Análisis

Evolución y uso de los lenguajes controlados en documentación informativa

Por **Lourdes Castillo** y **Alejandro de-la-Cueva**

Resumen: *La documentación periodística adolece de instrumentos adecuados de clasificación e indización de la información, a excepción del Subject Reference System del Internacional Press Telecommunications Council (IPTC), todavía en estado incipiente de desarrollo. Se hace una revisión de las contribuciones más relevantes sobre la clasificación e indización de noticias en los medios de comunicación, tanto españoles como internacionales. También se estudia la elaboración y uso de vocabularios controlados para el tratamiento de la información de actualidad y sobre todo de tesauros especializados. Por último, se destacan algunas características específicas de la documentación periodística que condicionan la utilización de tesauros para la indización y recuperación de información de actualidad.*

Palabras clave: *Documentación periodística, Lenguajes controlados, Tesauros, Clasificaciones, Indización de prensa, Servicios de documentación de medios de comunicación.*

Title: Evolution and use of controlled languages in news documentation

Abstract: *News documentation lacks suitable instruments of classification and indexing, the only exception being the Subject Reference System of the International Press Telecommunications Council (IPTC), which is not yet fully developed. The most relevant contributions to classifying and indexing news in Spanish-language media, both in Spain and internationally, is reviewed. We also studied the development and use of controlled vocabularies for processing news information, primarily specialized thesauri. Finally, news documentation characteristics specific to indexing and information retrieval for print news media are described in detail.*

Lourdes Castillo *es doctora en geografía e historia por la Universitat de València. Ha impartido clases en la diplomatura de biblioteconomía y documentación de la Universitat de Valencia y es documentalista en la Unidad de Documentación de RTVV.*

Alejandro de la Cueva*, doctor en cirugía y medicina (1987) por la Universidad de Valencia. Especialista en documentación médica por la Universidad de Valencia (1990). Profesor titular de biblioteconomía y documentación del Departamento de Historia de la Ciencia y Documentación de la Universidad de Valencia, desde 1988. Analista de la base de datos Índice Médico Español (IME) desde 1973 a 1992. Vicedirector de la misma entre 1979 y 1989.*

Keywords: *News documentation, Controlled languages, Thesauri, Classifications, News indexing, News reference services.*

Castillo, Lourdes; Cueva, Alejandro de la. "Evolución y uso de los lenguajes controlados en documentación informativa". En: *El profesional de la información,* 2007, noviembre-diciembre, v. 16, n. 6, pp. 617-626.

DOI: 10.3145/epi.2007.nov.08

Introducción

EN ESPAÑA SE EDITAN 95 PERIÓDICOS de los cuales 66 tienen edición electrónica, es decir un 69% del total[1]. En el mundo las ediciones digitales de la prensa en lengua española constituyen por su volumen el segundo segmento lingüístico aunque distantes en número de los 1.236 diarios en inglés (Díaz Nosty, 1999).

Un archivo periodístico de tamaño medio, como el de *Radiotelevisión Valenciana* (*RTVV*), selecciona un total de 289.631 noticias desde 1998 hasta 2005, con un crecimiento medio de más de 36.000 noticias al año. Un volumen mayor, de aproximadamente 139.404 noticias correspondientes a los años 2003 a 2005, contiene la base de datos de prensa de *Euskal Telebista*. Por otra parte, los servicios de seguimiento de prensa pueden almacenar volúmenes ingentes de noticias. Es el caso del servicio de información de prensa digital *iMente*, que diariamente captura unas 115.000 noticias diarias (**Guallar**, 2006).

El acceso inmediato a la información periodística y la interactividad son dos de las principales características que diferencian a los diarios digitales, pero no las únicas. La posibilidad de actualización y corrección continua; la sencillez para el usuario de copiar, editar y archivar los textos íntegros; el acceso a los números atrasados y a colecciones enteras de periódicos son otras de las posibilidades que ofrece la prensa distribuida a través de internet. Aparentemente el acceso casi generalizado a las telecomunicaciones parece solucionar el problema de la búsqueda de información; sin embargo los problemas en la recuperación de información en cuanto a calidad y pertinencia persisten y es necesario seguir investigando, además de estudiar cómo adaptar las herramientas tradicionales a los nuevos entornos y usuarios, los que se han dado en llamar "usuarios casuales", con habilidades de recuperación limitadas.

El control de vocabulario es una realidad en documentación científica y en otras áreas como la documentación administrativa o la legislativa y jurídica. Esa normalización se refleja en herramientas de control que se aplican en grandes sistemas como *PubMed* o *FSTA*. Algunos ejemplos de estas herramientas terminológicas son tesauros como el *MeSH*, el *Agrovoc*, el macrotesauro de la *OCDE*, *Eurovoc*, *Hasset*, *Sosig*, *Glin*, *Spines*, *Eric*, *Envoc*, y muchos más.

Frente a esta normalización contrasta la ausencia en documentación periodística de instrumentos generales consensuados respecto a la indización y clasificación de documentos. Cada medio de comunicación utiliza su propio lenguaje controlado, elaborado independientemente; y en muchas ocasiones se trata de vocabularios de bajo nivel como clasificaciones o listas de autoridades. A esto hay que añadir una tendencia que no ha beneficiado nada al desarrollo de los lenguajes controlados en documentación periodística: el empleo del lenguaje natural en la recuperación de información de actualidad.

"Cada medio de comunicación utiliza su propio lenguaje controlado, elaborado independientemente; y en muchas ocasiones se trata de vocabularios de bajo nivel, como simples listas"

Tan sólo el *Subject Reference System*[2], la clasificación temática desarrollada por el *Consejo Internacional de Telecomunicaciones en Prensa* (*Internacional Press Telecommunications Council*, *IPTC*), puede representar este instrumento consensuado, pero todavía poco desarrollado.

El presente trabajo tiene como objetivo hacer una revisión de estudios relevantes sobre la evolución de la clasificación e indización de los documentos de prensa y la elaboración y utilización de vocabularios controlados, especialmente del tesauro, en documentación informativa. También se resaltan algunas de las causas que se señalan en la literatura como condicionantes de la escasa utilización de los tesauros en este medio.

Estudios sobre la documentación y la indización de la prensa

El decano de la indización de prensa es **William Frederick Poole**, que en 1848 publicó *Index to subjects treated in the reviews and other periodicals.* En 1853 mejoró este primer esfuerzo con el título *Index to periodical literature*, obra considerada precursora del *The New York times index* (**Semonche**, 1993). En 1902 **Henry B. Wheatley** publicó *How to make an index*, que es una guía sobre el proceso de indización (**Semonche**, 1993). Hasta 1909 los estudios sobre los servicios de documentación periodística fueron escasos y esporádicos. Quizás el primer artículo sobre la actividad documental informativa fue el de **I. D. Marshall**, "Article on methods in newspaper libraries", publicado en una de las primeras revistas de periodismo, *Newspaperdom*, en su primer número de 1892. Este texto hace las primeras observaciones sobre la labor de clasificación y archivo de los materiales de referencia. En 1893 **Bardwell** se refirió también a las tareas de clasificación y archivo de recortes en el artículo *Scrapbooks*, *Clippings*, *etc*., publicado en *Library journal* (**Galdón**, 1986).

González-Quesada (1995) explica cómo se constituyó en 1923, dentro de la *Special Libraries Association*, un *Grupo de Periodismo* dedicado a analizar y sistematizar el funcionamiento de los servicios de documentación en las empresas periodísticas. La revista *Special libraries* permitió hacer públicos los estudios iniciales sobre los servicios de documentación periodística y contribuyó decisivamente a dar a conocer los métodos empleados en la labor de clasificación y archivo de diversos rotativos norteamericanos. **Hansen** y **Ward** (1991) publicaron un trabajo sobre el uso de bases de datos electrónicas y otras tecnologías de la información en 105 diarios de gran tirada en 1990. Por su parte **Hegg** (1991) hizo un estudio comparativo sobre el almacenamiento y los métodos de recuperación automatizados en pequeños diarios de EUA. El *Grupo de Periodismo* constituyó además un comité para elaborar una clasificación estándar, pero no lograron su propó-

sito. De hecho, en 1933 no había dos centros de documentación periodísticos con el mismo sistema de clasificación, aunque hubiese coincidencias parciales. **Galdón** (1994) enumera los criterios que seguían la mayoría: elaborar una clasificación específica para recortes periodísticos; acotarla en apartados de temas, personas y países; fijar encabezamientos y subdivisiones simples y fáciles de recordar y permitir la codificación por varias entradas.

En su capítulo sobre indización de prensa de 1993, **Semonche** destaca las aportaciones de **Desmond** y **Friedman**. **Robert W. Desmond** publicó en 1933 *Newspaper reference methods*. Este libro analiza y valora los aspectos de estructura y método, a la vez que hace hincapié en el tratamiento de los problemas derivados de la clasificación. **Harry Friedman** escribió en 1942 *Newspaper indexing*, otra obra sobre indización de prensa.

La *American Newspaper Publisher Association* (*ANPA*) publicó en 1974 *Guidelines for newspaper libraries*, reeditada en 1976. Las *Guidelines* constan de 16 capítulos sobre otros tantos aspectos que, según la *Newspaper Division* de la *Special Libraries Association*, se deben tener en cuenta al plantear los servicios de documentación periodística. Uno de estos capítulos está dedicado a la indización y expone los diferentes modos de realizar índices y la forma de usarlos.

En el ámbito europeo la obra más destacable a nivel teórico es la de **Geoffrey Whatmore**, que en 1964 publicó *News information: the organization of press cuttings in the libraries of newspaper and broadcasting services*, traducido en España en 1970 con el título *La documentación de la noticia: organización y métodos de trabajo para archivos de referencia de periódicos y agencias*. Este libro consta de 17 capítulos y en dos de ellos se hace referencia a los índices, pero siempre hablando de sistemas manuales. Expone los sistemas de índices de varios periódicos (*Financial times*, *The guardian*, *Daily mail*, *Daile sketch*, *Evening news*, *Glasgow herald*) en forma de tarjetas con marcadores de colores, tiras superpuestas con márgenes inferiores visibles, en volúmenes y hojas intercambiables.

"A nivel internacional, el tesauro de prensa más destacado por todos los autores es el empleado en la base de datos del diario *The New York times*"

En 1973, con ocasión del primer congreso sobre servicios de documentación periodística organizado en Gran Bretaña por la *Association of Special Libraries and Information Bureaux* (*Aslib*), **Whatmore** expuso, bajo el título *Classification for news Libraries*, los problemas de clasificación específicos de estos servicios. El mismo autor publicó en 1978 el libro *The modern news library*, en el que pone al día los métodos prácticos de archivo y organización de los servicios de documentación, pormenorizando las soluciones a las posibles dificultades prácticas y aportando una lista de pautas a seguir en las tareas de selección y clasificación. En 1979 se publicaron las *Normas para la presentación de índices analíticos en centros de documentación y archivos de publicaciones periódicas,* preparadas años antes por **Justo García-Morales**, director del *Centro Nacional del Tesoro Documental y Bibliográfico*. **García-Morales** (1979) afirma que la dificultad de organizar un centro de documentación de prensa estriba en la universal amplitud de las materias, en el tiempo y en el espacio geográfico que reflejan. Aconseja que el sistema adoptado sea ante todo práctico y rápido, considerando que el objeto de la elaboración de índices es facilitar la consulta urgente y la localización de cualquier dato o curiosidad, contenido en un diario o publicación de índole informativa.

En 1991 **Hans H. Wellisch** publicó *Indexing from A to Z*, en el que ofrece algunos consejos y consideraciones sobre la indización de la prensa. Dos años más tarde **Barbara P. Semonche** editó el libro *News media libraries: a management handbook*, en cuyo capítulo 19, "Newspaper indexing policies and procedures", escrito por la propia autora, repasa la historia y los tipos de índices de periódicos, ofreciendo pautas generales para indizar un periódico.

A nivel internacional, el tesauro de prensa más destacado por todos los autores es el empleado en la base de datos del diario *The New York times*, que data de 1971 y puede considerarse como el primero de los de información de actualidad. Incorpora más de 700.000 descriptores e incluye una relación alfabética de nombres de personas.

Alan R. Greengrass (1983), analiza el funcionamiento del departamento de indización de este diarioasí como la estructura del tesauro empleado y del índice impreso; describe el proceso de toma de decisiones para la incorporación de nuevos términos y los problemas de la elaboración del tesauro. Además estudia su base de datos de texto íntegro proporcionando ejemplos de resúmenes y de páginas del tesauro.

El tesauro del diario *The New York times* ha sido estudiado por varios autores. **Pastor** (1992) expone sus características principales y la

estructura y relaciones semánticas. También **Caridad** (1980) y **Coll-Vinent** (1982) han publicado trabajos que proporcionan información sobre el mismo.

Milstead (1983) estudia los índices de prensa de diarios norteamericanos y considera que, de entre los publicados, los más detallados son los de *The New York times index*, *The London times index* y *The official Washington post index*. Cada uno de ellos emplea su propio vocabulario controlado, con sus múltiples referencias cruzadas y sus subdivisiones jerárquicas. El *National newspaper index* proporciona sólo una lista de términos, basada en la *Library of Congress subject headings*. En este mismo artículo **Milstead** describe el caso concreto del índice del *The Washington post*, la especialización temática de los indizadores, el procedimiento de indización, el empleo de referencias cruzadas y subencabezamientos, la revisión mensual y anual de los índices, la consideración como esencial del vocabulario controlado, la inadecuación de otros índices y de las listas de encabezamientos a su nivel de especificidad y la estructura del índice.

Martínez-Pestaña (1986) describe 54 bases de datos periodísticas y solamente en tres de ellas se hace referencia a la indización: la del *The New York times*, en la que como ya se ha visto, se especifica la indización con vocabulario controlado, la de *The lexicon herald leader*, que incluye descriptores y la de *Ontap magazine index*.

Pastor (1992) detalla en unas tablas 51 bases de datos de prensa internacionales aparecidas entre 1971 y 1991 y especifica entre otros datos el tipo de lenguaje documental utilizado en el análisis y recuperación de la información. Sólo nueve de ellas, un 17%, disponen de un tesauro de información de prensa para el análisis y recuperación de la información: *The New York times; Gruner & jahr, Politiken dagbladet; St. Louis post dispatch; Documentation française, BIPA; Ringier & Co; Aftonbladet; Edi 7* y *Le monde*. Esta autora expone las principales características de los tesauros de *The New York times* y *Gruner & jahr* debido a que, durante mucho tiempo, han sido los únicos vocabularios adecuadamente estructurados y desarrollados en el área de la documentación periodística.

El tesauro de la base de datos periodística de la editorial alemana *Gruner & jahr* contiene un total de 5.000 descriptores relacionados entre ellos por referencias cruzadas. Se encuentran integrados en 32 grandes facetas o campos temáticos, subdivididos a su vez en varios apartados. Las relaciones que recoge son las de equivalencia y jerarquía.

En el *Segon seminari: l'experiencia multimedia*, **Fuentes** (1994) anota el empleo de descriptores en las bases de datos de *Le monde* y de *Bayard-presse*. También **Coll-Vinent** (1978) estudió el centro de documentación de *Le monde*.

El tesauro de *Le monde diplomatique*[3], pese a denominarse así, es más bien una clasificación o una lista de términos autorizados de temas amplios y no muy numerosos. Incluye también ámbitos geográficos.

González-Quesada (1995), en un estudio sobre la evolución histórica de la documentación periodística, comenta que en los años 70 cuarenta diarios norteamericanos y una docena de europeos –entre los que se hallaban los británicos *The times* y *The guardian*, el francés *Le monde*, el italiano *Corriere de la sera* y el alemán *Frankfurter allgemeine zeitung*– utilizaban índices como guía de los contenidos de la publicación. El mismo autor afirma que la *BBC* cuenta con un sistema de clasificación propio, el *Schedule of subject headings*.

En mayo de 1998 se inició el proyecto de la Unión Europea *Laurin* con el objetivo de desarrollar un modelo genérico para la digitalización de recortes de prensa. Gestionado y coordinado por el *Department of German Language and Literature* de la *Universidad de Innsbruck*, es poseedor de una de las más amplias colecciones de recortes de prensa de literatura y crítica procedentes de prensa de Austria, Alemania y Suiza. Otras siete colecciones de recortes de prensa sobre cultura, política y economía de otros tantos países europeos se unieron al proyecto. Todos esos casos junto a los nuevos que se añaden, forman la base de datos *Laurin*. Los recortes están enlazados con el tesauro multilingüe *Laurin* que está organizado por conceptos, además incluye los topónimos del *Getty thesaurus geographic names* y los de la *Nomenclature of territorial units for statistics.* (**Mühlberger**, 1999; **Calvanese**, et al., 2001).

El diario *Les echos* clasifica las noticias de acuerdo con unas listas predefinidas de materias y ámbitos geográficos. *Los Angeles times* admite en las búsquedas online emplear los descriptores que asignan los documentalistas (**Jiménez** et al., 2000).

El *Web thesaurus compendium*[4] no incluye ningún tesauro especializado en medios de comunicación o especialmente válido para la información de actualidad.

En 1999 el *IPTC* desarrolló la *Subject reference system* para permitir a los proveedores de información (principalmente agencias) acceder a un lenguaje independiente codificado para describir el contenido de las noticias y para facilitar su intercambio. El sistema se explica en la *Information interchange model guideline 3* (*IIMG3*)[5]. Se trata de un sistema clasificatorio jerárquico y consta de 17 niveles temáticos principales que, a su vez, se sub-

"Los estudios sobre lenguajes controlados en medios españoles evidencian una escasez en la elaboración y uso de tesauros debidamente estructurados"

dividen en términos relacionados jerárquicamente. Hay que recordar que en España el grupo *Prisacom* lo adaptó en 2001 para utilizarlo como sistema de indización y de recuperación para su hemeroteca digital.

Los vocabularios controlados en los medios de comunicación españoles

Los estudios sobre el empleo de lenguajes controlados en medios de comunicación españoles evidencian un uso predominante de glosarios y sistemas clasificatorios, además de una escasez en la elaboración y uso de tesauros debidamente estructurados.

Martín-Muñoz y **López-Pavillard** (1995) citan algunos de los tesauros, temáticamente independientes (geográfico, de deportes, de agricultura, de animales, de ciencia y técnica) que se emplean en las bases de datos de documentación audiovisual de *Radiotelevisión Española* (*RTVE*) y que fueron creados y desarrollados por el propio centro de documentación a lo largo de su actividad. Para la base de datos de documentación escrita de *RTVE*, *Basinfa* (*Base de Información de Actualidad*) se consideró imprescindible contar con dos tesauros; uno temático y otro geográfico. Ante la imposibilidad de elaborar un tesauro temático propio se optó por realizar una adaptación del elaborado por la *Unesco*.

En cuanto a *Telecinco*, para el análisis de la información audiovisual se emplea un tesauro que incluye 3.000 descriptores controlados y cerca de 1.000 no descriptores con sus correspondientes reenvíos. **Valle-Gastaminza**, **García-Jiménez** y un equipo de colaboradores (2001) analizan el estudio, la construcción y su puesta en funcionamiento.

Llobet y **Pañella** (1988) describieron el proceso de la elaboración del tesauro para la base de datos de imágenes de *Televisió de Catalunya* (*TV3*). Elaborado con el programa *Mistral*[6], se encuentra estructurado en dos partes: la primera incluye nombres propios de personas y entidades (diccionario de autoridades) con relaciones de sinonimia y notas explicativas. La segunda está dedicada a nombres comunes y geográficos que incluyen relaciones jerárquicas (genéricos y específicos), asociativas (términos relacionados) y de sinonimia. Se emplea también para la descripción de imágenes en la *Unidad de Documentación* de *Radiotelevisión Valenciana*. El *Departamento de documentació* de *Televisió de Catalunya* ha sido analizado en la tesis de **Codina** (1996).

El *Centro de documentación y archivo* de *Euskal Telebista* (*ETB*) utiliza para la recuperación temática de su base de datos de imágenes, un tesauro elaborado por su equipo de documentalistas a partir del de *TV3*. Consta de 12.000 términos en total: 3.500 son descriptores temáticos, 2.500 geográficos y 6.000 identificadores o descriptores onomásticos. Para la base de datos de noticias de prensa utiliza una lista de materias de elaboración propia realizada a partir de diversos listados y tesauros.

Hay que señalar que las bases de datos de imágenes de televisión solamente disponen de representaciones textuales de las imágenes. Las de prensa escrita sin embargo, disponen ya de los textos íntegros y por lo tanto de la posibilidad de utilizar todos los términos de los artículos. Esto significa que además de por la naturaleza de la tipología documental, las bases de datos de imágenes audiovisuales requieren unos vocabularios controlados menos conceptuales y no son totalmente válidos para la recuperación de información de noticias de prensa. Los vocabularios controlados para la descripción de imágenes de televisión están sometidos a un esquema de datos mínimos de representación recomendado internacionalmente, la *minimum data list* de la *FIAT/IFTA*[7], que contiene recomendaciones sobre la representación de contenidos.

Por otra parte, en los centros de documentación de diarios españoles la situación es similar a la de los medios audiovisuales. El servicio de documentación del diario *El correo español* fue analizado en la tesis de **Pastor-Ruiz** (1992). En la tercera parte de este trabajo se describe el lenguaje documental empleado en las bases de datos de este periódico. Se trata de un vocabulario controlado compuesto por dos listas alfabéticas construidas con las palabras clave registradas, durante la fase de indización, en los campos de materias y de descriptores. **Martín** (1994) también describe el centro de documentación y las bases de datos de este diario. El autor hace referencia al mantenimiento de cuatro listas: descriptores, personas, lugares y modos, que comparten las bases gráficas y de prensa. Comenta además que la falta de modelos y la inexistencia de herramientas de control de vocabulario propiciaron un crecimiento exagerado del número de términos, facilitado también por una filosofía de precoordinación muy acusada y redundante.

En la descripción realizada por **Aguado** (1995) del sistema de ar-

chivo y documentación del *Grupo Prensa Española*[8] (diario *ABC* y revista *Blanco y negro*) se mencionan los campos onomásticos, geográficos y temáticos en los que se emplean tablas de validación y tesauros especializados adaptados a las necesidades propias de *Prensa española*. No se dan más detalles de estos tesauros, sin embargo, se menciona en el mismo artículo la elaboración de uno de ellos, así como su aplicación a los índices como objetivo prioritario.

En una descripción del diseño y creación de la base de datos documental del *Grupo Godó*[9], **Salmurri** et al. (2002), mencionan estar desarrollando herramientas (listas de validación y tesauro) para la descripción documental y de contenidos.

En cuanto a *El país*, el diario español de mayor tirada, utiliza una extensa clasificación jerárquica. A partir de su base de datos documental se generaba un índice que se editó en papel semestralmente hasta 1996 y se empleaba para búsquedas retrospectivas en papel. En estos índices puede apreciarse la complejidad de su clasificación jerárquica de tipo precoordinada.

En febrero de 2001 se puso en marcha la versión electrónica bajo la denominación de *elpais.es*, el cual junto con el resto de medios del grupo de comunicación *Prisacom*[10], utilizan un nuevo sistema editorial y documental basado en xml. Además, *Prisacom* optó por el estándar *News Industry Text Format* (*NIFT*) desarrollado por el *International Press Telecommunications Council* (*IPTC*) y el *Subject Reference System*. Esta clasificación temática se amplió y adaptó a las intereses de los medios de la empresa. **Flora Sanz** (2003) comenta que al tratarse de una clasificación muy general y con categorías propias del ámbito norteamericano se vieron obligados a incluir nuevos términos y añadir un tercer nivel de jerarquía para adaptarlo a sus necesidades. Por otra parte, se redujeron drásticamente las entradas utilizadas en la clasificación respecto a las empleadas en la base de datos del diario *El país* en versión papel. En el proceso de migración de los contenidos de las bases de datos en papel (*Hércules*) a la digital (*Pegaso*), se establecieron equivalencias entre las 750.000 entradas normalizadas de grandes temas, materias, topónimos, personas y empresas de *Hércule*s, y las 2.500 categorías[11] de *Pegaso*, basadas en la clasificación del *IPTC*.

"Sólo 2 de 20 bases de datos periodísticas online españolas analizadas por *Pastor* emplean un auténtico tesauro"

De los 11 servicios de documentación de medios de comunicación en España analizados por **Fuentes** y **Conesa** (1994) sólo en dos casos, *El observador* y el de la base *Basinfa* de *RTVE*, se comenta la utilización de tesauro. El diario catalán *El observador*, que dejó de publicarse en 1993, organizaba el archivo manual de prensa de acuerdo con un tesauro que se iba ampliando según las necesidades.

De entre las ponencias presentadas en el *Segon seminari: la documentació als mitjans de comunicació*[12], cerca de 20 trataban el caso concreto de algún medio de comunicación español (prensa, revistas, agencias, radio), y sólo en el ya citado caso de diario *El correo español* se hace referencia al empleo de lenguaje controlado. En las descripciones de las bases de datos *Baratz*, la revista *El Temps* y el archivo de ilustraciones del diario *ABC*, se refleja el empleo de un campo para descriptores, así como otro para los ámbitos geográficos si bien no se menciona el empleo de ningún tesauro. En relación a la base de datos de información de actualidad *Baratz*, **Aquesolo** (1995) señala que cada documento es analizado en un registro independiente al que se le asignan descriptores temáticos, de personalidades y geográficos. El mismo autor comenta otras aplicaciones de consulta de prensa y agencias, además de mencionar *Egunez egun* (*Día a día*), base de datos de prensa histórica del siglo XIX que fue desarrollada sobre el programa *Knosys* y que utiliza el tesauro *Eurovoc*.

La situación de los centros de documentación de prensa diaria en Andalucía ha sido estudiada por **Aquesolo** (1996). En su artículo expone que el sistema de clasificación de los fondos es básico: carpetas clasificadas por temas o según la estructura temática definida por las secciones del diario.

Sólo dos de veinte bases de datos periodísticas online españolas analizadas por **Pastor** (1992a) emplean estrategias de interrogación valiéndose de un auténtico tesauro. Se trata de la herramienta confeccionada por la empresa catalana *Enfony* y de la del Gobierno Vasco, *Inforpo 2*. *Enfony* ha confeccionado para cada base de datos diseñada a medida[13], un tesauro con los términos específicos de la temática que cubren. *Inforpo 2* usa también un tesauro referido a todos los temas de prensa recogidos y cuenta ya con más de 1.500 descriptores.

Por otra parte, las ediciones en cd-rom de los diarios *El mundo*, *La vanguardia* y *El periódico de Cataluña* tampoco disponen de tesauro. En cuanto al caso de *El país* añade una clasificación temática donde se organizan los conceptos, de más generales a más específicos. Tampoco las hemerotecas electrónicas de la mayor parte de los diarios

españoles de información general incorporan ningún tipo de vocabulario controlado para la recuperación de la información por parte de los usuarios. En el caso de la versión digital de *El país* tampoco se emplea ni en la hemeroteca ni en el buscador, pero sin embargo se puede apreciar la utilización de una clasificación jerárquica a partir del servicio denominado "El índice"[14], que incluye onomástico, geográfico, categorías temáticas y otras opciones.

En el entorno de la administración, **Izquierdo-Arroyo** y **Moreno-Fernández** (1992) describieron la estructura del tesauro diseñado para la base de datos de información de actualidad de *La región de Murcia*. Engloba 21 grandes categorías temáticas que incluyen la de onomásticos y topónimos. Se compone de dos partes: sistemática, incluyendo relaciones jerárquicas, partitivas, asociativas y de monoequivalencia semántica, así como notas de aplicación y sistema de facetas (agente, instrumento, modo, acción, materia) y otra alfabética.

Álvaro, **Villagrá** y **Sorli** (1989a) evaluaron 47 tesauros disponibles en lengua española. Para conseguir la relación de los mismos examinaron trece directorios y bibliografías sobre el tema. Ninguno de estos tesauros pertenece a un medio de comunicación.

Lo que si es frecuente es que algunos medios de comunicación españoles adapten tesauros ya elaborados, como el de la *Unesco* y el *Eurovoc*. El primero fue desarrollado por dicha organización para la indización y recuperación de información de la red integrada de documentación de este organismo. Fue publicado por primera vez en 1977 y en 1995 vio la luz una segunda edición. Está formado por unos 8.500 términos distribuidos en cinco secciones correspondientes a las principales áreas de actividad de la *Unesco*: educación, información y comunicación, ciencias sociales, cultura y humanidades, ciencia y tecnología y otra sección general. En ediciones posteriores se añadieron otras áreas de conocimiento: política, derecho y economía. Además, incluye nombres de países.

El *Eurovoc* es el tesauro empleado para la confección de los índices del *Diario oficial de las Comunidades Europeas*. Se caracteriza por su condición multidisciplinar, ya que cubre los temas de interés, tanto en el ámbito de la actividad parlamentaria, como en el de las instituciones comunitarias y en el de los estados miembros. Además añade dos secciones que contienen listas de topónimos y nombres de organizaciones.

Condicionantes para la utilización de tesauros en medios de comunicación

La clasificación e indización de los artículos periodísticos presenta unas características específicas que hacen difícil la aplicación de las clasificaciones alfabéticas de materias que se utilizan habitualmente en las bibliotecas generales, ya que se adaptan mal a la información periodística y la mayoría de los medios que elaboran dossiers o índices han desarrollado las suyas propias (**Fuentes**; **Conesa**, 1994). En este sentido **Díaz** et al. (1986) consideran que el problema de los tesauros existentes en ámbitos no periodísticos reside en ser sistemas destinados a la indización del conocimiento científico. En cambio, en la información de actualidad la materia prima son acontecimientos o reflexiones sobre los mismos.

García-Gutiérrez (1999) señala, refiriéndose a *la Clasificación Decimal Universal* y a lenguajes documentales similares: "si la CDU tuvo y sigue teniendo una gran aceptación, en el mundo biblioteconómico, para el control bibliográfico superficial del ámbito científico, la extrapolación de su filosofía a la organización documental del discurso periodístico sería un error ya que el enciclopedismo aparece como único rasgo común y tan sólo en el nivel extensional. De hecho el enciclopedismo que interesa al mass media queda marcado por intereses e ideología institucionales de los que la CDU carece a pesar de ser un producto del pensamiento positivista. Las restantes características de la actualidad eliminan la posibilidad de adoptar esquemas encorsetados, codificados, y de imposible puesta al día" (p. 356).

Pastor (1992) estima que las dificultades existentes en la indización de prensa escrita se derivan en su mayor parte de las características de la información de actualidad.

Rodríguez-Vela (1992) afirma que los tesauros de información de actualidad escasean dada la universalidad que se exige a su cobertura. En el mismo sentido **García-Gutiérrez** (1999) señala que la extensión enciclopédica de la actualidad crea dificultades a la hora de compilar y estructurar el vocabulario, aunque este problema queda compensado por la superficialidad de su tratamiento.

Además, la prensa nacional y regional adapta sus contenidos al ámbito geográfico al que pertenece. Este factor de localismo, que hace relevantes determinadas noticias en un contexto concreto, dificulta en cierta medida el compartir lenguajes controlados entre medios de comunicación distintos.

Perpinyà (1995) expone las condiciones específicas de las empresas del área de los medios de comunicación que condicionan las características principales de uso de los lenguajes documentales:

– El sector está constituido por empresas privadas y no centros de investigación, en consecuencia no existe una investigación conjunta

para la elaboración y utilización de un lenguaje común. Cada medio utiliza el lenguaje documental que le parece más adecuado.

– Los intereses en cada caso son muy distintos. Los documentos que se recogen y organizan en agencias de publicidad, prensa de información general, prensa deportiva, económica, del corazón, medios audiovisuales, etc, son absolutamente diferentes entre sí y requieren de instrumentos de descripción particulares.

– El tipo de información que procesan, de carácter general (cuestiones muy diversas) y de actualidad (se producen nuevos temas constantemente y otros se quedan obsoletos con rapidez) es otro condicionante esencial, puesto que obliga a tener un lenguaje muy flexible que admita muchos cambios.

– El usuario (básicamente el periodista) necesita obtener la información de forma inmediata. Esto influye en el tiempo dedicado al análisis documental, que tiene que ser mínimo, puesto que la base de datos se tiene que mantener permanentemente actualizada (p. 130).

A estas condiciones se puede añadir que la estructuración del sector en grupos de comunicación, de los que dependen conjuntos de cabeceras periodísticas o emisoras de radio y televisión, no se ha traducido en la existencia de centros de documentación que sirvan al conjunto de medios ligados empresarialmente. Tan sólo el empleo de las bases de datos de *El país* las empresas del grupo *PRISA* se puede citar como modelo de funcionamiento distribuido.

También **Perpinyà** (1995) comenta que la realidad de los sistemas de documentación periodísticos es que cada medio utiliza su propio lenguaje documental, y que generalmente se prefiere la indización a través del lenguaje libre o de listas de materia, debido al ahorro de tiempo en la construcción y utilización. Esto es así en las bases de datos de distribución pública, tanto de texto íntegro (*The guardian*, *The times*, *The sunday times*, *The independent*) como referenciales (*Baratz*, *Documentación de Medios*).

"Hay que plantear un acuerdo entre los medios de comunicación españoles para crear un vocabulario controlado común"

Otra de las razones para rechazar la construcción de un tesauro de información de actualidad es la exigencia de dedicación y tiempo que requiere su elaboración y su mantenimiento. **García-Gutiérrez** (1999) señala que la información y el vocabulario periodístico quedan obsoletos rápidamente por lo que cualquier lenguaje documental necesitaría de un equipo humano que realizara las actualizaciones constantemente y esto supone una asignación presupuestaria que no todos los medios de comunicación pueden permitirse.

En las conclusiones de un estudio sobre la situación de los centros de documentación de medios de comunicación de Madrid se señalan, entre otras, que la inexistencia de tesauros o lenguajes documentales especializados en información de actualidad ha originado que cada centro desarrolle su sistema propio dependiendo de necesidades y recursos muy concretos (**Razquin**, 1993).

Fuentes (1994) concluye, tras quince años de estudio de la documentación periodística, que no hay ninguna uniformidad ni intento de normalización de las operaciones documentales (selección, tratamiento y recuperación) que permita un intercambio de información entre los diferentes medios. Sin embargo, la iniciativa de un formato normalizado para textos en la industria periodística –*NITF* (*News Industry Text Format*) auspiciado por la *Newspaper Association of America* y el *International Press Telecommunications Council*)– no sólo incluye un formato normalizado para el intercambio de noticias, sino una clasificación temática recomendada, la *Subject Reference System.*

Por último, los estudios de consumo de información periodística revelan que buena parte de las necesidades expresadas se refieren a protagonistas de la actualidad y personajes públicos; por tanto, se basan en identificadores. **Castillo-Blasco** (2001) cuantifica en su estudio un 30,66% de peticiones basadas en identificadores, un resultado muy próximo al de **Iturregui** (comunicación personal, servicio de documentación escrita de *Euskal Telebista*).

Conclusiones y propuestas

La escasez de instrumentos de control terminológico en el tratamiento documental de la información de actualidad descansa en una serie de argumentos de naturaleza operativa de cierta solidez. Dichas cuestiones se basan en el elevado coste de la elaboración de tesauros, en la naturaleza dinámica y amplia cobertura temática del lenguaje periodístico, en la celeridad exigida a un espacio informativo de frecuencia tan elevada y en la naturaleza de las demandas planteadas, basadas en el "quién" y el "dónde" de las noticias.

Sin embargo, el ingente volumen de la información de actualidad y su creciente presencia en el entorno de la información accesible a través de internet, al alcance de usuarios no especializados, son argumentos que también aconsejan el control de su contenido. Además, la propia estructuración del trabajo periodístico,

con profesionales encuadrados en secciones o unidades especializadas, aconsejan el empleo de algún instrumento que permita organizar no sólo los documentos, sino también las peticiones. De esta forma se permitiría, por ejemplo, la implantación de sistemas de alerta destinados a los profesionales de los medios y también al público en general.

Desde este punto de vista, no parece descabellado plantear la necesidad de un acuerdo de mínimos entre los medios de comunicación españoles que se tradujera en un vocabulario controlado común para la organización de sus contenidos. Planteado en términos de macrotesauro, cada uno de los medios o grupos de comunicación tendría la posibilidad de extender ese instrumento en función de sus propias necesidades. Tomar como punto de partida los *iptc-subjectcode* de los *newscodes* del *Internacional Press Telecommunications Council* parece un arranque conveniente para abordar un esquema mínimo consensuado de control del lenguaje periodístico.

Agradecimientos:

A **Jesús Andérez** y a **Marta Iturregi** (*Dpto. de documentación y archivo*, *Euskal Telebista*) por los datos proporcionados sobre el archivo de prensa de su institución y la revisión crítica del manuscrito.

Notas

1. Cifras del *Anuario El país* 2005.

2. El *Subject Reference System* está incluido en la *IIMG*: *Information Interchange Model Guideline*. Tiene 3 niveles de jerarquía.

3. *http://www.ina.fr/CP/MondeDiplo/Thesaurus/thesaurus.fr.htm*

4. *http://www.darmstadt.gmd.de/~lutes/thesoecd.html*

5. *http://www.iptc.org/IIM/3.0/specification/IIMV3.PDF*

6. Actualmente *AIRS*.

7. *http://www.fiatifta.org*

8. El *Grupo Correo* se fusionó en 2001 con el *Grupo Prensa Española* y a partir de 2003 pasó a denominarse *Vocento.*

9. Editor de los diarios *La vanguardia*, *Mundo deportivo* y revistas como *Magazine*, *Què fem?*, *Què más?* y el semanario del motor *Escape*.

10. *El país*, *As*, *Cinco días*, *Cadena SER* y *Los 40.com*.

11. El total de términos que incluye esta segunda clasificación es de aproximadamente 10.000 si se tienen en cuenta también los términos geográficos y los onomásticos de personas y empresas.

12. Celebrado en Valencia del 7 al 9 de marzo de 1994.

13. Estas bases de datos recopilan una selección de artículos de prensa sobre un tema concreto de interés para el cliente.

14. Con esta opción se recupera a partir de los índices las noticias de ese día que coinciden con el epígrafe del índice solicitado. En el caso de los onomásticos y geográficos se permite ampliar la búsqueda a otros documentos del retrospectivos del archivo. En todos los casos la visualización del documento final es solo para suscriptores.

Referencias bibliográficas

Anuario El país 2005. Madrid: Ediciones El País, 2005. ISBN 84-95595-12-5.

Aguado-González, Francisco-Javier. "Organización del sistema de archivo y documentación de Prensa Española, S. A. (ABC y Blanco y negro)". En: *Revista general de información y documentación*, 1995, n. 2, pp. 203-208.

Álvaro-Bermejo, Concha; **Villagrá-Rubio, Ángel**; **Sorli-Rojo, Ángela**. "Desarrollo de lenguajes documentales formalizados en lengua española: evaluación de los tesauros disponibles en lengua española". En: *Revista española de documentación científica*, 1989, n. 3, pp. 283-305.

Aquesolo-Vegas, José. "Situación de los servicios de documentación de la prensa diaria de Andalucía". En: *Cuadernos de documentación multimedia,* 1996, n. 5. Consultado en: 20-04-00. *http://www.ucm.es/info/multidoc/multidoc/revista/cuadern5/aquesolo.htm*

Calvanese, Diego; **Catarti, Tiziana**; **Santucci, Giuseppe**. "Laurin: a distributed digital library of newspaper clippings". En: *World wide web*, 2001, n. 4, pp. 5-20.

Caridad-Sebastián, Mercedes. "Estructura general del banco de datos del New York times". En: *Documentación de las ciencias de la información*, 1980, n. 4, pp. 139-155.

Castillo-Blasco, Lourdes; **Doménech-Vidal, Soledad**; **Soler-Monreal, Concepción**; **Amat, Carlos B.** "Demanda de información de actualidad en un servicio de referencia periodística. Análisis descriptivo de 4.160 solicitudes". En: *Revista española de documentación científica*, 2001, v. 24, n. 1, pp. 36-50.

Codina, Lluís. *Teoria de sistemes, teoria de recuperació d'informació i documentació periodística*. Barcelona: Universitat Autònoma de Barcelona, 1996. ISBN 84-490-0725-9; 978-84-490-0725-5.

Coll-Vinent, Roberto. *Teoría y práctica de la documentación*. Barcelona: A.T.E., 1978. ISBN 84-7442-030-X; 978-84-7442-030-2.

Coll-Vinent, Roberto. *Teoría de la teledocumentación*. Barcelona: A.T.E., 1982. ISBN 84-7442-164-0; 978-84-7442-164-4.

Díaz-Arias, Rafael, et al. "La base de información de actualidad (Basinfa) de RTVE, un sistema automatizado de documentación periodística". En: *Segundas jornadas españolas de documentación automatizada*, 1986, pp. 175-183.

Díaz-Nosty, Bernardo. **"**La difusión de la prensa diaria en lengua española". En: *El español en el mundo. Anuario del Instituto Cervantes 1999*. Barcelona: Plaza y Janés, pp. 65-130. Consultado en: 04-11-06. *http://cvc.cervantes.es/obref/anuario/anuario_99/*

Fuentes-Pujol, Maria-Eulàlia. "Evolució de la documentació periodística a Espanya durant els darrers cinc anys i algunes experiències europees". En: *Segon seminari: la documentació als mitjans d'informació. L'experiència multimèdia. Ponències i conclusions*, 1994, pp. 17-28.

Fuentes-Pujol, Maria-Eulàlia; **Conesa, Alicia**. *La documentació periodística. Catalunya, Espanya i altres experiències europees*. Barcelona: Generalitat de Catalunya, 1994.

Galdón-López, Gabriel. *Perfil histórico de la documentación en la prensa de información general 1845-1984.* Pamplona: Eunsa, 1994.

Galdón-López, Gabriel. *El servicio de documentación de prensa: funciones y métodos.* Barcelona: Mitre, 1986.

García-Gutiérrez, Antonio-Luis. "Lenguajes documentales e información de actualidad". En: **García-Gutiérrez, A. L.** (ed.). *Introducción a la documentación informativa y periodística*. Sevilla: Editorial MAD, 1999, pp. 351-372.

García-Morales, Justo. "Normas para la preparación de índices analíticos en centros de documentación y archivos de publicaciones periódicas". En: *Documentación de las ciencias de la información*, 1979, n. 3, pp. 71-111.

Gómez, Bernardo; **Paniagua, Francisco**. "Las ediciones digitales de los diarios españoles. Nacimiento y consolidación de un sector en auge". En: *Razón y palabra*, n. 47, 2005. Consultado en: 04-11-06. *http://www.cem.itesm.mx/dacs/publicaciones/logos/anteriores/n47/gomezpaniagua.html*

González-Quesada, Alfons. "La evolución histórica de la documentación periodística". En: **Fuentes-Pujol, M. E.** (ed.). *Manual de documentación periodística.* Madrid: Síntesis, 1995, pp. 23-39.

Greengrass, Alan R. "Indexing at the New York times information service". En: *Indexing specialized formats and subjects*. Metuchen: Scarecrow Press, 1983, pp. 180-188.

Guallar, Javier. **"**iMente, servicios de información de actualidad en línea". En: *El profesional de la información*, 2006, v. 15, n. 6, pp. 426-435. Consultado en: 04-01-07. *http://eprints.rclis.org/archive/00007856/01/epi06_guallar_imente.pdf*

Guallar, Javier. **"**Mètodes i tècniques de recerca en els articles de documentació periodística a Espanya (1997-2002)". En: *BiD*, 2003, n. 11. Consultado en: 04-01-07. *http://www2.ub.es/bid/consulta_articulos.php?fichero=11gualla.htm*

Hanse, Kathleen A.; **Ward, Jean**. "Information technology changes in large newspaper libraries". En: *Special libraries*, 1991, v. 82, n. 4, pp. 267-273.

Hegg, Judith L. "Small newspaper libraries: the libraries that time (and automation) passed by". En: *Special libraries*, 1991, v. 82, n. 4, pp. 274-281.

International Press Telecommunications Council. *Information Interchange Model Guideline 3*, 1999. Consultado en: 19-07-01.
http://www.iptc.org

International Press Telecommunications Council. *Subject Reference System Guidelines. Versión 3*, 2003. Consultado en: 21-02-04.
http://www.iptc.org

Izquierdo-Arroyo, José-María; **Moreno-Fernández, Luis-Miguel**. "Diseño de una base de datos de prensa controlada por un lenguaje facetado de estructura combinatoria ("thesaurus")". En: *Revista española de documentación científica*, 1992, v. 15, n. 1, pp. 44-63.

Jiménez, Àngels; **González, Alfons**; **Fuentes-Pujol, Maria-Eulàlia**. "Las hemerotecas digitales de la prensa en internet". En: *El profesional de la información*, 2000, v. 9, n. 5, pp. 15-24.

Llobet, Montserrat; **Pañella, Imma**. "Un thesaurus aplicat a un mitjà de comunicació audiovisual. Experiència a TV3". En: *Item*, 1988, n. 2-3, pp. 51-60.

Martín, Mauricio. "El correo. Documentación de prensa. Aprovechando recursos". En: *Segon seminari la documentació als mitjans d'informació. L'experiència multimèdia. Ponències i conclusions*, 1994, pp. 54-62.

Martín-Muñoz, Javier; **López-Pavillard, Jacobo**. "La documentación audiovisual en RTVE". En: *Documentación de las ciencias de la información*, 1995, n. 18, p. 143-171.

Martínez-Pestaña, M. Jesús. "Estructura de los bancos y bases de datos de prensa". En: *Documentación de las ciencias de la información*, 1986, n. 10, pp. 159-212.

Milstead, Jessica L. "Newspaper indexing: The Official Washington Post Index". En: *Indexing specialized formats and subjects*. Metuchen: Scarecrow Press, 1983, pp. 189-204.

Mühlberger, Günter. "Newspaper clippings in a digital world: the Laurin project". En: *Exploit interactive*, 1999, n. 2, 20 July. Consultado en: 13-04-00.
http://www.exploit-lib.org/issue2/laurin/

Pastor-Ruiz, Fátima. *La irrupción de las nuevas tecnologías en la documentación periodística*. Bilbao: Universidad del País Vasco, 1992.

Perpinyà-Morera, Remei. "Los lenguajes documentales". En: **Fuentes-Pujol, M. E.** (ed.). *Manual de documentación periodística*. Madrid: Síntesis, 1995, pp. 111-132.

Razquin, Pedro. "Situación de los centros de documentación en los medios de comunicación de Madrid". En: *Cuadernos de documentación multimedia*, 1993, n. 2. Consultado en: 30-06-00.
http://www.ucm.es/info/multidoc/multidoc/revista/num2/prazquin.html

Rodríguez-Vela, Cristina. "Los lenguajes documentales en las bases de información política y de actualidad". En: *Revista española de documentación científica*, 1992, n. 1, pp. 13-23.

Sanz-Calama, Flora. "La hemeroteca digital de El país". En: *IV jornadas de bibliotecas digitales*, 2003. Consultado en: 21-06-04.
http://imhotep.unizar.es/jbidi/jbidi2003/14_2003.pdf

Semonche, Barbara P. "Newspaper indexing policies and procedures". En: *News media libraries. A management handbook*. Westport: Greenwood Publishing Group, 1993. Consultado en: 19-01-02.
http://metalab.unc.edu/journalism/indexing.html

Whatmore, Geoffrey. *La documentación de la noticia: organización y métodos de trabajo para archivos de referencias de periódicos y agencias.* Pamplona: Universidad de Navarra, 1990.

Wellisch, Hans H. *Indexing from A to Z.* New York: Wilson Company, 1991.

Lourdes Castillo, *Unidad de Documentación de RTVV, Pista de Ademuz s/n, 46100 Burjassot, Valencia.*
macas@uv.es

Alejandro De-la-Cueva*, Departamento de Historia de la Ciencia y Documentación, Facultad de Medicina y Odontología, Av. Blasco Ibáñez 15, 46010 Valencia.*
alejandro.cueva@uv.es