especiales funciones de difusión y su sensibilidad a las transformaciones en el soporte y el medio, es propicio para proponer sistemas de organización del conocimiento más adaptados al medio digital. Veremos cómo las últimas tendencias de investigación en documentación pueden ayudar en el tratamiento digital de la información en redacciones de prensa, especialmente aquellos saberes relacionados con el análisis de contenido y los lenguajes controlados. Un primer punto de partida para aplicar las aportaciones de la documentación a la organización y gestión de la documentación digital en prensa estaría centrado en las hemerotecas. Y más concretamente desde la disciplina del análisis de contenido, es de particular interés en los lenguajes de marcado, aplicados a la gestión interoperable y difusión distribuida de las noticias, los esquemas de metadatos y los valores asignados a cada una de sus propiedades: los lenguajes controlados. Concretamente nos vamos a centrar en el lenguaje controlado utilizado en NewsML: Topicsets. Su objetivo no es otro que proporcionar un listado de valores normalizados para los metadatos empleados por este lenguaje. Concluimos valorando las propiedades que como vocabulario controlado presentan los Topicsets.

Palabras clave: Hemeroteca, IPTC, NewsML, TopicSets, Vocabularios controlados.

Abstract

The rapid penetration and diffusion of the technologies of the information in the different professional areas has provoked a series of theoretical and methodological lagoons that eventually prevent his tidy development, managing to be even an obstacle for the adjustments that constant are appearing. The field of the communication, for his special functions of diffusion and his sensibility to the transformations in the support and the way, is propitious to propose systems of organization of the knowledge more adapted to the digital way. We will see how the last trends of investigation in documentation can help in the digital treatment of the information in drafts of press, specially those knowledge related to the analysis of content and the controlled languages. The first point of item to apply the contributions of the documentation to the organization and management of the digital documentation in press would be centred on the newspaper libraries. And more concretly from the discipline of the analysis of content, it is of particular interest in the languages of marked, applied to the interoperable management and diffusion distributed of the news, the schemes of metadatos and the values assigned to each of his properties: the controlled languages. Concretly we go away to centring in the controlled language used in NewsML: Topicsets. His aim is not other one that to provide a list of values normalized for the metadata used by this language. We end up by valuing the properties that as controlled vocabulary they present the Topicsets.

Keywords: Controled vocabulary, IPTC, NewsML, Newspaper archive, TopicSets.

1 Introducción

La introducción de las nuevas tecnologías de la información en el sector mediático ha marcado un nuevo ciclo evolutivo en la actividad periodística. La prensa digital permite una difusión sin precedentes, facilidad de acceso, inmediatez, economía de medios, extensión ilimitada, e información a la carta, eliminando las restricciones de tiempo y espacio, y alterando las rutinas de los equipos de redacción (Franco Álvarez, 2005, p. 168). Se han transformado profundamente las formas de comunicación entre los distintos agentes especializados que participan en la industria mediática, con la puesta en funcionamiento de nuevas infraestructuras, protocolos y estándares de intercambio, para la distribución de paquetes informativos a través de distintos canales y formatos de transmisión. Sin embargo, se hace necesario coordinar el empleo de todas las tecnologías empleadas en los medios para poder mejorar la interoperabilidad entre sistemas tanto de información como de recuperación.

A nivel interno, la tendencia en los productores de información apunta a la adopción de plataformas integradas que proporcionen soporte a todo el ciclo de elaboración, gestión y publicación de contenidos, abarcando la recepción de información externa (p.e. noticias de agencia), la elaboración de contenidos propios, maquetación, documentación o gestión de archivo. Consecuencia de esta transformación tecnológica en la industria mediática ha sido la emergencia de un nuevo mercado de servicios online para la redistribución, sindicación, agregación, y distribución de noticias de archivo. Así, por ejemplo, podemos citar iniciativas como las bases de datos de noticias de prensa a partir de las cuales se pueden realizar búsquedas por distintas materias[1], o iniciativas de bibliotecas y centros de documentación públicos como la BritishLibrary.

[1] La base de datos NewsLibrary proporciona y vende noticias publicadas fundamentalmente en prensa estadounidense. En: www.newslibrary.com. Consultado el 12/12/2006.

La BritishLibrary proporciona desde su portal el acceso a un proveedor de servicios propio, a modo de gateway o pasarela. Uno de estos enlaces apunta a un directorio de recursos sobre prensa digital denominado: "News and media resources". El sistema de recuperación por navegación es particularmente adecuado para materias interdisciplinarias y cuando el objetivo es obtener una visión global acerca de una materia o bien mantener informado al usuario sobre las actualizaciones de las noticias (Marchionini, 1997, p. 100). No obstante, la novedad de este recurso es que a través de este directorio de prensa en línea, tanto internacional como británica, no sólo se permite la recuperación por navegación sino que es posible acceder a sus respectivas hemerotecas[2]. Sin embargo, presenta el problema de no ser capaz de realizar búsquedas simultáneas en cada uno de los medios por materias, sino que nos obliga a acceder a cada una de las hemerotecas para poder recuperar las noticias registradas por cada medio sobre una materia. Es decir, carece de un sistema de información que unifique la recuperación desde una misma interfaz en las distintas hemerotecas a las que se encuentra suscrita la BritishLibrary y el primer problema para poder manejar de forma simultánea los sistemas de búsqueda en todas las hemerotecas es la falta de interoperabilidad entre las mismas.

La hemeroteca de un periódico es un producto informativo muy valioso para una amplísima variedad de consumidores, pero también para los propios redactores de noticias. Las noticias recogidas de la publicación diaria en medios de comunicación representan un enorme volumen de información muy vagamente organizada, comparada con una biblioteca o un centro de documentación tradicional, que crece sin una planificación previa a medida que se incrementa con nuevos documentos informativos de estructura y formato heterogéneo. Esta falta de organización, tanto documental como de contenidos, va a obstaculizar la definición de una web cooperativa (Gayo Avello, 2002) que facilite no sólo la difusión compartida entre agencias y medios de comunicación, sino el diseño de un marco común de interoperabilidad para la aplicación de sistemas de organización del conocimiento en red (NKOS).

Este corpus en constante crecimiento de noticias archivadas resulta del trabajo coordinado, pero en buena medida autónomo, de un grupo de reporteros cuyo objetivo primario no es construir una hemeroteca, sino proporcionar el mejor producto informativo posible para el consumo inmediato. Los periodistas suelen estar asistidos por documentalistas que ayudan a clasificar, indexar y anotar noticias cuando éstas pasan al archivo, utilizando software específico para la gestión de hemerotecas. Además se necesitan mecanismos potentes de búsqueda y exploración para que los consumidores de información puedan encontrar su camino en este espacio informativo. La tecnología actual proporciona sistemas de búsqueda basada en palabras clave (a menudo por campos: titular, resumen, cuerpo, sección), funcionalidades de exploración dentro de un ejemplar de un periódico y, en ediciones web, navegación a través de hiperenlaces creados a mano entre noticias como pueden ser los enlaces a antecedentes de una noticia.

Sin embargo, se constatan problemas importantes en las capacidades de los sistemas de gestión documental aplicados a las hemerotecas como la incapacidad de buscar documentos que carezcan de información textual (Salazar, 2005, p. 160), lo cual es muy frecuente en medios de comunicación audiovisual, o la incapacidad de recuperar tanto imagen fija como en movimiento por el contenido o las propiedades intrínsecas de las mismas. Posibilidad que ya

[2] El directorio de la BritishLibrary es el más importante recurso de acceso gratuito a hemerotecas en Europa. En: www.britishlibrary.net Consultado el 12/12/2006.

se está ofreciendo en información museística con el motor de recuperación de imágenes QBIC[3] de IBM.

Existe aún un amplio margen para aprovechar las posibilidades que ofrece el medio digital para la explotación de una hemeroteca. Los aspectos que pueden ser mejorados incluyen: a) la búsqueda basada en palabras clave resulta limitada en capacidad expresiva; b) débil interrelación entre elementos de la hemeroteca: los usuarios pueden necesitar combinar varias consultas indirectas manualmente para obtener respuestas a consultas complejas; c) falta de un estándar comúnmente adoptado para la representación y distribución de noticias entre periódicos; d) falta de consenso interno entre periodistas y documentalistas sobre la terminología de descripción de contenidos; e) falta de implicación de los redactores en el proceso de archivo. Una de las novedades más prometedoras en la línea de superar el tipo de limitaciones señaladas, para un sistema de gestión de información de las características descritas, son las emergentes tecnologías de la web semántica, que propone nuevas técnicas, paradigmas y estándares para la representación del conocimiento que faciliten la localización, distribución e integración de recursos en la W3C.

La implantación de las nuevas tecnologías de la información en las redacciones en prensa ha supuesto una profunda renovación de los sistemas de gestión documental de sus centros de documentación. Sin embargo, estos nuevos programas en muchas ocasiones son deficientes en cuanto a las posibilidades de recuperación documental ofrecida. Deficiencias que son más evidentes cuando se trata de la recuperación de documentos multimedia; documentos con unas características intrínsecas que hacen poco adecuados los tesauros por su estructura jerárquica. Esto nos lleva a reformular los lenguajes documentales tradicionales para poder fomentar el empleo de un esquema de relaciones asociativas que facilite una recuperación por materias más dinámica y no tan sintagmática tal y como propuso en su momento Maniez (2002). Esta propuesta sería desarrollada más adelante por profesionales de la biblioteconomía y organización del conocimiento como Rebecca Green (2004) y Carol Bean (2004) cuando plantean en el marco del congreso internacional de ISKO en el año 2004 el empleo de indicadores semánticos a partir de estructuras verbales previamente definidas. Esta línea de investigación sería tomada por Moreiro (2004) y Llorens para plantear un tesauro que hiciera uso de algún tipo de forma verbal capaz de sistematizar la recuperación en un nuevo lenguaje documental, ya definido por entonces como ontología, denominado Topic Maps (Pepper).

Por otra parte, los trabajos realizados en tesauros y sistemas de clasificación están siendo recogidos por los especialistas en computación, lingüística computacional e inteligencia artificial para proponer las ontologías como esquema interactivo y navegable para la organización del conocimiento y la recuperación de información en red a fin de poder proporcionar herramientas de recuperación por contexto, o mas bien por la semántica de los textos.

[3]Hermitage Museum. En: http://hermitagemuseum.org/fcgi-bin/db2www/browse.mac/category?selLang=English. Consultado el 12/12/2006.

2 Método. Normalización para la edición de noticias compartibles e interoperables entre agencias de noticias y medios de comunicación

El consorcio IPTC es una federación[4] de las mayores agencias de noticias del mundo, entre las que se encuentra la agencia EFE. Su meta es la definición y edición de normas para la transmisión normalizada de las noticias editadas por cada una de las agencias de noticias con independencia de la lengua, el formato o la plataforma que empleen para su edición.

El IPTC se constituyó en 1965 a partir de un grupo de organizaciones de agencias de noticias tales como Alliance Européenne des Agences de Presse, ANPA (ahora NAA), FIEJ (ahora WAN) y la North American News Agencies (un comité conjunto de Associated Press, Canadian Press y United Press International) con el objetivo de salvaguardar los intereses que en telecomunicaciones podía tener la prensa internacional.

Desde finales de los años 70, la actividad del IPTC se ha centrado en primer lugar en el desarrollo y edición de normas para el intercambio de noticias.

Los objetivos para los que se creó el IPTC son[5]:

(a) Investigación y desarrollo en telecomunicaciones para mejorar el flujo de noticias.
(b) Formulación de las exigencias que la prensa hace de las telecomunicaciones y llamar la atención de las autoridades en telecomunicaciones a fin de que valoren la introducción de las demandas que las agencias de noticias hacen de las telecomunicaciones.
(c) Que los medios de prensa estén representados en organizaciones nacionales e internacionales o comités que tratan sobre telecomunicaciones.
(d) Publicación de toda aquella información relativa al progreso técnico y al desarrollo en el campo de las telecomunicaciones.

3 Discusión. El lenguaje controlado en NewsML

NewsML proporciona un marco estable para los metadatos donde los valores que le pueden ser asignados son básicamente de dos tipos: controlados y no controlados. Una colección de valores controlados por una organización se la conoce como “vocabulario controlado”.

El Consorcio IPTC tiene una lista de vocabularios controlados a fin de poder normalizar la asignación de valores a sus metadatos por parte de las distintas agencias de noticias. A este vocabulario controlado se le conoce como IPTC NewsCodes[6].y, concretamente, en su epígrafe de Vocabularios Controlados los denomina TopicSets o colecciones de Topics al igual que hace la norma ISO 13250 de Topic Maps[7].

[4] IPTC Home. En: http://www.iptc.org/pages/index.php Consultado el 20/03/06.
[5] IPTC. En: http://www.iptc.org/pages/about_wpgme.php. Consultado el 12/12/2006
[6] IPTC NewsCodes. En: http://www.iptc.org/metadata Consultado el 04/12/2006
[7] ISO/IEC 13250:2000 Topic Maps. En: www.y12.doe.gov/sgml/sc34/document/0129.pdf Consultado el 12/12/2006.

El empleo de un vocabulario controlado significa que los elementos del vocabulario están controlados por una organización, que en este caso es un Consorcio para la normalización. El término de "vocabulario controlado" significa que los valores a asignar a los metadatos deben ser coherentes con la lista de términos del vocabulario pero no que sean idénticos.

La definición más básica de un vocabulario controlado que podemos dar es que se trata de una lista de valores, en la que cada uno de ellos es expresado por medio de una cadena de caracteres (Gil Urdiciain, 2004). Sin embargo para el caso de los metadatos empleados en NewsML se especifica la notación a emplear en la DTD de NewsML, en el apartado referido al atributo Formalname donde se dice que una lista de valores controlados es "una cadena de caracteres cuyo significado está determinado por un vocabulario controlado". A pesar de que cualquier proveedor de noticias puede proporcionar una lista de valores referenciales para un vocabulario controlado, el consorcio IPTC recomienda el empleo de los TopicSets como vocabulario controlado.

El consorcio IPTC define un TopicSet como una colección de topics con un nombre[8]. Pero ¿qué es un topic? NewsML recogió la definición que se da de topic en la norma ISO 13250 y dice que un topic no es más que cualquier cosa o concepto del mundo real que pueda ser referenciada en una noticia. Por tanto, un elemento topic no es mas que una representación de un concepto o algo real, pero un topic no es capaz de expresar nada fuera de la fuente o noticia a la que está referenciada. La función del FormalName es la de ser una referencia externa de la que dispone un topic. Aparte del formalname los topics cuentan con elementos tales como tipo de identificación, descripciones y propiedades. noticia[9].

En muchas ocasiones los vocabularios controlados proporcionan valores para los metadatos que son importantes para la interpretación global de una noticia. Por coherencia con el tratamiento de las noticias por parte de diferentes proveedores es fundamental que sean adoptados valores comunes en determinados campos clave tales como: NewsItemType, Status, Location o Subject. De hecho algunos de estos campos son considerados obligatorios por parte de las agencias de noticias a fin de poder efectuar la sindicación de contenidos.

A modo de conclusión podemos ver el ejemplo práctico siguiente tomado para una noticia de prensa[10], donde se refleja en la figura 1 la lista de metadatos empleada por NewsML en lengua española. Donde Subject, SubjectMatter y SubjectDetail son tipos de topic[11], los valores asignados a cada uno de estos tipos de topic constituyen los topicset que vemos insertados con la etiqueta <topicset> en el lenguaje NewsML.

[8] NewsML Guidelines 1.2. En: http://www.newsml.org/pages/index.php . Consultado el 10/12/2006
[9] Id.
[10] Id.
[11] IPTC News Code. En: http://www.iptc.org/NewsCodes. Consultado el 12/12/2006

```
<NewsML>
    <Catalog Href="http://www.iptc.org/IPTC/catalog/catalog.IptcMasterCatalog.xml"/>
    <NewsEnvelope> ... </NewsEnvelope>
    <NewsItem>
        <Identification>...</Identification>
        <NewsManagement>...</NewsManagement>
        <TopicSet Duid="iptc.status" Scheme="IptcTopicType" FormalName="Status">
            <Comment xml:lang="en-GB">The current usability of a NewsItem.</Comment>
            <Comment>xml:lang attribute values updated</Comment>
            <Topic Duid="stat1">
                <TopicType Scheme="IptcTopicType" FormalName="Status"/>
                <FormalName Scheme="IptcStatus">Usable</FormalName>
                <Description variant="Name" xml:lang="en-GB">Usable</Description>
                <Description variant="Explanation" xml:lang="en-GB">The NewsItem and its content may be
published without restriction.</Description>
                <Description variant="ChangeComment" xml:lang="en-GB">none</Description>
                <Description variant="ChangeVersion">0</Description>
            </Topic>
            <Topic Duid="stat2">
                <TopicType Scheme="IptcTopicType" FormalName="Status"/>
                <FormalName Scheme="IptcStatus">Embargoed</FormalName>
                <Description variant="Name" xml:lang="en-GB"> Embargoed</Description>
                <Description variant="Explanation" xml:lang="en-GB">Neither the NewsItem nor its content may be
published until released for publication by the provider.</Description>
                <Description variant="ChangeComment" xml:lang="en-GB">none</Description>
                <Description variant="ChangeVersion">0</Description>
            </Topic>
            <Topic Duid="stat3">
                <TopicType Scheme="IptcTopicType" FormalName="Status"/>
                <FormalName Scheme="IptcStatus">Withheld</FormalName>
                <Description variant="Name" xml:lang="en-GB"> Withheld</Description>
                <Description variant="Explanation" xml:lang="en-GB">Neither the NewsItem nor its content may be
published until further notice.</Description>
                <Description variant="ChangeComment" xml:lang="en-GB">none</Description>
                <Description variant="ChangeVersion">0</Description>
            </Topic>
            <Topic Duid="stat4">
                <TopicType Scheme="IptcTopicType" FormalName="Status"/>
                <FormalName Scheme="IptcStatus">Canceled</FormalName>
                <Description variant="Name" xml:lang="en-GB">Cancelled</Description>
                <Description variant="Explanation" xml:lang="en-GB">Neither the NewsItem nor its content may be
used under any circumstances. If the NewsItemor its content has been published the publisher must take immediate
action to withdraw or retract it, as may be legally necessary.</Description>
                <Description variant="ChangeComment" xml:lang="en-GB">none</Description>
                <Description variant="ChangeVersion">0</Description>
            </Topic>
        </TopicSet>
    </NewsItem>
</NewsML>
```

Fig. 1. El vocabulario controlado inserto en el lenguaje NewsML a través de la etiqueta TopicSet.

4 Conclusión

En este trabajo hemos querido hacer una primera valoración de las capacidades de un nuevo recurso orientado a servir como lenguaje controlado en una hemeroteca: los topic sets. Su propósito es facilitar la recuperación de información contextual en las redacciones de medios de comunicación y agencias de noticias. Los topic sets, tal y como su nombre nos hace pensar son una red semántica o una representación gráfica conceptual, pero tienen un potencial muy superior de organización y gestión del conocimiento para entornos de información virtual distribuidos. Por tanto, una ventaja del empleo de los topic sets en las agencias de noticias es que se encuentran en formato XML, lo cual facilita considerablemente la interoperabilidad de la información entre sistemas. Esto está en línea con las actuales tendencias sobre el desarrollo de las tecnologías web orientadas hacia la elaboración de ontologías y sistemas interoperables como paso previo para la consecución de la web semántica.

Hay una obvia necesidad de que existan entornos especializados diseñados para permitir al redactor elaborar noticias basadas en web sobre la base de las necesidades de difusión que cada usuario demande sin necesidad de conocer los lenguajes controlados que pueda haber detrás de cada sistema de gestión documental.

Actualmente quedan pendientes de investigación cuestiones tales como: cómo dotar a los redactores de las noticias de una serie de componentes basados en topic sets para ser empleados en distintos sistemas desde los cuales las noticias puedan ser rápidamente recuperadas desde la hemeroteca del medio; o cómo permitir que el usuario pueda personalizar las facetas para un mismo material informativo; y cuáles son los interfaces más idóneos para la visualización y acceso a la información en distintos contextos periodísticos.

Todas estas cuestiones son por las que pasarán las próximas investigaciones en el empleo de topic sets para la elaboración de materiales informativos. Un área en incipiente surgimiento, con un gran potencial de desarrollo en el contexto de la Sociedad del Conocimiento que esperamos se vaya consolidando a partir de acuerdos para su aplicación en el seno del consorcio IPTC entre las distintas agencias y principales medios de comunicación.

Bibliografía citada

BEAN, C. Representation of medical knowledge for automated semantic interpretation of clinical reports. En: MCILWAINE, Ia C. *Knowledge Organization and the Global Information Society*. Londres: Ergon, 2004.

FRANCO ÁLVAREZ, G. *Tecnologías de la comunicación: producción, sistemas y difusión digital*. Madrid: Fragua, 2005.

GAYO AVELLO, D. *Web cooperativa* [recurso electrónico]. Oviedo: Universidad de Oviedo, 2002. <www.di.uniovi.es/~dani/research/trabajo-investigacion.pdf > [Consultado: 18 dic. 2006]

GIL URDICIAIN, B. *Manual de lenguajes documentales*. Gijón: Trea, 2004.

GREEN, R. Patterns in verbal polysemy. En: MCILWAINE, Ia C. *Knowledge Organization and the Global Information Society*. Londres: Ergon, 2004.

IPTC NewsCodes List [recurso electrónico]. <http://www.iptc.org/NewsCodes/nc_ts-table01.php> [Consultado: 12 dic. 2006].

MANIEZ, J. *Actualité des langages documentaires*. Paris: ADBS, 2002.

MARCHIONINI, G. *Information seeking in electronic environments*. Cambridge: Cambridge University Press, 1997.

MOREIRO GONZÁLEZ, J. A. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural*. Gijón: Trea, 2004.

MOREIRO GONZÁLEZ, J. A. [et al.]. Nuevos patrones en la representación y la visualización de la información para entornos distribuidos: del tesauro al Topic Map. *Códice: Revista de la Facultad de Sistemas de Información y Documentación*, 2005, n. 1.

NewsML Guidelines 1.2. [recurso electrónico] <http://www.newsml.org/pages/index.php>. [Consultado: 10 abr. 2006]

PEPPER, S. The TAO of Topic Maps [recurso electrónico] <http://www.ontopia.net/topicmaps/materials/tao.html> [Consultado: 12 dic. 2006]

SALAZAR, I. *Las profundidades de Internet*. Gijón: Trea, 2005.