Estudio sobre la indización/etiquetado y los lenguajes documentales en cinco diarios españoles

Study on indexing/tagging and controlled vocabularies in five Spanish newspapers

Antonio García-Jiménez (1) David Rodríguez-Mateos (2) Beatriz Catalina-García (3)

(1) Facultad de Ciencias de la Comunicación, Universidad Rey Juan Carlos, c/ Camino del Molino, s/n, Campus de Fuenlabrada, 28943; antonio.garcia@urjc.es (2) Departamento de Periodismo y Comunicación Audiovisual, Universidad Carlos III de Madrid, c/Madrid,128, 28903 Getafe (Madrid), david.rodriguez@uc3m.es (3) Facultad de Ciencias de la Comunicación, Universidad Rey Juan Carlos; beatriz.catalina@urjc.es

Resumen

Esta investigación parte de la preocupación sobre el modo en el que se organiza el conocimiento en los diarios españoles. Tiene un doble objetivo: por un lado, determinar cómo se desarrolla el análisis, indización y etiquetado de los contenidos periodísticos. El segundo es la delimitación de la existencia, características y aplicación operativa de vocabularios/lenguajes documentales. Se ha remitido un cuestionario a un profesional especializado de los siguientes medios: El País, El Mundo, ABC, La Vanguardia y Heraldo de Aragón. Se observa divergencias en las tareas de indización y etiquetado, la conexión con técnicas SEO, así como la continuación en el empleo de lenguajes documentales, si bien no en todos los casos.

Palabras clave: Documentación informativa. Etiquetado. Indización. Periodismo. Medios. Lenguajes documentales. España.

Abstract

This research comes from a major issue, knowledge organization at Spanish newspapers, with two main aims. First, the way its contents are analysed, indexed and tagged. Second, the use of controlled vocabularies, its presence, main features and implementation is studied. A survey was conducted among some news librarians from El Pais, El Mundo, ABC, La Vanguardia and Heraldo de Aragón. Some indexing and tagging variations were found, and there is a relation to SEO techniques. Controlled vocabularies are also applied, although there are some differences among those media.

Keywords: Media libraries. Tagging. Indexing. Journalism. Media. Controlled vocabularies. Spain.

1. Introducción y objetivos

En el contexto de los cambios vertiginosos y de calado que se están produciendo en el ámbito mediático-periodístico y, por ende, en lo que tiene que ver con la documentación informativa, esta investigación se preocupa por los procedimientos e instrumentos con los que se describen los contenidos periodísticos para su posterior enlazado y/o recuperación al objeto de contribuir a las tareas informativas (García Jiménez, 2016).

Dos son los objetivos específicos. El primero es determinar cómo se aplica en los diarios digitales españoles los procedimientos de análisis, indización y etiquetado de los contenidos periodísticos. El segundo es la delimitación de la existencia, características y aplicación operativa de vocabularios/lenguajes documentales. Conectado a estos objetivos, también son de interés el marco profesional de las funciones documentales de los diarios españoles. Este es un trabajo (1) de naturaleza aplicada y a medio camino entre el ámbito de la Documentación Informativa y el de la Organización del Conocimiento. Aunque entendemos que sería pertinente en este caso, no se planteará ninguna reflexión conceptual sobre el significado, y su correspondiente transformación (lo que daría muestra de un cambio de paradigma) de algunos términos. El ejemplo más claro lo tenemos en las interacciones, solapamientos y fugas existentes entre "análisis documental" y "etiquetado". Tampoco se hará una valoración crítica de las consecuencias de naturaleza epistemológica e ideológica de las actividades convertidas aquí en objeto de estudio.

2. Estado de la cuestión

2.1. Situación del periodismo

Es comúnmente aceptado que el periodismo vive un momento convulso. Y esto se debe a la casi absoluta digitalización de sus diferentes procesos. Un dominio en continua interacción con los

medios o redes sociales (Fondevila, 2017), donde los procedimientos y técnicas SEO tienen un peso relevante (Codina et al., 2017), lo que influye en la actividad periodística, y por ende, documental. Lo que también se trasluce en los términos-etiquetas y lenguajes con los que se organiza el conocimiento.

Otra de las características del periodismo moderno tiene que ver con los crecientes procesos de automatización y "algoritmización" (Saad y Bertocchi, 2012; Jung et al., 2017). De hecho, es cada vez mayor la aportación del Big Data (Stone, 2014; Renó y Renó, 2017), por ejemplo, en términos de análisis automatizado de titulares, adaptación a las audiencias, segmentación y personalización de contenidos, seguimiento de noticias en tiempo real o generación de relatos. Unido a este fenómeno, se añade la mayor presencia de una corriente periodística que tiene una gran conexión con la Documentación, concretamente el periodismo de datos (López-García et al., 2016).

2.2. Documentación Informativa

El ámbito primario donde se ubica este trabajo es el de la Documentación Informativa. En este sentido, son varias las tendencias que se advierten en la literatura especializada, con datos complementarios y en ocasiones, contradictorios. En el trabajo centrado en Andalucía de Meléndez-Malavé y Hirschfeld-Suárez (2016), se observa que no todos los medios gozan de un centro de documentación. Se podría hablar incluso de una lenta, aunque progresiva, desaparición de los mismos. Asimismo, es destacable el descenso del número de documentalistas que trabajan en los medios, lo que podría vincularse a la doble crisis tanto económica como la relativa al propio del negocio de la prensa. De acuerdo con este estudio, la información más requerida por los periodistas se basa en noticias y fotos de archivo, especialmente vinculadas al ámbito local.

Por su parte, en la investigación de Marcos-Recio y Edo (2015) se afirma que un 33% de los periodistas españoles ven el servicio de documentación como imprescindible y el 50% como necesario. Al mismo tiempo, se detectan porcentajes relevantes de periodistas que usan de forman intensiva el centro de documentación, si bien son los recursos electrónicos los que se constituyen como las fuentes con un mayor índice de consulta. En este sentido, el estudio de Pintado (2013) confirma que habitualmente los periodistas solo se documentan cuando tienen dudas sobre un determinado aspecto de la información que están elaborando, y no aluden a esta actividad como un modo genérico de comportamiento. Destaca el hecho de que el uso de fuentes, que en cualquier caso son mayoritariamente institucionales y electrónicas, se diferencia por el género del texto y por la temática tratada.

2.3. Etiquetado y lenguaje

Los continuos cambios que tienen lugar en los modos de producción y difusión de la información periodística tienen sus consecuencias en la organización del conocimiento. Influyen en la asignación terminológica correspondiente y en los instrumentos que la gestionan, y que vinculamos al subcampo teórico que se ocupa de herramientas como las clasificaciones, los tesauros, las folksonomías y, más tangencialmente, las ontologías.

2.3.1. Vocabularios y lenguajes documentales

Uno de los primeros aspectos que se tendrá en cuenta en esta investigación es si, por un lado, el medio analizado emplea un lenguaje/vocabulario, así como su naturaleza y sus características. En primer lugar, nos parece útil, de acuerdo con Soler y Gil (2010), y conforme a unas categorías válidas para nuestra investigación, la determinación del nivel de complejidad de los lenguajes a los que aquí hacemos referencia: de menor a mayor, las folksonomías, las taxonomías, los tesauros y las ontologías.

En este sentido, resulta relevante la defensa por parte de autores como Yedid (2013) de las folksonomías. Desde su punto de vista, no existen grandes diferencias en relación a los términos-etiquetas de otros vocabularios controlados, a pesar del mayor grado de inconsistencias existentes en el uso del plural-singular. Asimismo, desde su perspectiva, los términos en uso están más relacionados con el vocabulario que se emplea realmente por parte de los usuarios, lo que sugeriría una mejora en la recuperación de información, puesto que se deriva de un proceso de selección-elaboración más intuitivo. Junto al menor costo que supone, hay quien aduce, a su vez, una mayor democratización en el sentido de uso de un conocimiento compartido y colaborativo, frente a una visión más jerárquica. En cualquier caso, el debate sigue abierto (du Preez, 2015).

Inicialmente, desechamos el uso real y operativo de instrumentos de gran complejidad. Es el caso de las ontologías. Frente a estas, y en un plano puramente documental, diversos autores (García Jiménez, 2004; Codina y Pedraza, 2011; Mendes et al., 2015) apuntan la validez de los tesauros, demostrada en una trayectoria ya larga. Precisamente, desde una mirada vinculada a las clasificaciones documentales, Szostak (2014) hace hincapié en las dificultades que encuentran las ontologías para su aplicación documental con-

creta: por un lado, por la pobre concepción terminológica subyacente, y por otro, sus diferentes planteamientos y bases, lo que complica en exceso sus opciones de interoperabilidad.

Además, en gran medida gracias a SKOS (Simple Knowledge Organization System), el tesauro ya goza de opciones claras de formalización, aunque condicione su estructura relacional en el contexto de la web semántica (Sánchez, Colmenero y Moreiro, 2012). De hecho, tal y como señalan Pastor et al. (2012), al analizar el funcionamiento de diversos lenguajes controlados que incorporan SKOS, los tesauros, junto a las clasificaciones, obtienen buenos resultados, salvo en lo que se refiere a la interoperabilidad y la interconexión entre ellos mismos.

Si bien hasta el momento no se han descrito procesos de automatización en los instrumentos empleados en prensa, en cualquier caso, en una revisión de este tipo, se tendrá que tomar en consideración los gestores de tesauros. Precisamente, Martínez y Alvite (2014) aplican un método para la evaluación de este tipo de instrumentos. Algunos de los puntos de análisis son: el propósito, la disposición en términos de integridad y coherencia, así como los elementos asociados a la interoperabilidad, integración y a la compatibilidad con los estándares propios de la web semántica (RDF/SKOS).

2.3.2. Análisis, etiquetado y formalización

Son escasas las investigaciones centradas en el análisis-etiquetado humano. En este sentido, podemos acudir a Pérez et al. (2014) cuando hace mención al etiquetado social, que hace posible que los términos sean más concretos y están más pegados a los intereses de los usuarios. Al mismo tiempo, se adapta más a los ámbitos colaborativos. Incluso se vinculan con comunidades de usuarios concretas. Por su parte, sí encontramos algunos textos relevantes sobre la formalización del análisis y marcado semántico. Así, la investigación de Pastor (2013) se centra en la situación y aplicabilidad de Schema.org, los datos estructurados en Xhtml, los microdatos, microformatos o el RDFa.

2.4. Etiquetado, indización y lenguajes en el periodismo

En primer lugar, acudimos al trabajo realizado por Rubio Lacoba (2012) que analiza lo que podríamos definir como folksonomía controlada del diario *El País*. En efecto, se trata de una elaboración que tiene como fuente fundamental la indexación social realizada sobre los productos periodísticos. Una tarea iniciada por los periodistas cuando incorporan las etiquetas, que son filtradas posteriormente por los documentalistas con vistas a evitar la sinonimia y la polisemia, y para mejorar las relaciones entre las citadas etiquetas. En cualquier caso, son procesos que se realizan tomando en consideración las técnicas SEO.

Se trata de un lenguaje que se ha alimentado de las bases de datos propias y de tesauros especializados. Su estructura radica en diferentes áreas: temas, personajes, organizaciones, lugares y eventos. También hay que contar con la conexión automatizada de las noticias con las etiquetas o temas. Finalmente, este lenguaje está vinculado a un editor que gestiona: la solicitud e incorporación de etiquetas, las acciones de desambiguación alrededor de fechas, cargos y siglas, así como el establecimiento de etiquetas, que no se situarán en el lenguaje principal pero que pertenecerán a lo que denominan "conceptos editoriales".

En un plano más teórico, destacan varias líneas de trabajo vinculadas con el periodismo. Así, Baños (2013) aborda las posibilidades que ofrecen las noticias de divulgación científica en lo que se refiere a la actualización de los tesauros. Aunque muestran algunas limitaciones, sí se sugiere su capacidad para colaborar con la actualización de estas herramientas, fundamentalmente como "yacimiento" de términos. Dentro también del ámbito periodístico, el estudio de Baños et al. (2015) revisa otro espacio relevante: los modelos de representación de los metadatos. Fundamentalmente destaca la ausencia de "uniformidad" lo que dificulta que se aplique realmente la interoperabilidad. En este trabajo se advierte que, frente a la mayor dedicación en la literatura especializada sobre NewsML y NITF, en el campo profesional predomina en los códigos fuente schema.org junto a los esquemas, Twitter Cards y Open Graph Protocol (en Facebook).

Con un planteamiento más especulativo, aunque con orientación operativa mantenida en el tiempo, García Gutiérrez (2011) propone las exomemorias. Con una base crítica, transcultural y participativa, intenta superar la investigación descriptiva y desprovista de una perspectiva positiva y estereotipada. A su vez, (García Gutiérrez, 2014) desarrolla un modelo de análisis de textos periodísticos, para su posterior recuperación, que se centra en los siguientes puntos: acción (a través de una macroproposición global), sujetos, objetos, elementos asociativos, situación, causa, finalidad, consecuencia, modo, instrumento, lugar y tiempo. No deja de sorprender que sea de los pocos investigadores preocupados por los "matices" ideológicos y el "impacto" en la sociedad de este tipo de tareas, más si cabe cuando nos referimos al dominio periodístico.

Finalmente, traemos a colación un estudio (Søbak y Pharo, 2017) que, aunque centrado en televisión, aborda el etiquetado-indización en una cadena, en este caso la Norwegian Broadcasting Corporation (NRK). En su dinámica de trabajo, y aunque no se cuenta con un vocabulario controlado, los indizadores tienen acceso a las etiquetas que han sido previamente usadas. Uno de los hallazgos más relevantes es el referido a la influencia de la indización descentralizada, tanto en los porcentajes de temas y nombres de personas o instituciones representados como en la variación de los niveles de exhaustividad, o la presencia de un bajo nivel de cobertura, en gran medida a partir de las diferencias de prácticas profesionales (con diferentes niveles de formación) de etiquetado encontradas.

3. Metodología

Esta investigación presenta un análisis preliminar, en el marco de la actividad documental en algunos de los principales medios periodísticos españoles en versión digital (tanto si tienen además versión impresa como si no), el uso de vocabularios/lenguajes y del análisis/descripción documental o etiquetado en relación a las noticias que publican. Por una parte, se pretende delimitar si el etiquetado es realizado con fines documentales, esto es, si ha sido empleado para describir cada contenido periodístico (noticia, reportaje, artículo de opinión, etc.) con el fin de facilitar la relación y asociación entre distintos contenidos, así como su recuperación en un futuro. Ello supone previamente averiguar si en estos medios existen profesionales de la documentación y, a continuación, saber si estos profesionales participan en la definición de los textos empleados como etiquetas. Es también posible que ese etiquetado sea realizado por otro tipo de profesionales, ya sean los mismos periodistas que elaboran los contenidos, o bien especialistas en la creación de términos que sirvan para destacar la presencia de estos medios, a corto plazo, en otros recursos de Internet, tanto en buscadores como en distintas redes sociales, mediante técnicas como el search engine optimization, o SEO (Carroll, 2010).

Se ha pretendido revisar, de acuerdo con la información ofrecida por estos medios, quién y cómo define esta descripción mediante etiquetas: si se plantea la generación de esos descriptores basados en lenguajes documentales, y de qué tipos, o bien, si se realiza mediante lenguaje libre; cómo se evalúa la calidad de ese etiquetad y su efectividad; y además, qué profesionales generan, de hecho, esos contenidos.

En relación a los vocabularios/lenguajes documentales empleados, se analizan los siguientes aspectos: su existencia, su nivel de control terminológico, el número de términos existente, su estructura, las fuentes que lo alimentan, las interacciones y conexiones con otros instrumentos, su base tecnológica, los modos de integración en el plano documental y en la redacción, su impacto en la recuperación de información, los profesionales que se ocupan de ellos, los proceso de actualización, y el papel de los protocolos de evaluación y determinación de su calidad.

Para cumplir con los objetivos previamente marcados, se ha empleado una doble estrategia: inicialmente, se ha tratado de localizar y contactar, para cada medio analizado, a algún miembro del mismo que pudiera estar vinculado o participar en el etiquetado de sus contenidos. El contacto se ha realizado vía telefónica, a través del correo electrónico, o mediante redes sociales (como Linkedin). Se ha pretendido obtener a algún profesional de la documentación dentro del medio o, secundariamente, a algún otro posible implicado en la generación de etiquetas (sea un responsable de la redacción, un experto en SEO, etc.). En aquellos casos en los que se ha conseguido ese contacto, se ha realizado una encuesta detallada, mediante Google Docs, empleando una combinación de preguntas tanto cerradas como abiertas, con el fin de amplificar las respuestas. La extracción de datos de ese cuestionario se presentan en el apartado de resultados.

La base de trabajo ha estado compuesta por los 15 periódicos españoles de mayor audiencia, medida en el número de visitas a su versión digital, de acuerdo con los baremos oficiales aceptados por los propios medios en los meses de julio, agosto y septiembre de 2017, de acuerdo con la empresa Comscore. Dado que esta compañía solo hace públicos sus datos a sus asociados, se han consultado las referencias publicadas en uno de los periódicos analizados (OK Diario 2017a, 2017b, 2017c). Se han tomado datos de tres meses, dado que algunos de los medios no aparecían en todas las mediciones, por lo que se han dado por válidos cuando aparecían al menos en dos de los meses analizados. Durante la investigación, no ha sido posible contar con respuestas concretas de muchos de estos medios. Sí que han respondido a esta cuestión cuatro de los diarios con mayor difusión digital: *El Mundo, El País, La Vanguardia* y *ABC*, en los cuales existe, aún, un servicio de documentación como tal. En cuanto al quinto medio que respondió, 20 Minutos, no consta que mantenga ya dicho servicio, aunque sí lo hace el que es ahora el diario matriz de su grupo, *Heraldo de Aragón*, a quien se le ha aplicado también la encuesta.

	El País	*ABC*	*El Mundo*	*La Vanguardia*	*Heraldo de Aragón*
Utilización	Sí	Sí	No	No	Sí
Control de términos	Sí	Sí			Sí
Estructura	Colaburario	Tesauro		Taxonomía	Tesauro
Fuentes	Antiguas del grupo Prisa. Estándar IPTC. Tesauros especializados: CSIC y CINDOC.	Clasificación propia. Tesauros externos referenciales		www.classora.com	Tesauros UNESCO
Nº términos	Más de 6000	Más de 6000		Menos de 1000	Más de 6000
Impacto en la recuperación 1-5	5	4		3	4
Fórmulas para desambiguar	Notas aclaratorias Nombres normalizados (URI)	Relaciones tipo		No se emplean	Notas aclaratorias
Relaciones semánticas entre términos	Específicos y generales	Específicos y generales		Específicos y generales	No hay relaciones
Tipo de categorización	Personas Lugares Organizaciones Eventos Temas	Por asignación Basado en metadatos		Ciudades Ubicaciones Equipos Nombres propios, etc	Facetada
Profesionales encargados	Documentalistas	Documentalistas		Documentalistas y programadores informáticos	Documentalistas y programadores informáticos
Frecuencia de actualización	Menos de un mes	Menos de un mes		Menos de un mes	Mensualmente
Conexión con otros lenguajes	RR. internas con Enciclomedia (PRISA Radio) Futuro:Linked Data	Etiquetado de contenidos de página web		No hay conexión	No hay conexión
Modelo-Base para el lenguaje	Procedente del Colabulario con metas DC y Schema	UNE 50106:1900		No hay modelo	No hay modelo
Evaluación funcionamiento y calidad	Humana Automática	Humana		No se evalúa	No se evalúa
Coste		10% capital humano		No disponible	Actualmente mínimo
Integración en actividad cotidiana	Publicación diaria de contenidos. Arquitectura de información Propagación en RR.SS. Estratégica para negocio online.	Catalogación de registros		Creación de temas para mejorar el enlazado SEO	Imprescindible para búsquedas de envergadura
Elementos para marcado semántico	Estrategia aplicada por equipo SEO	No registrado	Palabras claves en el resumen	No registrado	Resumen documental y descriptores (papel). Taxonomía SEO (web)
Contenidos etiquetados	Los publicados por el medio en su web	Los publicados por el medio en su web	Contenidos en papel	Los publicados por el medio en su web	Los publicados por el medio en su web
Etiquetas en medios y redes sociales	No emplean redes	No emplean redes	No emplean redes	No emplean redes	No emplean redes
Responsables de etiquetado	Periodistas Especialistas en medios sociales	Periodistas Documentalistas	Documentalistas	Herramienta de Classora	Periodistas Documentalistas
Relación etiquetado-lenguaje documental	El Colabulario genera una página optimizada por equipo SEO. El periodista tiene posibilidad de enlazar contenidos relacionados- Algoritmos para establecer potenciales intereses del lector	El etiquetado se desarrolla a partir del lenguaje	Ninguna	Ninguna	Ninguna
Modo de evaluar el etiquetado	Estadísticas internas Herramientas propias del equipo SEO	Indicadores de posicionamiento web	No	Prevalecen términos genéricos para que no dupliquen contenidos clasificados y carezcan de valor SEO	No

Tabla I. Resultados de la encuesta

4. Resultados

4.1. Los centros de documentación

En todos los diarios analizados se mantiene la figura del documentalista que, según los casos, desempeña su labor junto a profesionales de otros ámbitos; principalmente programadores y/o informáticos que están presentes en tres de los centros; además de un periodista, un diseñador gráfico, un especialista en medios sociales y, finalmente, un responsable de métrica web.

En la mayoría de los casos, no son más de tres personas las encargadas de este cometido; solo en uno de los centros (*El País*) trabajan más de 5 profesionales que, atendiendo al volumen de información generada y recibida por los periódicos analizados, suponen una cantidad adecuada para la gestión documental en un medio de comunicación. La duda que surge en este punto es la relación entre número de profesionales y el desempeño de las funciones correspondientes.

La principal tarea encomendada se relaciona, en todos los casos, con la gestión y actualización del fondo documental. Y en cuatro de los medios el trabajo documental también se orienta a la búsqueda de información y el etiquetado de documentos. El asesoramiento en la recuperación de información a periodistas, el diseño y mantenimiento de lenguajes y la elaboración de productos documentales y/o periodísticos son tareas que se desempeñan en tres de estos centros. Solo uno de ellos dedica parte de su trabajo a la digitalización y conservación del fondo antiguo. Este último resultado sugiere que, o bien ya tienen todo el material analógico reproducido en los nuevos formatos, o bien que estos centros carecen de un archivo antiguo porque su creación se produjo en la era digital.

Cuatro de los cinco periódicos encuestados afirman emplear una base de datos para el desarrollo de su trabajo cotidiano. Dos de ellos (*El País* y *El Heraldo de Aragón*) mantienen un sistema de gestión propio, aunque el utilizado por el periódico nacional es compartido con todos los medios que conforman el *Grupo PRISA*. En *ABC* y *El Mundo* se utiliza el archivo digital multimedia *Quay,* y en este último diario se combina con otras bases como la plataforma diseñada por la *Agencia EFE (EFE Data*), *Informa* y la estadounidense *Factiva.*

4.2. Vocabularios/lenguajes documentales

Tanto *El Mundo* como *La Vanguardia* reconocen en el cuestionario no disponer de un tipo de lenguaje documental establecido. No obstante, por las respuestas recibidas de este último medio, entendemos que emplean un lenguaje que no es propio, o bien que no tiene un control documental real, y que puede basarse en una mera clasificación temática que reutilizan en algunas ocasiones. En principio, se han aceptado las respuestas en este sentido, si bien este punto necesitaría de una ulterior revisión. El resto, *El País* y *ABC*, en el ámbito nacional, y *El Heraldo de Aragón* (aquí y en adelante), en el regional, aplican en todos ellos un lenguaje controlado.

Se utilizan diferentes tipos de lenguajes: taxonomía en *La Vanguardia,* tesauros aplicados por *ABC* y *Heraldo de Aragón*, y en *El País* se maneja lo que en la actualidad denominan Colabulario, similar a los tesauros pero con una adaptación mayor a los entornos digitales.

La diversidad que caracteriza a los periódicos estudiados se manifiesta también en las fuentes que utilizan para el enriquecimiento de su lenguaje documental. *El País* y *ABC* recurren a una mayor diferenciación de procedencias, mientras que los diarios de ámbito regional se ciñen respectivamente a una sola fuente: el catalán se sustenta en la base digital y abierta de Classora (http://www.classora.com/), y el aragonés se apoya en tesauros de la UNESCO. Igualmente se encuentran diferencias en función del ámbito que cubren los periódicos en lo que respecta a los profesionales encargados de estos lenguajes: si bien en todos los diarios esta labor es ejercida por documentalistas, en el caso de *La Vanguardia* y *El Heraldo de Aragón* se incorporan también a este cometido programadores informáticos.

Se observan divergencias también entre las distintas coberturas geográficas: en los regionales no se establece conexión con otros lenguajes, no existe un modelo que sirva como base y tampoco se aplican protocolos de evaluación y calidad del leguaje empleado. En el caso de *El País* se recurre nuevamente al Colabulario para establecer un modelo del lenguaje, y para las conexiones lingüísticas se desarrollan relaciones internas con la Enciclomedia (procedente de PRISA Radio). *ABC* conecta con el etiquetado de contenidos en la página web, y el modelo-base que utiliza es el establecido por la Asociación Española de Normalización, UNE. En ambos periódicos de ámbito nacional son los responsables del centro quienes evalúan y supervisan el funcionamiento de los lenguajes, aunque en *El País* se apoyan también en sistemas automáticos.

Otras cuestiones puntuales que marcan la diferencia entre unos periódicos y otros se derivan del número de términos empleados. Todos manejan una cantidad superior a los 6.000 a excepción de *La Vanguardia,* cuyo número no alcanza el millar (y tampoco este periódico aplica fórmulas para eliminar ambigüedades terminológicas). Por su parte,

El Heraldo de Aragón no marca relaciones semánticas entre términos, mientras que los otros tres establecen tanto conexiones específicas como generales. En todos ellos se pretende organizar los términos alrededor de categorías, aunque no coincidan entre sí: por ejemplo, "personas", "ciudades", etc. Igualmente, es el periódico aragonés el que registra diferencias en cuanto a la actualización del lenguaje, que desarrolla este cometido una vez al mes, mientras que el resto de periódicos lo hace con una mayor frecuencia.

Por otra parte, los profesionales que cumplimentaron el cuestionario perciben un importante impacto de los vocabularios/lenguajes en la recuperación de información del periódico. En una escala de grado creciente entre el 1 y 5, el único que presenta una visión más modesta (3) es *La Vanguardia*; mientras que en *El Heraldo de Aragón* y *ABC* entienden que el impacto es notable (4) y muy grande (5) en *El País*. Estos índices se relacionan también con la integración en la actividad cotidiana de cada periódico analizado. Al respecto, el centro de *Prisa* utiliza su lenguaje documental para múltiples cometidos: la publicación diaria de contenidos, el diseño de la arquitectura de la información, la difusión en redes digitales e incluso como un elemento más de la estrategia global para el negocio online. En el resto de medios se reconoce también un cierto nivel de integración, aunque en menor medida y en aspectos más puntuales: en *ABC* para la catalogación de registros; como vía para crear temas que mejoren el enlazado SEO dentro de *La Vanguardia*; y, finalmente, la incorporación de lenguajes es clave en *El Heraldo de Aragón* para búsquedas con cierta envergadura.

4.3. Procedimientos de análisis y etiquetado

La única coincidencia detectada en todos los periódicos es la ausencia generalizada de control documental sobre el etiquetado de sus contenidos presentados en medios y redes sociales. En el resto de ítems establecidos se observan diferencias entre todos los centros analizados. Igualmente, en función de la cobertura geográfica, no existe una relación significativa en ningún aspecto. Como punto preliminar, el marcado semántico de *El País* y *El Heraldo de Aragón* es aplicado en el entorno de la estrategia del equipo SEO, aunque en el formato papel, y el periódico aragonés coincide con *El Mundo* al utilizar los descriptores incluidos en los resúmenes. Este diario se diferencia del resto en la selección de contenidos que son etiquetados, correspondientes a los incluidos en la versión impresa, mientras que los otros diarios se centran en los publicados en su web.

Los responsables de este cometido son documentalistas en el caso de *El Mundo, ABC* y *El Heraldo de Aragón;* aunque en estos dos últimos también participan periodistas. Precisamente son estos profesionales, junto a los especialistas en medios sociales, los encargados de esta tarea en *El País.* En *La Vanguardia,* esta responsabilidad se delega en la herramienta aportada por Classora.

Dos de los diarios nacionales, *ABC* y especialmente *El País,* establecen relaciones entre el marcado semántico y el lenguaje documental. Este último aplica las vinculaciones mediante tres sistemas: a través del Colabulario que se proyecta en la web con la optimización del SEO, los enlaces aportados por los periodistas y, finalmente, los algoritmos que se aplican en función de los intereses detectados en los lectores. En *ABC* el etiquetado se desarrolla a partir del lenguaje documental; mientras que en los otros tres periódicos no se establece relación alguna.

Finalmente, la evaluación del etiquetado no se desarrolla en todos los periódicos y los que lo hacen, lo establecen de forma diferente. A partir de las estadísticas internas y de herramientas SEO es la línea que *El País* aborda para esta cuestión*; ABC* consulta indicadores de posicionamiento web, mientras que *La Vanguardia* se rige por la prevalencia de términos generales con el fin de evitar duplicaciones en contenidos ya clasificados. Tanto *El Mundo* como *El Heraldo de Aragón* manifiestan no aplicar ningún sistema de evaluación.

5. Discusión y conclusiones

En buena parte de los periódicos analizados, se mantiene la presencia documental sobre los contenidos. Los responsables documentales continúan siendo conscientes de su importancia para la recuperación de información, y se mantiene el uso de lenguajes documentales. En este sentido, a la diversidad tipológica detectada se añade la que se observa en lo tocante a sus fuentes de alimentación. Nos encontramos ante instrumentos que suelen tener un número alto de términos, y que suelen contar con relaciones semánticas, aunque no en todas las opciones estudiadas. Cuentan con diferenciados objetivos y en su elaboración y mantenimiento intervienen documentalistas y, en ocasiones, otros profesionales. Con una actualización, por regla general, en un tiempo no superior al mes, no emplean ningún protocolo concreto de evaluación ni de determinación de su calidad. En cualquier caso, sí ofrecen un claro impacto en la recuperación de información.

No obstante, esa actividad documental parece tener un reflejo cada vez menos perceptible en el etiquetado real de los contenidos en la red. Las

etiquetas son aplicadas por documentalistas en la versión web, pero no son empleadas en las versiones de esos contenidos en medios o redes sociales. En tres de los cinco casos analizados, ese etiquetado no está realmente relacionado con un lenguaje documental, tampoco se corresponde con un marcado semántico en detalle de los contenidos y, cuando ocurre, está enfocado más bien a la actividad del medio en SEO (es decir, a mejorar su difusión a corto plazo en la red). Principalmente, se emplea para una descripción de personas, lugares y de ciertos temas. Tampoco se emplean, en su mayoría, estrategias de evaluación y de control, desde el punto de vista documental sobre el etiquetado realizado.

Otro hallazgo de esta investigación es que en solo dos de los cinco medios, la actividad documental tiene algún influjo real en la actividad digital del medio. Una de las principales motivaciones de estos resultados radica en el escaso número de profesionales, con responsabilidades sobre la documentación, existentes en los medios analizados.

Como conclusión general, se puede afirmar que los servicios de documentación que permanecen en los periódicos analizados están centrados, fundamentalmente, en tareas de recuperación de información que en la actividad digital del medio, y apenas participan en el etiquetado de contenidos, al menos, de cara al exterior. No ha sido posible encontrar referencias sobre servicios de documentación en periódicos nativos digitales, aunque este extremo requiere de una investigación más en profundidad: por un lado, para confirmar si esta ausencia es real y masiva; y por otro, para confirmar hasta qué punto la descripción y recuperación de información en estos medios, tanto interna (en la generación de contenidos) como externamente (en el etiquetado de los contenidos que se añade a las noticias para sus lectores), se ha visto afectada por esta situación.

Finalmente, esta investigación presenta diferentes limitaciones. En primer lugar, los datos expuestos proceden de 5 medios de comunicación que coinciden en una misma trayectoria: disponen de edición digital y de edición impresa, por lo que su servicio documental ya existía antes de su presencia en la red y se ha adaptado, en lo posible, al entorno digital. En segundo lugar, se ha aplicado un cuestionario, que es una técnica útil para los objetivos propuestos, si bien para una profundización en algunos aspectos sería útil acudir a otras aproximaciones metodológicas para completar algunos de los resultados y para despejar algunas dudas. A las dificultades aquí resaltadas conviene añadir que nos situamos en un momento de evidente cambio en la realidad documental.

Agradecimientos

Los autores agradecen la colaboración de los diarios analizados y especialmente a los profesionales que han colaborado.

Apéndice: Cuestionario empleado

Datos generales

Medio:

Nombre de la persona que responde este cuestionario:

Puesto que ocupa en el medio:

Sección 1: Preguntas genéricas para todos los medios

1. Si su medio dispone de versiones en papel y digital, ¿las redacciones digital e impresa están integrada? Sí / No
2. ¿Qué perfiles profesionales de su medio están vinculados a la actividad documental? Periodistas / Documentalistas / Programadores/informáticos / Gestores de la información digital / Especialistas en medios sociales (SEO, SEM, etc.) / Gestores de medios/redes sociales / Diseñadores gráficos/responsables de visualización / Responsables de métrica web / Otra…
3. ¿Su medio dispone de un servicio de documentación? Sí / No
4. En caso de respuesta afirmativa a la anterior cuestión, ¿cuáles son sus tareas? Asesoramiento en la recuperación de información a periodistas / Gestión y actualización del fondo documental / Elaboración y mantenimiento de lenguajes / vocabularios documentales / Etiquetado de documentos / Búsqueda de información / Elaboración de productos documentales / Elaboración de productos periodísticos / Otra…
5. ¿Cuántos documentalistas componen el servicio de documentación, si existe? 1 persona / Entre 1 y 3 personas / Entre 3 y 5 personas / Más de 5
6. ¿Qué bases de datos se emplean en vuestro trabajo cotidiano? (Respuesta abierta)
7. ¿Qué tipo de fondo documental se genera y se mantiene en vuestro medio? (Respuesta abierta)

Sección 2: Cuestiones sobre el lenguaje / vocabulario documental

8. ¿Su servicio documental dispone de algún tipo de lenguaje documental? Sí / No

(El resto de esta sección, preguntas 9 a 26, se responderán solo en caso afirmativo a la pregunta anterior)

9. ¿De qué tipo de lenguaje se trata, teniendo en cuenta el control de los términos empleados? Lenguaje libre / Lenguaje controlado
10. ¿De qué tipo es ese lenguaje documental, teniendo en cuenta su estructura? Taxonomía / lenguaje de clasificación / Folksonomía / Tesauro / Ontología / Otra…
11. ¿Cuáles han sido las fuentes iniciales para la construcción del lenguaje documental? (Respuesta abierta)
12. ¿Cuál es el número de términos (aproximado) del lenguaje documental? Por debajo de 1000 términos / Entre 1001 y 3000 términos / Entre 3001 y 6000 términos / Más de 6000 términos
13. ¿En qué medida el lenguaje documental tiene impacto en la recuperación de información de los contenidos del medio? (1 = Nada / 5 = En gran medida)
14. ¿Qué fórmulas se emplean para la desambiguación de términos? No se emplean / Mediante notas aclaratorias / Mediante relaciones del tipo "usado por/use" / Otra…
15. El lenguaje documental ¿plantea relaciones semánticas entre los términos? Sí / No
16. En caso afirmativo a la pregunta anterior, ¿qué relaciones semánticas se emplean? Términos relacionados / Términos específicos/generales / Otra…
17. Ese lenguaje ¿tiene formas de categorización? Sí / No
18. En caso afirmativo, ¿qué tipo de categorización se emplea? (Pregunta abierta)
19. ¿Qué profesionales son los encargados de la elaboración y mantenimiento del lenguaje documental? Periodistas /

Documentalistas / Programadores/informáticos / Gestores de la información digital / Especialistas en medios sociales (SEO,SEM, etc.) / Gestores de medios/redes sociales / Diseñadores gráficos/Responsables de visualización / Responsable de métrica web / Otros (Si es así, ¿cuales?)

20. ¿Con qué frecuencia se actualiza el lenguaje documental? En un lapso menor a un mes / Mensualmente / Trimestralmente / Anualmente / Más de un año / El lenguaje no se ha actualizado desde su creación
21. ¿Ese lenguaje tiene conexión con otro tipo de lenguajes? Sí / No
22. En caso afirmativo a la pregunta anterior, ¿con qué otros lenguajes? (Respuesta abierta)
23. ¿El lenguaje está basado en algún modelo general de lenguajes de marcado semántico como: Dublin Core / OWL / Otra…
24. ¿Se evalúa de alguna forma el funcionamiento y la calidad del lenguaje documental? Sí / No

24B. En caso de respuesta afirmativa a la anterior, ¿cómo se evalúa? (Respuesta abierta)

25. ¿Cuál es el coste de la elaboración y mantenimiento del lenguaje documental? (Respuesta abierta)
26. ¿De qué forma se integra este lenguaje en la actividad cotidiana del medio? (Respuesta abierta)

Sección 3: Preguntas sobre el marcado semántico de los contenidos del medio

27. ¿Qué elementos y categorías de descripción semántica se emplean para realizar el marcado semántico de cada documento? (Respuesta abierta)
28. ¿Qué contenidos del medio son etiquetados? Contenidos publicados por el medio en su sitio web / Contenidos publicados por el medio en medios sociales/redes sociales / Otra…
29. Si se etiquetan contenidos publicados en medios o redes sociales, ¿qué redes o medios se emplean? Facebook / Twitter / YouTube / Instagram / Google+ / Snapchat / Tumblr / Pinterest / Otra u otras [Cuál o cuáles]
30. ¿Quiénes son los encargados de etiquetar los contenidos del medio? Periodistas / Documentalistas / Programadores/informáticos / Gestores de información digital / Especialistas en medios sociales (SEO, SEM, etc.) / Diseñadores gráficos y responsables de visuaización / Responsables de métrica web / Otra…
31. ¿Qué relación existe entre el etiquetado o marcado semántico y el lenguaje documental que usa el medio? (Respuesta abierta)
32. ¿Existe alguna forma de evaluar la efectividad del etiquetado? Sí / No

32. En caso afirmativo, ¿cuál es el procedimiento para medir esa efectividad? (Respuesta abierta)

Referencias

Baños Moreno, María José (2013). Fuentes para la actualización de macrotesauros: noticias de divulgación científica. // Cuadernos de Gestión de Información. 3 (2013) 13-24.

Baños Moreno, María José; Felipe, Eduardo R.; Pastor Sánchez, Juan Antonio; Martínez Béjar, Rodrigo y Lima, Gercina (2015). Metadatos en noticias: un análisis internacional para la representación de contenidos en periódicos. // XII Congreso ISKO España y II Congreso ISKO España-Portugal, 19-20 de noviembre, 2015.

Carroll, Nicholas (2010): Search Engine Optimization. // Bates, Marcia J. (coord). Encyclopedia of library and information sciences. Vol. 6, 4613-4629.

Codina, Lluís; Gonzalo-Penela, Carlos; Pedraza-Jiménez, Rafael; Rovira, Cristòfol (2017). Posicionamiento Web y Medios de Comunicación: Ciclo de Vida de una Campaña y Factores SEO. Barcelona: Departamento de Comunicación. Serie Editorial DigiDoc, 2017.

Codina, Lluís; Pedraza Jiménez, Rafael (2011). Tesauros y ontologías en sistemas de información documental. // El profesional de la información, 20:5 (sept-oct, 2011) 555-563.

du Preez, Madely (2015). Taxonomies, folksonomies, ontologies: what are they and how do they support information retrieval?. // The Indexer, 33:1, 29-37.

Fondevila Gascón, Joan Francesc (2017). Algoritmos sobre el impacto de los medios de comunicación en medios sociales: estado de la cuestión. // Icono14. 15:1 (2017) 21-41.

García Gutiérrez, Antonio (2011). Epistemología de la Documentación. Barcelona: Stonberg, 2011.

García Gutiérrez, Antonio (2014). Análisis documental de noticias de prensa en sistemas de información factual. // Revista Española de Documentación Científica. 37:2 (abril-junio, 2014) e046.

García Jiménez, Antonio (2004). Instrumentos de representación del conocimiento: tesauros versus ontologías. // Anales de Documentación. 7 (2004) 79-95. http://revistas.um.es/analesdoc/article/view/1691/1741

García Jiménez, Antonio (2016). Organización del conocimiento para la documentación en periodismo: situación y prospectiva. // Scire. 22:2 (jul.-dic. 2016) 21-28.

Jung, Jaemin; Song, Haeyeop; Kim, Youngju; Im, Hyunsuk; Oh, Sewook (2017). Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. // Computers in Human Behavior. 71 (2017) 291-298.

López-García, Xosé; Toural-Bran, Carlos; Rodríguez-Vázquez, Ana Isabel (2016). Software, estadística y gestión de bases de datos en el perfil del periodista de datos. // El profesional de la información (EPI). 25:2 (2016) 286-294.

Marcos-Recio, Juan Carlos; Edo, Concha (2015). Análisis de la nueva perspectiva de la documentación periodística en los medios de comunicación españoles. // Revista general de información y documentación. 25:2 (2015) 389-423

Martínez González, M. Mercedes; Alvite Díez, M. Luisa (2014). Propuesta metodológica de evaluación de gestores de tesauros compatibles con la web semántica. // Anales de Documentación. 17:1 (2014).

Meléndez-Malavé, Natalia; Hirschfeld-Suárez, Rocío (2016). Situación de los centros de documentación escritos andaluces. // El profesional de la información. 25:4 (julio-agosto 2016) 606-615

Mendes, Paula Raphisa; Martins dos Reis, Raquel; Coura Moreira dos Santos Maculan, Benildes (2015). Tesauros no acesso à informação: uma retrospecção. // Revista ACB: Biblioteconomia em Santa Catarina. 20:1 (2015) 49-66.

OK Diario (2017a). Nuevo récord de visitas de OKDIARIO: 33.089.131. // OK Diario. 19/8/2017. https://okdiario.com/audiencia/2017/08/19/comscore-julio-2017-1251511

OK Diario (2017b). Nuevo récord de visitas: 37.545.412 y de usuarios únicos: 9.160.000. // OK Diario. 22/09/2017. https://okdiario.com/audiencia/2017/09/22/comscore-agosto-2017-1343824

OK Diario (2017c). Nuevo récord de visitas de OKDIARIO: 39.380.408 y ya es sexto en el Top 10 de periódicos generalistas. // OK Diario. 20/10/2017. https://okdiario.com/audiencia/2017/10/20/comscore-septiembre-2017-1433604

Pastor Sánchez, Juan Antonio (2013). Marcado semántico: tecnologías y aplicación para la representación de sistemas de organización del conocimiento en el contexto Linked Open Data. // Scire. 19:2 (2013) 55-68.

Pastor Sánchez, Juan Antonio; Martínez Méndez, Francisco Javier; Rodríguez Muñoz, José Vicente (2012). Aplicación de SKOS para la interoperabilidad de vocabularios

controlados en el entorno de linked open data. // El profesional de la información. 21:3 (may-jun, 2012) 245-253.

Pérez Sanchidrián, Elaine; Campos Posada, Raúl; Campos Posada, Gloria Elisa (2014). Etiquetado social: un modelo de representación de la información en la blogosfera. // Biblios. 56 (2014) 19-27.

Pintado Navarro, Rocío (2013). El comportamiento informacional de los periodistas en la Región de Murcia. // Cuadernos de Gestión de Información. 3 (2013) 25-51.

Renó, Denis; Renó, Luciana (2017). Algoritmo y noticia de datos como el futuro del periodismo transmedia imagético. // Revista Latina de Comunicación Social. 72, (2017) 1.468-1.482. http://www.revistalatinacs.org/072paper/1229/79es.html

Rubio Lacoba, María (2012). Nuevas destrezas documentales para periodistas: el vocabulario colaborativo del diario El País. // Trípodos. 31 (2012) 65-78.

Saad Corrêa, Elizabeth; Bertocchi, Daniela (2012). A cena cibercultural do jornalismo contemporâneo: web semântica, algoritmos, aplicativos e curadoria. // MATRIZes. 5:2 (ene-jun, 2012) 123-144

Sánchez Cuadrado, Sonia; Colmenero Ruiz, María Jesús; Moreiro, José Antonio (2012). Tesauros: estándares y recomendaciones. // El profesional de la información. 21:3 (may-jun, 2012) 229-235.

Søbak, V.; Pharo, N. (2017). Decentralized subject indexing of television programs: The effects of using a semicontrolled indexing language. // Journal of the Association for Information Science and Technology. 68:3 (2017) 739-749.

Soler Monreal, Concha; Gil Leiva, Isidoro (2010). Posibilidades y límites de los tesauros frente a otros sistemas de organización del conocimiento: folksonomías, taxonomías y ontologías. // Revista Interamericana de Bibliotecología. 33:2 (jul-dic, 2010) 361-377.

Stone, Martha L. (2014). Big data for Media. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Big%20Data%20For%20Media_0.pdf

Szostak, Rick (2014). Advances in Classification Research Online 2013. Classification, Ontology, and the Semantic Web. // Advances In Classification Research Online. 24:1 (2014) 30-37.

Yedid, Nadina (2013). Introducción a las folksonomías: definición, características y diferencias con los modelos tradicionales de indización. // Información, Cultura y Sociedad. 29 (dic. 2013) 13-26.

Enviado: 2018-04-12. Segunda versión: 2018-10-14.
Aceptado: 2018-10-31.