Metadatos en noticias: un análisis internacional para la representación de contenidos en periódicos

Metadata in news: an international review for representing the content of newspapers

María-José Baños-Moreno (1), Eduardo R. Felipe (2), Juan Antonio Pastor-Sánchez (3), Rodrigo Martínez-Béjar (4), Gercina Lima (5)

(1) Universidad de Murcia, Campus de Espinardo, C.P. 30100, mbm41963@um.es,
(2) Universidad Federal de Minas Gerais, erfelipe@hotmail.com,
(3) Universidad de Murcia, Campus de Espinardo, C.P. 30100, pastor@um.es,
(4) rodrigo@um.es,
(5) Universidad Federal de Minas Gerais, glima@eci.ufmg.br

Resumen

Los periódicos trabajan con un gran volumen de información que necesita ser descrita adecuadamente. Para ello, las etiquetas “title”, “keywords” y “description” son muy utilizadas en el código fuente de las noticias online. Sin embargo, estas no resultan suficientemente descriptivas. Así, surgen estándares de metadatos, con el fin de facilitar la interoperabilidad y profundizar en la descripción. Actualmente, las etiquetas HTML y diversos estándares conviven en el sector periodístico, con diversos grados de implantación. Se analiza el código fuente de diarios internacionales de información general y se lleva a cabo una profunda revisión bibliográfica sobre estándares de metadatos. El propósito es conocer qué estándares existen, y evaluar su uso en los códigos fuente de una muestra de periódicos. Para ello se identifican los metadatos de contenido semántico de los códigos fuente. Además se desarrolla el software MetadadosHTML. Como conclusiones destacan la gran distancia entre los estándares recogidos en la bibliografía y los mostrados en los código fuente. En el primer caso, los más referenciados son los formatos NewsML y NITF, implementados por algunos medios y agencias de prensa, al menos a nivel interno. Por el contrario, en el código fuente los más habituales son schema.org y dos esquemas para mostrar información en redes sociales, Open Graph Protocol (usado por Facebook) y Twitter Cards. Esto, evidencia la convivencia de diversos estándares de metadatos en el ámbito de los medios de comunicación y pone de relieve la falta de uniformidad en su uso. Para alcanzar el ideal de interoperabilidad de los contenidos, es preciso utilizar tecnologías de la Web Semántica. En este sentido, se debería tender a definir ontologías o vocabularios RDF para las diferentes propuestas analizadas.

Palabras clave: Periódicos, metadatos, schema.org, Open Graph Protocol, NITF, NewsML

Abstract

Newspapers use a large volume of information that must be described adequately. To do this, the tags “title”, “keywords” and “description” are widely used in the source code of online-news items. However, these are not descriptive enough for the news items. So, metadata standards are created in order to facilitate the interoperability and make a deeper description of them. Currently, HTML tags and several standards live together in the newspaper industry, with different degrees of implementation. In this paper, we analyze the source code of international general-information newspapers. Moreover, we conducted an in-depth literature review on metadata standards. The aim is to analyze what standards exist and how they are used in a sample of newspapers’ source codes. For that, we identify the semantic metadata in the source codes and design the software MetadadosHTML. As conclusions, the great distance between metadata standards identified in the literature review and those in the source codes is clearly shown. In the former, the most cited metadata are NewsML and NITF formats, implemented at least in an internal level by some media and press agencies. By the contrary, schema.org and two social media schemas (Open Graph Protocol for Facebook and Twitter Cards) are the most common in the latter one. The coexistence of different metadata standards in the media sector is exhibited, stressing the lack of uniformity in their use. To achieve the ideal of interoperability between contents, the use of Web Semantic technologies is needed. In this regard, to define ontologies and RDF vocabularies for the different proposals analyzed should be a tendency.

Keywords: Newspapers, metadata, schema.org, Open Graph Protocol, NITF, NewsML

Baños-Moreno, María-José; Felipe, Eduardo R.; Pastor-Sánchez, Juan-Antonio; Martínez-Béjar, Rodrigo; Lima, Gercina. *Metadatos en noticias: un análisis internacional para la representación de contenidos en periódicos.* En XII Congreso ISKO España y II Congreso ISKO España-Portugal, 19-20 de noviembre, 2015, Organización del conocimiento para sistemas de información abiertos. Murcia: Universidad de Murcia

Agradecimientos

Este trabajo ha sido posible gracias a la financiación de la Universidad de Murcia a través del programa de becas predoctorales con resolución R–406/2011. Agradecimentos ao apoio das Agências de fomento FAPEMIG e CNPq/Brasil

1. Introducción

Internet, que ha propiciado un profundo cambio cultural, ha puesto al alcance de los ciudadanos el mayor volumen de información de todos los tiempos, aunque gran parte de forma poco estructurada. Esto se traduce en la dificultad para descubrir conocimiento relevante (Paepen et al., 2002; Pereira y Baptista., 2003a; Castells et al, 2006; Silva et al., 2008; Kallipolitis et al., 2012; Nies et al., 2012; Díaz Nosty, 2013; Mannens et al., 2013), afectando de forma considerable a la prensa escrita, el dominio más rico en contenidos en la Web (Saleh y Al-Khalifa, 2009, Abadal et al., 2014). Este crecimiento y disponibilidad contribuye a la necesidad de agrupar la información a nivel semántico (Paepen et al., 2002; Pereira y Baptista, 2003a; Pereira y Baptista, 2004). Buena parte de los métodos de recuperación de información, fundamentados en algoritmos de localización de términos o cálculo de frecuencias, entre otros, son útiles, pero insuficientes para lograr un resultado relevante y preciso. Los metadatos, un conjunto común de elementos que representan un documento (Yaginuma et al., 2003a; Shoval et al., 2008) pueden suplir las carencias de estos métodos. Los metadatos pueden concebirse como una forma efectiva de gestionar, de forma estandarizada, enormes cantidades de información (Baptista y Machado, 2001; Paepen et al., 2002; Martínez-Fernández et al., 2004; Wong et al., 2010).

Si bien el análisis e identificación de los elementos formales de una noticia están resueltos, el empuje de la Web Semántica en el ámbito periodístico vendría del marcado del contenido de los artículos. Desde principios de siglo, las principales líneas de investigación para este propósito se centran en la indización humana y automática, procesamiento de lenguaje natural (PLN), métodos estadísticos, etc. Sin embargo, los problemas de interoperabilidad y la falta de un estándar de representación comúnmente aceptado (Castells et al., 2006; Troncy, 2008) persisten en tanto no se afianzan las técnicas y herramientas de anotación semántica. La utilización de metadatos de presentación del contenido es un paso intermedio.

Con la llegada de los medios de comunicación a Internet, muchas noticias se publican en línea y en diferentes formatos (Paepen et al., 2002; Agarwal et al., 2012). Los periódicos se transforman aprovechando las posibilidades que las nuevas tecnologías les brindan. Por esto, no es extraño encontrar proyectos que aplican tecnologías de la Web Semántica en los procesos de producción y difusión de información de actualidad, con experiencias basadas en una cierta automatización asistida por personal del periódico, como Rubio Lacoba (2012), Fernández et al. (2006, 2007a, 2007b, 2010 y 2012) ó Castells et al. (2006).

Una de las bases de la Web Semántica es la reutilización de productos previos. En este sentido, es posible aprovechar aquellos metadatos definidos por el propio medio. En función de las necesidades y posibilidades de los periódicos, existen diferentes opciones para expresar estos elementos. Además, hay que tener en cuenta los estándares propios del estado del arte, tales como NITF o NewsML. Estos son utilizados por agencias y periódicos para hacer accesibles sus contenidos y pueden ser fuentes para una ontología de dominio (Fernández et al., 2007; 2010). De hecho, este trabajo se enmarca en un proyecto de modelado de una ontología de dominio en política y economía. La ventaja de los metadatos es su potencialidad para la recuperación de información utilizando los principios de interoperabilidad semántica.

El objetivo de este trabajo es conocer qué estándares de metadatos se utilizan para representar el contenido de una noticia, tanto en periódicos como en la bibliografía especializada. En este sentido y dentro del contexto de los medios de comunicación escrita, se entiende noticia como un texto que relata objetivamente un hecho (noticioso) susceptible de ser de interés público, por ser algo novedoso, destacable y/o actual. La noticia es el género informativo por excelencia y el de uso más extendido (García Gutiérrez, 2014). Para ello, se ha realizado una búsqueda en la literatura científica sobre este tema, al tiempo que se ha utilizado una muestra internacional de periódicos de información general para el análisis del código fuente y conocer los metadatos utilizados.

El trabajo se organiza de la siguiente manera: primero se muestra qué metadatos para prensa son referenciados en la bibliografía; a continuación se describe la metodología para la selección de diarios de información general, la toma de datos y la comparación entre los metadatos identificados en los trabajos previos y los presentes en los códigos fuente de noticias de diarios de la muestra; después se plantean algunos de los resultados obtenidos; finalmente se exponen conclusiones y propuestas de trabajos

futuros. Entre las conclusiones, destaca la gran distancia entre los estándares recogidos en la bibliografía y los mostrados el código fuente de los periódicos. En el primer caso, los más referenciados son los formatos NewsML y NITF, implementados por algunos medios y agencias de prensa a nivel interno. En proyectos desarrollados en el ámbito académico para la gestión de noticias como Omnipaper, ePaper o News, también son frecuentes. Por el contrario, en el código fuente los metadatos más habituales son Schema.org y los esquemas para visualizar información en redes sociales, *Open Graph Protocol* (usado por Facebook) y *Twitter Cards*. Esto, evidencia, por un lado, la convivencia de diversos estándares de metadatos en el ámbito de los medios de comunicación y pone de relieve la falta de uniformidad en el uso de estos. Cabe preguntarse el porqué del salto entre lo referenciado y lo expuesto en los códigos fuente. Es posible que se deba a ciertas reticencias a una exposición pública de los mecanismos de representación de noticias ante un potencial uso por otros agentes, o simplemente a un modelo de gestión que distingue claramente entre lo interno y lo externo.

2. Metadatos en prensa

Los metadatos son el conjunto de atributos o elementos que representan un recurso, facilitando su identificación y distinguiendo documentos relevantes de los no relevantes (Yaginuma et al., 2003b; Hillmann, 2005; Abbud Grácio y Fadel, 2010, p. 10-11; Kallipolitis et al., 2012; Asociación Española de Normalización y Certificación, UNE-ISO/TR 23081-3:2012). Más específicamente:

> "Los metadatos son descripciones estructuradas y codificadas que describen características y propiedades de objetos y recursos para facilitar su localización, recuperación, valoración, administración, persistencia e interoperabilidad" (Pastor-Sánchez, 2013, p. 22).

Centrándonos en los que tienen por objeto la representación del contenido de un documento, podemos definirlos como el "conjunto de propiedades asociadas a la faceta descriptiva del contenido de una noticia, esto es, datos que se pueden deducir desde el consumo (lectura, escucha o visionado) de la noticias (IPTC, 2009, pp. 20). Se excluyen, por tanto, aquellos datos formales de la producción o difusión de la noticia, como autor, medio, fecha de publicación [que no responden] al análisis estructural de contenido que nos ocupa, pues pertenecen a otro plano de descripción" (García-Gutiérrez, 2014). En agencias y medios de comunicación los metadatos semánticos se han ido añadiendo manualmente para, en las búsquedas, filtrar mediante palabras clave y categorías (Fernández et al., 2006). Además del código fuente, estos metadatos pueden encontrarse en un documento independiente (al que se referencia), o bien almacenados en bases de datos externas. Este meta-contenido, inicialmente se refleja en su código fuente mediante las etiquetas <meta> "title", "keywords" y "description" (1) (Alesso y Smith, 2004, pp. 60):

- "Title" provee información general sobre el documento, es usada para indexar la página basándose en palabras clave del título
- "Description" es el resumen de la página web que el motor de búsqueda mostrará en la página de resultados
- "Keywords" permite listar explícitamente las palabras clave que se asocian con la página web, es usada por los motores de búsqueda para la indexación. Desde 2009, esta etiqueta dejó de ser utilizada por Google (Cutts, 2009) siendo reemplazada por <news_keywords> para las noticias (Galfi, 2012; McGee, 2012).

Con el tiempo fue evidente la necesidad de definir conjuntos de elementos comunes (Yaginuma et al., 2003a; Pereira y Baptista, 2004) que facilitaran el intercambio de información a nivel semántico, profundizando en la descripción de los recursos. Así, surgen estándares para metadatos como *Dublin Core Metadata Iniciative* (DCMI) y otras opciones más específicas para las necesidades concretas de un área de conocimiento, una tipología documental, una organización, etc. En el ámbito periodístico existen diversas posibilidades. En este sentido, las entidades más relevantes son la *Newspaper Association of America* (en adelante NAA) y la *International Press Telecommunications Council* (en adelante IPTC), un consorcio internacional de agencias de noticias, editores y distribuidores de periódicos (Pellegrini, 2012).

La forma de indicar las propiedades de los recursos, en este caso noticias, se basa en una idea similar a las sentencias RDF (*Resource Description Framework*) en las que los recursos son descritos a partir de tripletas compuestas por un sujeto, un predicado y un objeto (Pastor Sánchez, 2013, p. 56). Con los estándares de metadatos, aparece un elemento intermedio que busca sistematizar y ahondar en las posibilidades de descripción, como se muestra a continuación:

Sujeto	*Predicado*	*Objeto*
Noticia	Elemento de metadato (propiedad)	Valor que toma el elemento
Noticia	Keywords	Keyword1, Keyword2, etc.

↓

Sujeto	*Esquema metadatos*	*Elemento metadato*	*Valor*
Noticia	p.e. Dublin Core	Keywords	Keyword1, Keyword2, etc.

Tabla 1. Formas de describir una noticia en el código fuente

Se indican, por orden alfabético, los metadatos para representar noticias en medios de comunicación de acuerdo con la bibliografía. En el anexo se muestra las referencias que citan estas especificaciones:

ANPA 89-3 (The ANPA Wire Service Guidelines for 1200 Baud transmission). Estas guías, también conocidas como ANPA 1312 y NAA 89-3, fueron desarrolladas al amparo de NAA como formato para la codificación y transmisión de información en los servicios de noticias y los métodos por los cuáles los periódicos podían utilizarlos. Su última actualización es de 1989.

Dublin Core (DC) es un simple pero efectivo conjunto de metadatos para la descripción de un amplio rango de recursos online. Este estándar es creado y mantenido por un grupo internacional e interdisciplinar de profesionales de diversas áreas. DC trabaja a dos niveles: simple, con 15 elementos básicos de descripción; y cualificado, que incluye otros tres elementos y una serie de cualificadores. Algunas iniciativas se han centrado en la adaptación de DC a sectores concretos como DC-Lib, para bibliotecas. Allen et al. (2007), Fernández et al., (2006), Castells et al. (2006) o Yaginuma et al. (2003) son sólo algunos de los trabajos que utilizan este esquema, combinado con otros como NITF, para proyectos como NEWS, Neptuno u Omnipaper, por ejemplo.

Information Interchange Model (IMM) fue definido por IPTC para el intercambio de contenido binario de noticias, principalmente fotografías (Fernández et al, 2007a).

IPTC 7901 fue la primera recomendación de la IPTC para el intercambio de mensajes sobre noticias entre periódicos y agencias de noticias. Se desarrolló con el propósito de facilitar la comunicación, superando la rigidez del ANPA 89-3, en la que se inspira. Fue actualizado por última vez en 1995. Según IPTC todavía es el formato más utilizado por las agencias para enviar artículos de texto a clientes junto con ANPA 1312 (en América).

Los microdatos son frameworks para introducir contenido semántico en el código HTML, un grupo de pares nombre-valor (Hickson, 2013) donde se identifica qué esquema se utiliza, qué recurso se describe (itemscope itemtype) y sus características (itemprop). Entre los referenciados por la bibliografía destacan: rNews, un modelo compuesto de clases conceptuales con un nivel suficiente de expresividad semántica en términos de vocabulario y tipos de datos que cubren los atributos más importantes de la noticia para su marcado semántico (Pellegrini, 2012; Heravi y McGinnis, 2013); vCard es un modelo para el intercambio de información de datos de identificación y localización de una persona o empresa, tales como nombre, apellidos, domicilio o email. La última versión de este modelo, RFC6350 se ha adaptado a la Web Semántica mediante su representación en una ontología por el W3C (Iannela y McKinney, 2013).

Microformatos "son pequeños patrones en HTML para representar cosas habitualmente publicadas como personas, eventos, posts de blogs, revisiones y etiquetas en la web" (http://microformats.org/wiki/introduction). Estos patrones, embebidos en el código HTML, permiten marcar semánticamente cierto contenido de la web. En esencia se utiliza el atributo "class" para identificar vocabularios, clases y propiedades (Pastor-Sánchez et al., 2013). Existen varios tipos, dependiendo de la información a describir. En la literatura se hace mención de hNews, un microformato proporcionado por la agencia de noticias Associated Press (Pellegrini, 2012) que establece un número de campos para describir completamente el trabajo periodístico. En la última modificación de este estándar también participaron autores de la Media Standards Trust, una organización independiente para fomentar la calidad, transparencia y rendición de cuentas en la era digital. Para información más general es necesario recurrir a un microformato más general, hEntry, que sustituye a hAtom.

NewsForm es presentado por Mueller (2000) como un nuevo formato para representar información sobre eventos en noticias, pues NITF y XMLNews eran demasiado generales. Este formato es compatible con los dos anteriores.

News Industry Text Format (NITF): desarrollado por la IPTC y la NAA (Mueller, 2000), es un formato estándar abierto y público, en XML, que define la estructura de artículos de noticias, facilitando la reutilización de información. Tam-

bién permite la anotación de entidades que aparezcan en el contenido de los artículos (inline metadata). Entre otros, según IPTC ([2014]) es utilizado por The Associated Press, Dow Jones, The New York Times y Deutche Presse-Agentur (2). Es el primer estándar de la nueva generación tras ANPA 89-3, IPTC 7901 e IIM (Fernández et al., 2007a). La última versión fue liberada en junio de 2012.

NewsML Architecture (NAR): también diseñado por la IPTC, es un estándar de intercambio para noticias en general utilizado para representar noticias como paquetes multimedia (García et al., 2006). NewsML apoya la expresión de metadatos e información de noticias electrónicas y las relaciones entre datos de la noticia (Kodama et al., 2008), facilitando el reaprovechamiento de contenidos descriptos a partir de metadatos. La última versión desarrollada es NewsML-G2 2.19. Se complementa con NITF, ya que presentan diferentes opciones y posibilidades. Este formato es utilizado, por ejemplo, por Reuters (Tenenboim et al., 2008) y en el proyecto ePaper como una de las opciones, junto a RSS. Entre otros NewsML ofrece soporte para el contenido de noticias en cualquier formato y metadatos para la descripción de una noticia como un todo. EventsML, diseñado por IPTC es un estándar para transmitir información sobre eventos en la industria informativa (Biyun et al., 2009). Actualmente se integra en NewsML.

News Value Markup Language (NVML) es un lenguaje desarrollado por Kodama et al. (2008) para la descripción de metadatos para la transmisión de documentos con NewsML de forma independiente a este estándar, permitiendo el control de la distribución y gestión estos documentos. Entre los metadatos definidos para NVML en el cuerpo del documento se incluye nvValue para el contenido de la noticia, que incluye, entre otros, resumen y categoría. El sistema de edición permite la clasificación automática de las noticias en categorías de IPTC SubjectCodes en función de la relevancia de cada palabra del texto y de las noticias ya almacenadas en el sistema.

Publishing Requirements for Industry Standard Metadata (PRISM) es una iniciativa de PRISM Working Group y nextPub/PRISM Source Vocabulary Working Group que aúna un conjunto de esquemas y, sobre todo, vocabularios, que los publicadores pueden utilizar para la descripción de contenidos a descargar o intercambiar mediante sindicación, agregación, etc., tales como noticias, blogs, libros, revistas, newsletter, entre otros. Para ello, usa como base Dublin Core (DC) y Dublin Core Metadata Iniciative (DCMI). Uno de los usos más frecuentes de PRISM es la codificación tanto de metadatos como del texto de un artículo en XML, usando para ello PRISM Aggregator Message (PAM) para enviar contenidos entre miembros de la cadena de suministro de información. Para el marcado en el propio texto, también cuenta con PRISM Inline Markup (PIM). La última versión (3.1.) fue publicada a finales de 2014.

Really Simple Syndication (RSS) es un formato para la sindicación de contenidos y metadatos de descripción. “La sindicación de contenidos ayuda al usuario a estar actualizado sobre nuevos contenidos publicados en diferentes y distribuidas fuentes de información y mejora la visibilidad dichos contenidos, garantizando que los usuarios serán advertidos cuando un nuevo contenido sea publicado” (Pereira y Baptista, 2009). Es utilizado, entre otros, por Saias et al. (2006) para obtener una primera clasificación de las noticias de última hora recibidas en un medio de comunicación.

XMLNews fue creado por WavePhore y mantenido por Megginson technologies Ltd., se dividía en dos especificaciones. XMLNews-Story para marcar el contenido de noticias en formato texto mediante un subconjunto de etiquetas de NITF. XMLNews-Meta era un formato de metadatos que podía aplicarse sobre una noticia o cualquier otro recurso, sobre el que aportaba información. Las últimas actualizaciones se realizaron en 1999.

Algunas de estas iniciativas se solapan en ciertos aspectos, aunque sus propósitos son diferentes. En otras ocasiones, se complementan. De hecho, los grupos de trabajo de PRISM y de IPTC están trabajando juntos para definir un formato y vocabulario de metadatos comunes. Por ejemplo, PRISM y IPTC NewsCodes pueden ser utilizados en NewsML.

Merece la pena destacar NewsCodes, un conjunto de vocabularios diseñados por la IPTC, una jerarquía de clasificación de materias con 3 niveles y 17 categorías en su primer nivel (García et al. 2006; Troncy, 2008). Anteriormente fue conocida como Topic Sets (Fernández et al., 2007). Son utilizados, entre otros, por Pereira et al. (2003), Pereira y Baptista (2003) y Schranz (2005) en el proyecto Omnipaper; Fernández et al. (2006), Fernández et al. (2007b) y Fernández et al (2010) en NEWS; Tenenboim et al. (2008) en ePaper; y Kallipolitis et al. (2012) en World News Finder (WNF). En todos los casos, esta taxonomía fue utilizada para la construcción de las ontologías que permitían la categorización de noticias. NewsCodes están compuestos por términos de diversos dominios, clasificados

jerárquicamente en tres niveles, de más general a más específico: subject, subject matter, subject detail. Estos vocabularios aportan valores para los elementos y atributos, por ejemplo, de documentos NewsML y NITF. Algunas taxonomías que destacan de NewsCodes son:

- Subject Reference System (SRS), creadas por IPTC y NAA. Ofrecen un esquema en el que se define valores para clase de objeto, atributo, referencia de materia (subject reference), sinónimos y cualificadores. NITF, NewsML e IIM fueron concebidos en 2003 como extensiones de SRS. Está compuesto por Subject Code, Subject Qualifier, Media Type, NewItem y Genre
- SubjectCodes es una taxonomía centrada en el texto, desarrollada desde 1990 hasta 2010, cuando es reemplazada por una nueva extensión, Media Topics, actualizada recientemente

Aunque la mayor parte de la bibliografía y la muestra seleccionada se centra en periódicos, los metadatos también son utilizados en agencias de noticias. AFP, por ejemplo, según Guerrillot (2006) implementa el formato NewsML y hasta aboga por la implicación de la agencia en el desarrollo de NewsML-G2. También EFE y ANSA (Sánchez-Fernández et al. (2005; Fernández et al., 2010). La Agencia EFE, de hecho, muestra en su sitio web algunos documentos sobre cómo utilizar NewsML y NITF en sus artículos.

Algunos autores han reutilizado parte de estos metadatos para el desarrollo de proyectos de gestión de artículos periodísticos. ePaper (Tenenboim, 2008; Shoval et al., 2008) es un periódico personalizado basado en la categorización de noticias y perfiles de usuarios usando una ontología; Omnipaper (Yaginuma et al., 2003a, 2003b, 2004; Pereira y Baptista, 2003a, 2003b, 2004; Ariza Ávila y Baptista, 2004; Schranz, 2005), fue un proyecto desarrollado en el seno de la Unión Europea para la creación de un perfil de aplicación que resultara útil para el acceso a artículos periodísticos de diferentes medios europeos; Neptuno (Castells et al., 2006) tiene por objetivo desarrollar un archivo semántico de alta calidad para Diari SEGRE facilitando la descripción y anotación de artículos y la búsqueda de información integrando una ontología en el sistema; NEWS (*News Engine Web Services*) tiene como propósito desarrollar herramientas que asistan a las agencias de noticias en sus procesos de producción y distribución de artículos. También usa una ontología como herramienta principal del proyecto (Fernández-García et al, 2007, 2010); WNF (World News Finder) vincula una herramienta de extracción de conocimiento con una ontología para la extracción de información a partir de documentos de noticias del mundo en HTML y extrae conocimiento que coincide con la estructura de clasificación dada (Kallipolitis et al., 2012).

3. Metodología

Para configurar la muestra se llevó a cabo una selección continua de zonas geopolíticas, países y periódicos, siguiendo un proceso de muestreo no aleatorio intencional. Así, se dividió el mundo en 9 zonas, de acuerdo a sus competencias geopolíticas, históricas y socioeconómicas. A continuación, en cada área se seleccionaron aquellos países con mayor Producto Interior Bruto (PIB). Después, se escogió el diario generalista de cobertura nacional o internacional más leído, de acuerdo a los datos de accesos Web y compra de edición impresa de 4International Media & Newspaper (2012).

Para esta selección, se siguieron los siguientes criterios: Edición en inglés, español, portugués o francés. En Alemania no se halló ningún medio representativo escrito en un idioma distinto al alemán; Los periódicos deben contemplar de manera habitual noticias de política y economía, la temática de los artículos que queremos componer como muestra. De hecho, se han incluido en la muestra los tres diarios económicos más importantes a nivel internacional: Financial Times, The Economist (Reino Unido) y The Wall Street Journal (Estados Unidos); Los periódicos deben ser de tipo generalista, ya que facilita la comprensión de las noticias, con la excepción de los tres especializados ya indicados. En el caso particular de Australia, la ausencia de medios generalistas de referencia, hizo que se optara por un medio de corte económico.

La siguiente tabla muestra los veintiún periódicos analizados:

Parte del mundo	*País*	*Periódico*
UE y Norte de Europa	Alemania	Süddeutsche Zeitung
	Francia	Le Monde
	Reino Unido	The Daily Telegraph; Financial Times; The Economist
Norte de América	Estados Unidos	The New York Times; The Wall Street Journal
Sudamérica	Brasil	O Globo
	México	El Universal
ExRepúblicas URSS	Rusia	Pravda

África subsaharian a	Nigeria	Nigerian Tribune
	Sudáfrica	Independent Online
Asia	China	China Daily
	India	The Times of India
	Japón	Asahi Shimbun
Oriente Medio	Arabia Saudí	Arab News
	Emiratos Árabes Unidos	Gulf News
	Israel	Yedioth Aharonot
	Turquía	Today's Zaman
Oceanía	Australia	The Australian Financial Review
Norte de África	Egipto	The Daily News Egypt

Tabla 2. Muestra de periódicos

La toma de datos se basó en el análisis de los códigos fuente de tres noticias de cada uno de los diarios de la muestra. En diciembre de 2014, febrero y abril de 2015 se accedió a una noticia, seleccionada al azar, de la portada de cada periódico. Posteriormente se analizaron los metadatos de descripción de contenidos comprendidos entre las etiquetas <head> y </head>, esto es, etiquetas meta. También se tuvo en cuenta los metadatos utilizados para anotar el texto en el cuerpo de la propia noticia. Finalmente todos los datos obtenidos se recogieron en un documento Excel que facilitó la comparación entre lo mostrado por la bibliografía y los códigos fuente.

Finalmente, el software MetadadosHTML, una aplicación creada ad hoc y autoejecutable en Object Pascal permitió la descarga del código fuente, en HTML, de las 5582 noticias. Después se extrajeron diversos metadatos, a partir de los esquemas y patrones detectados.

4. Resultados

4.1. Aclaraciones terminológicas

Con una estructura jerárquica de metadatos es posible representación información en varios grados de detalle (Jokela et al., 2001) y con diferentes funciones que conviene aclarar: Metadatos en HTML, que aparecen en el código fuente de las noticias, mediante las etiquetas "title", "keywords", "news_keywords" y "description" o utilizando microformatos para la descripción semántica de información muy concreta. Fueron los primeros en aparecer y aún son muy utilizados; Vocabularios, tales como taxonomías, tesauros y ontologías, permiten indizar el contenido de una noticia a partir de una selección previa de conceptos y relaciones entre los mismos, como NewsCodes; Espacios de nombre ("namespace"), compuestos por un grupo de elementos y atributos. Un elemento existe en un espacio de nombre particular y debe ser validado contra el espacio de nombre referenciado. Así identificamos de dónde procede cada elemento y evitamos colisiones por el uso de elementos comunes (Alesso y Smith, 2004, pp. 70 y 200). Son imprescindibles para conocer la procedencia de esquemas y perfiles de metadatos; Esquemas y perfiles de metadatos, que definen elementos concretos de descripción de contenido de las noticias. Son insertados en el código fuente mediante espacios de nombre; Formatos de marcado semántico, como los que enumeran, Pastor-Sánchez et al. (2013) y Sporny et al., 2014: microformatos, RDFa y microdatos. En la figura 1, en anexo, se puede observar cómo se utilizan en una noticia seleccionada al azar.

4.2. Metadatos contenidos en el código fuente

Los estándares identificados fueron Dublin Core (descrito en el apartado 2), Schema.org, Open Graph Protocol (para distribuir información en Facebook), Twitter Cards. Además de las etiquetas meta "title", "keywords" y "description" y los microformatos de HTML, "news_keywords" resultó ser muy utilizada.

Schema.org es una colección de esquemas para marcar páginas HTML. Es usado por motores de búsqueda como Bing, Google Yahoo! y Yandex para mejorar la visualización y recuperación de recursos. El esquema más general es "Thing", que se aplica a cualquier tipo de recurso, pero lo habitual es utilizar uno más específico. Entre los diarios de la muestra se utilizan dos esquemas, Breadcrumb (6 diarios) y News Article (9 diarios, 2 de ellos hasta el nivel más específico). Breadcrumbs permite definir jerarquías temáticas o secciones. Funciona mediante la definición de enlaces, de más general a más específico; Para noticias, de más genérico a específico, encontramos los esquemas Thing > Creative Work > Article > News Article. Las instancias de News Article deben aparecer como valores de la propiedad associated Article del tipo Media Object. Esta clase contiene derivados de las propiedades de rNews, de IPTC. Para contenidos de noticias, a nivel del esquema Creative Work, la etiqueta <meta itemprop=keywords>, muestra las palabras clave, delimitadas por comas.

El Protocolo Open Graph permite que cualquier página web se convierta en un objeto rico, en un

gráfico social que, entre sus objetivos, persigue facilitar la visualización de sus contenidos en Facebook. Creado originalmente por los desarrolladores de esta red social, se inspira en Dublin Core, link-rel canonical, Microformatos y RDFa y actualmente es mantenida por la OWF (Open Web Fundation, 2014). Para la descripción de un recurso, cuenta con cuatro propiedades básicas (title, type, url e image), y siete opcionales (por ejemplo, site_name o description). Se definen en la cabera del código fuente "head" mediante la etiqueta <meta property="og:propiedad" >.

Con Twitter Cards, los desarrolladores de esta red de microblogging han definido algunos metadatos para visualizar información relativa a los recursos multimedia adjuntos en los tuits. Los veintiocho metadatos se organizan en siete subclases entre las que "Summary" se utiliza por defecto, como un resumen, y emplea cinco propiedades; "Summary Card with Large Image" es similar a la anterior, pero destaca una imagen; "Product card" se utiliza para productos informativos, con un total de 11 propiedades. Algunos tags a resaltar, por ejemplo, son <twitter:title>, <twitter:description>. Se definen en la cabera del código fuente "head" mediante la etiqueta <meta property="twitter:propiedad" >.

Google News utiliza el esquema sitemaps para marcar noticias y proporcionar metadatos a Google sobre ésta (Google Inc., 2014b). <news:keywords> consiste en "una lista separada por comas de palabras clave que describan el tema del artículo. Las palabras claves se pueden extraer de la lista de palabras clave existente de Google Noticias, aunque no es la única fuente de donde se puede obtener (Google Inc., 2014a)" (4)

4.3. Comparación bibliografía – códigos fuente

Los datos, de acuerdo con el siguiente gráfico, muestran la escasa coincidencia entre los metadatos referenciados en el estado del arte y los utilizados en el código fuente de las noticias de los diarios que componen la muestra.

Respecto a la bibliografía, destacan NewsML y NITF, ambos creados por IPTC como los estándares más referenciados. También Dublin Core es citado en un buen número de trabajos, la mayoría de ellos vinculados al proyecto Omnipaper. Es utilizado por un par de periódicos de la muestra. En contraste, en todos los códigos fuente aún se utilizan etiquetas meta para reflejar parte del contenido de las noticias (título, resumen y palabras clave). Open Graph Protocol, el esquema empleado para Facebook, y Twitter Card, son también muy utilizados. En este caso, el fin es facilitar la visualización de parte de las noticias en estas redes sociales. schema.org, que cuenta con un esquema propio para noticias (News Article) es aplicado por más de la mitad de los diarios de la muestra.

Los datos extraídos a partir del software MetadadosHTML facilitó el análisis de esta información. Se comprobó que título, resumen y palabras clave, los datos más comunes entre los estándares, coincidían. Esta información es utilizada en un trabajo posterior.

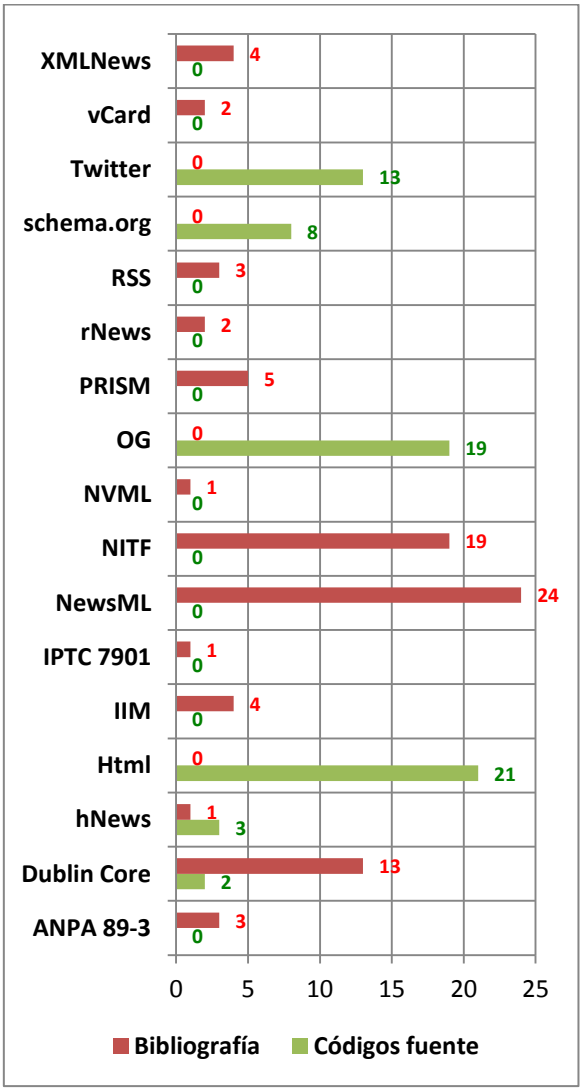

Gráfico 1. Comparación entre metadatos en bibliografía y códigos fuente de noticias

5. Conclusiones y propuestas de futuro

En el ámbito periodístico existen diversas formas de representar metadatos para el contenido, desde las más generales (las propias etiquetas de HTML, Dublin Core o los microdatos) a otras específicas y complejas como NITF y NewsML. La mayor parte de los estándares

específicos, recogidos en la bibliografía, permanecen ocultos al usuario final. El mapeado del código fuente de noticias, muestra que estos metadatos no se muestran en la parte de visible de las webs de los diarios. Por otro lado, aquellos metadatos utilizados en las webs apenas aparecen referenciados en el estado del arte.

Además, la diversidad de esquemas y tecnologías dificultan un mayor aprovechamiento y pueden hasta ir en contra de la estandarización u homogeneización de las formas de representar un documento. Si bien algunos esquemas generales, especialmente los basados en la difusión de información en redes sociales, se han propagado en todos los medios, la escasa presencia de esquemas más específicos en el código fuente de los 21 diarios, en los que se identifique algo más que palabras clave y resúmenes, apunta hacia dos posibles causas: no los utilizan y/o no los visibilizan. Esta conclusión coincide con lo indicado por Pellegrini (2012), que confiaba en el impulso de rNews para solventarlo. En cualquier caso, desde el punto de vista de la reutilización de la información sería deseable que, de entre los sets de metadatos para la representación de noticias, se alzara uno que fuera empleado por todos los medios de comunicación (García et al, 2006). Un modelo estándar o, como mínimo, consensuado por usuarios y desarrolladores (Pastor-Sánchez, 2013, p. 21) facilitaría un mayor intercambio y reutilización de información. Precisamente, González Cristóbal et al. (2002) advertían de la imposibilidad de "definir un recurso accesible de forma unificada debido a la heterogeneidad en los formatos de almacenamiento de noticias y en la meta-información que las describe". Por otro lado, para alcanzar el ideal de interoperabilidad de los contenidos, es preciso utilizar tecnologías de la Web Semántica. En este sentido, se debería tender a definir ontologías o vocabularios RDF para las diferentes propuestas analizadas en este trabajo como, por ejemplo, vCard.

Este trabajo representa un avance en el modelado de una ontología de dominio de política y economía, utilizando como base noticias. La principal ventaja de la información de actualidad es su abundancia en Internet y permanente actualización, lo que asegura una continua renovación de términos, conceptos y relaciones (Baños-Moreno et al., 2013). Esta idea no es nueva, pues ya existen ontologías que son punto de partida y también producto de noticias de prensa, tales como Saleh y Al-Khalifa (2009) y Fernández et al. (2006, 2007b, 2010 y 2012). El siguiente paso es la extracción de la información recogida por los metadatos del código fuente. Para ello, el software metadadosHTML, permite obtener un listado de términos sobre política y economía, de acuerdo con el contenido de las noticias. El análisis en profundidad de algunos de los esquemas de la bibliografía facilitará la identificación de entidades. Esta ontología cubriría "el hueco que existe entre la necesidad de los usuarios para la selección personalizada de contenidos y lo que la industria mediática puede ofrecer" (Fernández et al., 2007b), entre otros. En muchos casos, además, se reutilizan herramientas como IPTC Subject Codes o tesauros de una agencia o medio de comunicación. Castells et al. (2006), por ejemplo, parten de IPTC Subject Codes y aprovechan la experiencia de documentalistas y periodistas del Diari SEGRE; Tenenboim (2008) usa IPTC Subject Codes para la ontología implementada en ePaper; Biyun et al. (2009) parten de EventsML, NewsML y NITF con el fin de desarrollar una ontología para el análisis de noticias e implementar algoritmos para alimentar dicha ontología con nuevos elementos.

Una limitación importante de este trabajo es que, más allá de la identificación de espacios de nombre, prefijos y elementos visibles en el código fuente, no se alcanza a conocer qué estándares están siendo utilizados a nivel interno en cada medio de comunicación o para el intercambio con agencias de noticias. Entrevistas o encuestas a conocedores de los métodos de trabajo y de difusión de noticias de los diarios podría solventar esta cuestión.

Otra posibilidad habría sido utilizar las fuentes RSS de los diarios, o bien aquellos contenidos producidos por los lectores, ya convertidos en prosumidores (Toffler, 1980). La tendencia observada es la asunción de las posibilidades de la web 2.0. Facebook, de hecho, se ha convertido en un contenedor más de información, en un canal de comunicación entre el medio y los lectores. Tanto es así que, recientemente, The New York Times, The Guardian y la BBC, entre otros, han comenzado a publicar algunos de sus contenidos directamente en la red social (Jiménez Cano y Abad Liñán, 2015)

Notas

(1) También se utiliza "http-equiv" en lugar de "name", que "permite que los servidores que funcionan con el protocolo de transferencia de hipertexto (HTTP), recopilen la información para ofrecer los encabezados del mensaje de respuesta" (Lamarca Lapuente, 2013).

(2) Más información de organizaciones que usan NITF en https://www.iptc.org/site/News_ Exchange_Formats/NITF/Who's_using_it/

(3) El recurso de Google Testing-tool permite identificar el esquema utilizado y ver a qué nivel del esquema se ha implementado.

(4) Este servicio ha desaparecido en España, para diarios españoles, tras imponer el pago a editores en concepto de derechos de autor por la inclusión de artículos en este servicio desde el 16 de enero de 2014.

Referencias

4IMN. (2012). 4International Media & Newspaper. Retrieved October 11, 2012, from http://www.4imn.com

Abadal, E., Guallar, J., & Codina, L. (2014). Sistemi di documentazione della stampa periodica: quali sono e come valutarli? *AIB studi*, *54*(1), 75–86. http://doi.org/10.2426/aibstudi-9486

Abbud Grácio, J. C., & Fadel, B. (2010). Estratégias de preservação digital. En *Gestão, mediação e uso da informação* (pp. 58–83). São Paulo: Editora UNESP; Cultura Acadêmica. Recuperado de http://books.scielo.org/id/j4gkh/pdf/valentim-9788579831171-04.pdf

Agarwal, S., Singhal, A., & Bedi, P. (2012). Classification of RSS Feed News Items Using Ontology. En A. Abraham, A. Zomaya, S. Ventura, R. Yager, V. Snasel, A. K. Muda, & P. Samuel (Eds.), *International Conference on Intelligent Systems Design and Applications, ISDA* (pp. 491–496). New York: Ieee. http://doi.org/10.1109/ISDA.2012.6416587

Alesso, H. P., & Smith, C. F. (2004). *Developing Semantic Web Services*. Natick, Mass: A K Peters/CRC Press.

Allen, R. B., Japzon, A., Achananuparp, P., & Lee, K. J. (2007). A framework for text processing and supporting access to collections of digitized historical newspapers. En M. J. Smith & G. Salvendy (Eds.), *Human Interface and the Management of Information: Interacting En Information Environments, Pt 2, Proceedings* (Vol. 4558, pp. 235–244). Berlin: Springer-Verlag Berlin. Recuperado de https://scholarworks.iupui.edu/bitstream/handle/1805/4552/allen-2007-framework.pdf?sequence=1&isAllowed=y

Ariza Ávila, C. E., & Baptista, A. A. (2004). Uso de RDF y bases de datos de metadatos nativas dentro del proyecto Omnipaper. En *XATA 2004 : actas da 2ª conferência nacional* (Vol. 2, pp. 166–169). Porto: FEUP. Recuperado de http://repositorium.sdum.uminho.pt/handle/1822/2249

Asociación Española de Normalización y Certificación (2012). Información y documentación. Metadatos para la gestión de documentos. Parte 3: Método de auto-evaluación. UNE-ISO/TR 23081-3. Madrid: AENOR

Baños-Moreno, M.-J., Pastor-Sánchez, J.-A., & Martínez-Béjar, R. (2013). Propuesta de actualización de macro-tesauros a partir de noticias de divulgación científico-tecnológica. En *Informação e/ou Conhecimento: as duas faces de Jano* (pp. 99–112). Porto (Portugal): Faculdade de Letras da Universidade do Porto / CETAC.MEDIA. Recuperado de http://hdl.handle.net/10760/20684

Baptista, A. A., & Machado, A. (2001). Metadata Usage En an Online Journal - An application Profile. En A. Hübler, P. Linde, & J. W. . Smith (Eds.), *Electronic Publishing '01 - 2001 En the Digital Publishing Odyssey* (pp. 59–64). Kenterbury, UK: University of Kent. Recuperado de http://elpub.scix.net/cgi-bin/works/Show?200106

Biyun, H., Jun, W., & Yiming, Z. (2009). Ontology design for online news analysis (Vol. 4, pp. 202–206). http://doi.org/10.1109/GCIS.2009.78

Castells, P., Perdrix, F., Pulido, E., Rico, M., Fuentes, J. M., Benjamins, R., … Granollers, T. (2006). Newspaper Archives on the Semantic Web. En R. Navarro-Prieto & J. L. Vidal (Eds.), *HCI related papers of Interacción 2004* (pp. 267–276). Springer Netherlands. Recuperado de http://link.springer.com/chapter/10.1007/1-4020-4205-1_22

Cutts, M. (2009, September 21). Google does not use the keywords meta tag En web ranking. Recuperado de http://googlewebmastercentral.blogspot.com/2009/09/google-does-not-use-keywords-meta-tag.html

Díaz Nosty, B. (2013). *La prensa en el nuevo ecosistema informativo. «¡Que paren las rotativas!». La transición al medio continuo*. Barcelona: Ariel, Fundación Telefónica, Planeta. Recuperado de http://www.fundacion.telefonica.com/es/arte_cultura/publicaciones/detalle/238

Fernández, N., Arias Fisteus, J., Sánchez, L., & López, G. (2012). IdentityRank: Named Entity Disambiguation En the News Domain. *Expert Syst. Appl.*, *39*(10), 9207–9221. http://doi.org/10.1016/j.eswa.2012.02.084

Fernández, N., Blázquez, J. M., Fisteus, J. A., Sánchez, L., Sintek, M., Bernardi, A., … Ben-Asher, Z. (2006). NEWS: Bringing Semantic Web Technologies into News Agencies. En I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, … L. M. Aroyo (Eds.), *The Semantic Web - ISWC 2006* (pp. 778–791). Springer Berlin Heidelberg. Recuperado de http://link.springer.com/chapter/10.1007/11926078_56

Fernández, N., Blázquez, J. M., Sánchez, L., & Bernardi, A. (2007). IdentityRank: Named entity disambiguation En the context of the NEWS project (Vol. 4519 LNCS, pp. 640–654). Recuperado de http://www.scopus.com/inward/record.url?eid=2-s2.0-34548061897&partnerID=40&md5=09bf835b0eb41d2e206ff59b97ec5ca3

Fernández, N., Fuentes, D., Sánchez, L., & Fisteus, J. A. (2010). The NEWS ontology: Design and applications. *Expert Systems with Applications*, *37*(12), 8694–8704. http://doi.org/10.1016/j.eswa.2010.06.055

Fernández, N., Sánchez-Fernández, L., Blázquez-del-Toro, J. M., & Villamor-Lugo, J. (2007). The News Ontology for Professional Journalism Applications. En R. Sharman, R. Kishore, & R. Ramesh (Eds.), *Ontologies* (pp. 887–919). Springer US. Recuperado de http://link.springer.com/chapter/10.1007/978-0-387-37022-4_32

Galfi, R. (2012, September 19). Google News Blog: A newly hatched way to tag your news articles. Recuperado de http://googlenewsblog.blogspot.com.es/2012/09/a-newly-hatched-way-to-tag-your-news.html

García Gutiérrez, A. (2014). Análisis documental de noticias de prensa en sistemas de información factual. *Revista Española de Documentación Científica*, *37*(2), e046. http://doi.org/10.3989/redc.2014.2.1094

García, R., Perdrix, F., & Gil, R. (2006). Ontological Infrastructure for a Semantic Newspaper. En *In "Semantic Web Annotations for Multimedia Workshop, SWAMM 2006". 15th World Wide Web Conference*.

González Cristóbal, J. C., Villena Román, J., Bueno Carrillo, F. J., García Serrano, A. M., Ruiz Cristina, A., & Martínez Fernández, P. (2002). OmniPaper : acceso inteligente a periódicos europeos. Recuperado de http://rua.ua.es/dspace/handle/10045/1752

Google Inc. (2014a). Creating a Google News Sitemap. Retrieved November 30, 2014, from https://support.google.com/news/publisher/answer/74288?hl=en

Google Inc. (2014b). Learn about sitemaps. Retrieved February 12, 2014, from https://support.google.com/webmasters/answer/156184?hl=en

Guerrillot, S. (2006). Use of semantic technologies at Agence France-Presse (AFP). Presented at the Semantic Technology Conference, San José. Recuperado de http://ceur-ws.org/Vol-194/paper8.pdf

Heravi, B. R., & McGinnis, H. (2013). A Framework for Social Semantic Journalism. En *First International IFIP Working Conference on Value-Driven Social & Semantic Collective Intelligence (VaSCo)*. Paris, France. Recuperado de http://members.deri.ie/~bahher/Publications/A%20Framework%20for%20Social%20Semantic%20Journalism%20Final.pdf

Hickson, I. (2013, October 29). HTML Microdata: W3C Working Group Note 29 October 2013. Recuperado de www.w3c-prg/TR/microdata/

Hillman, D. (2005, November 7). Using Dublin Core (DCMI Recommended Resource). Recuperado de http://dublincore.org/documents/usageguide/

Iannella, R., & McKinney, J. (2014). vCard Ontology - for describing People and Organizations. W3C Interest Group Note 22 May 2014. Recuperado de http://www.w3. org/TR/vcard-rdf

IPTC (International Press Telecommunications Council). (2009). *NewsML G2. Specification Version 2.4. Power Conformance Level* (No. Document Revision 1). Recuperado de https://www.iptc.org/std/NewsML-G2/2.4/specification/NewsML-G2_2.4-spec-PCL.pdf

IPTC (International Press Telecommunications Council). (2014). NITF News Industry Text Format. Recuperado December 17, 2014, de https://www.iptc.org/site/News_Exchange_Formats/NITF/

Jiménez Cano, R., & Abad Liñán, J. M. (2015, May 13). "The New York Times", "The Guardian" y la BBC publican directamente en Facebook desde hoy. *El País*. San Francisco / Madrid. Recuperado de http://tecnologia.elpais.com/tecnologia/2015/05/13/actualidad/1431490102_473389.html

Jokela, S., Turpeinen, M., Kurki, T., Savia, E., & Sulonen, R. (2001). The role of structured content En a personalized news service. *Acta Polytechnica Scandinavica Mathematics and Computing Series*, (114), XX–XXI. Recuperado de http://www.computer.org/csdl/proceedings/hicss/2001/0981/07/09817044.pdf

Kallipolitis, L., Karpis, V., & Karali, I. (2012). Semantic search En the World News domain using automatically extracted metadata files. *Knowledge-Based Systems*, *27*, 38–50. http://doi.org/10.1016/j.knosys.2011.12.007

Kodama, M., Ozono, T., Shintani, T., & Aosaki, Y. (2008). Realizing a news value markup language for news management systems using newsML. En F. Xhafa & L. Barolli (Eds.), *2nd International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 249–255). Los Alamitos: Ieee Computer Soc. http://doi.org/10.1109/CISIS.2008.70

Lamarca Lapuente, M. J. (2013, August 12). *Hipertexto, el nuevo concepto de documento en la cultura de la imagen: Metadatos en HTML* (tesis). Universidad Complutense de Madrid, Madrid. Recuperado de http://www.hipertexto.info/documentos/meta_html.htm

Mannens, E., Coppens, S., De Pessemier, T., Dacquin, H., Van Deursen, D., De Sutter, R., & Van de Walle, R. (2013). Automatic news recommendations via aggregated profiling. *Multimedia Tools and Applications*, *63*(2), 407–425. http://doi.org/10.1007/s11042-011-0844-8

Mannens, E., Troncy, R., Braeckman, K., Van Deursen, R. D., Van Lancker, V. W., De Sutter, R., & Van De Walle, R. (2009). Automatic metadata enrichment En news production. En *10th International Workshop on Image Analysis for Multimedia Interactive Services* (pp. 61–64). New York. http://doi.org/10.1109/WIAMIS.2009.5031432

Martínez-Fernández, J. L., García-Serrano, A., Martínez, P., & Villena, J. (2004). Automatic Keyword Extraction for News Finder. En A. Nürnberger & M. Detyniecki (Eds.), *Adaptive Multimedia Retrieval* (pp. 99–119). Springer Berlin Heidelberg. Recuperado de http://link.springer.com/chapter/10.1007/978-3-540-25981-7_7

McGee, M. (2012, September 19). Google Announces News Keywords Meta Tag For Publishers. Recuperado de http://searchengineland.com/google-announces-news-keywords-metatag-133759

Mueller, E. T. (2000). Making news understandable to computers. *arXiv*. Recuperado de http://arxiv.org/html/cs/0003001

Nies, T. de, D'heer, E., Coppens, S., Van Deursen, D., Mannens, E., & Van de Walle, R. (2012). Bringing newsworthiness into the 21st century. En *Web of Linked Entities, Workshop proceedings* (pp. 106–117). Recuperado de http://ceur-ws.org/Vol-906/paper11.pdf

Open Web Fundation (OWF). (2014, October 20). The Open Graph protocol. Retrieved December 3, 2014, from ogp.me

Paepen, B. E. (2002). Omnipaper: Bringing electronic news publishing to a next level using XML and Artificial Intelligence. En J. A. Carvalho, A. Hübler, & A. A. Baptista (Eds.), *Proceedings of the 6th International ICCC/IFIP Conference on Electronic Publishing*. Karlovy Vary, Czech Republic: VWF Berlin. Recuperado de http://elpub.scix.net/cgi-bin/works/Show?02-29

Pastor-Sánchez, J. A. (2011). *Tecnologías de la Web Semántica* (Edición: 1). Barcelona: Editorial UOC, S.L.

Pastor-Sánchez, J. A., Orduña-Malea, E., & Saorín Pérez, T. (2013). Marcado semántico automático en gestores de contenidos: integración y cuantificación. *El Profesional de la Informacion*, *22*(5), 381 – 391. http://doi.org/10.3145/epi.2013.sep.02

Pellegrini, T. (2012). Semantic metadata En the news production process: achievements and challenges. En *Proceedings of the 16th International Academic MindTrek Conference 2012* (pp. 125–133). Tampere; Finland: ACM Press. http://doi.org/10.1145/2393132.2393158

Pereira, T., & Baptista, A. A. (2003). Omnipaper: descrição de recursos de notícias digitais em RDF. En J. C. Ramalho, Pedro Rangel Henriques, G. R. Librelotto, & G. V. Arnold (Eds.), *XML, aplicações e tecnologias associadas*. Braga, Portugal: Universidade do Minho. Recuperado de http://repositorium.sdum.uminho.pt/handle/1822/283

Pereira, T. S. M., & Baptista, A. A. (2009). The instantiation of OmniPaper RDF prototype En the context of scientific publications. *Electronic Library*, *27*(5), 767–778. http://doi.org/10.1108/02640470910998506

Rubio Lacoba, M. (2012). Nuevas destrezas documentales para periodistas: el vocabulario colaborativo del diario El País. *Trípodos*, *31*, 65–78. Recuperado de http://www.raco.cat/index.php/Tripodos/article/view/262073/349255

Saias, J., & Quaresma, P. (2006). A proposal for an ontology supported news reader and question-answer system. En S. Oliveira Rezende (Ed.), *Proceedings of International Joint Conference, 10th IBERAMIA, ICMC-USP*. Ribeirao Preto, Brazil. Recuperado de http://ceur-ws.org/Vol-199/wonto-05.pdf

Saleh, L. M. B., & Al-Khalifa, H. S. (2009). AraTation: An Arabic semantic annotation tool (pp. 447–451). http://doi.org/10.1145/1806338.1806421

Sánchez-fernández, L., Bernardi, A., & Fuentes, M. (2005). An experience with Semantic Web technologies En the news domain. En *Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005. The 2nd European Workshop on the (Ref. No. 2005/11099)* (pp. 455–459). London.

Schranz, M., Dustdar, S., & Platzer, C. (2005). Building an Integrated Pan-European News Distribution Network. En L. M. Camarinha-Matos, H. Afsarmanesh, & A. Ortiz (Eds.), *Collaborative Networks and Their Breeding Environments* (pp. 587–596). Springer US. Recuperado de http://link.springer.com/chapter/10.1007/0-387-29360-4_62

Shoval, P., Maidel, V., & Shapira, B. (2008). An Ontology-Content-based Filtering Method. *International Journal ITA*, *15*(4), 303–314. Recuperado de http://sci-gems.math.bas.bg:8080/jspui/handle/10525/88

Silva, D. L. da, Souza, R. R., & Almeida, M. B. (2008). Ontologias e vocabulários controlados: comparação de metodologias para construção. *Ciência da Informação*, *37*(3), 60–75. Recuperado de http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652008000300005

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014, January 16). JSON-LD 1.0: A JSON-based Serialization for Linked Data. Recuperado de http://www.w3.org/TR/json-ld/

Tenenboim, L., Shapira, B., & Shoval, P. (2008). Ontology-based classification of news En an electronic newspaper (pp. 89–97). Presented at the INFOS 2008: Intelligent Information and Engineering Systems, Varna, Bulgaria. Recuperado de http://sci-gems.math.bas.bg/jspui/bitstream/10525/1035/1/IBS-02-p12.pdf

Toffler, A. (1980). *The third wave: The classic study of tomorrow*. New York, NY, USA: Bantam.

Troncy, R. (2008). Bringing the IPTC News Architecture into the Semantic Web. En A. Sheth, S. Staab, M. Paolucci, D. Maynard, T. Finin, & T. Krishnaprasad (Eds.), *Semantic Web - Iswc 2008* (Vol. 5318, pp. 483–498). Berlin: Springer-Verlag Berlin. http://doi.org/10.1007/978-3-540-88564-1-31

Wong, W., Liu, W., & Bennamoun, M. (2010). An ontology-based interface for improving information exploration. En *Proceedings of the first international workshop on Intelligent visual interfaces for text analysis* (pp. 29–32). New York, USA: ACM. http://doi.org/10.1145/2002353.2002364

Yaginuma, T., Pereira, T., & Baptista, A. A. (2003a). Design of metadata elements for digital news articles En the omnipaper project. En S. M. de Souza Costa, J. A. Carvalho, A. A. Baptista, & A. C. Santos Moreira (Eds.), *From information to knowledge: 7th ICCC/IFIP International Conference on Electronic Publishing* (pp. 132–139). Minho, Portugal: Universidade do Minho. Recuperado de http://repositorium.sdum.uminho.pt/handle/1822/170

Yaginuma, T., Pereira, T., & Baptista, A. A. (2003b). Metadata elements for digital news resource description. En *Proceedings CLME'2003 - 3º Congresso Luso-Moçambicano de Engenharia* (pp. 1317–1326). Maputo. Recuperado de http://repositorium.sdum.uminho.pt/handle/1822/279

Anexo 1

Especificaciones	*Citados en bibliografía*	Nº
ANPA 89-3	Sánchez-Fernández et al., 2005; Fernández-García et al., 2007; Pellegrini, 2012	3
Dublin Core	Yaginuma, Pereira y Baptista, 2003; Pereira y Baptista, 2003a; Pereira y Baptista, 2003b; Pereira et al., 2003; Ariza y Baptista, 2004; Martínez-Fernández et al., 2004; Schranz, 2005; Fernández et al., 2006; Allen et al., 2007; Fernández-García et al., 2007; Fernández et al., 2010; Moreira et al., 2012; Pellegrini, 2012	13
hNews	Pellegrini, 2012	1
IIM	García et al., 2006; Fernández-García et al., 2007; Tse-Yi y Teong, 2008; Fernández et al., 2010	4
IPTC 7901	Fernández-García et al., 2007	1
NewsML (NAR)	Pereira y Baptista, 2003a; Pereira y Baptista, 2003b; Pereira et al., 2003; Yaginuma, Pereira y Baptista, 2003; Martínez-Fernández et al., 2004; Yaginuma, Pereira y Baptista, 2004; Sánchez-Fernández et al., 2005; Schranz, 2005; Schranz et al., 2005; Fernández et al., 2006; García et al., 2006; Guerrillot, 2006; Fernández-García et al., 2007; García et al., 2008; Kasper et al., 2008; Kodama et al., 2008; Tenenboim et al., 2008; Troncy, 2008; Tse-Yi y Teong, 2008; Biyun et al., 2009; Mannens et al., 2009; Fernández et al., 2010; Kallipolitis et al., 2012; Mannens et al., 2013	24
News Value Markup Language (NVML)	Kodama et al., 2008	1
NewsCodes (SRS)	Pereira y Baptista, 2004; Sánchez-Fernández et al., 2005; Fernández et al., 2006; García et al., 2006; Fernández-García et al., 2007; García et al., 2008; Kasper et al., 2008; Shoval et al., 2008; Tenenboim et al., 2008; Troncy , 2008; Tse-Yi y Teong, 2008; Heravi y McGinnis., 2013; Kallipolitis et al., 2012; Heravi et al., 2012; Chy et al., 2014	15
NITF	Mueller, 2000; Pereira y Baptista, 2003a; Pereira y Baptista, 2003b; Pereira et al., 2003; Yaginuma, Pereira y Baptista, 2003; Ariza y Baptista, 2004; Martínez-Fernández et al., 2004; Yaginuma, Pereira y Baptista, 2004; Sánchez-Fernández et al., 2005; Schranz, 2005; Schranz et al., 2005; Fernández et al., 2006; García et al., 2006; Fernández-García et al., 2007; Fernández et al., 2010; García et al., 2008; Tse-Yi y Teong, 2008; Biyun et al., 2009; Pellegrini, 2012	19
PRISM	Fernández-García and Sánchez-Fernández, 2004; Fernández et al., 2006; Fernández-García et al., 2007; Tse-Yi y Teong, 2008; Fernández et al., 2010	5
rNews	Heravi et al., 2012; Pellegrini, 2012; Heravi y McGinnis, 2013	2
RSS	Tenenboim et al., 2008; Fransincar et al., 2009; Pereira y Baptista, 2009	3
vCard	Pereira y Baptista, 2003a; Pereira y Baptista, 2003b	2
XMLNews	Mueller, 2000; Yaginuma, Pereira y Baptista, 2003; Yaginuma, Pereira y Baptista, 2004; Fernández-García et al., 2007	4

Tabla 3. Especificaciones de metadatos referencias por la bibliografía analizada

Especificaciones	*Referencias*	Nº
Etiquetas meta (html)	Süddeutsche Zeitung; Le Monde; The Daily Telegraph; Financial Times; The Economist; The New York Times; The Wall Street Journal; O Globo; El Universal; Pravda; Nigerian Tribune; Independent Online; China Daily; The Times of India; Asahi Shimbun; Arab News; Gulf News; Yedioth Aharonot; Today's Zaman; The Australian Financial Review; The Daily News Egypt	21
Dublin Core	Le Monde; The Telegraph	2
hNews	Nigerian Tribune; Independent Online; Asahi Shimbun	3
schema.org	Süddeutsche Zeitung; Le Monde; The Daily Telegraph; The New York Times; The Wall Street Journal; O Globo; Today's Zaman; The Daily News Egypt	8
Open Graph Protocol	Süddeutsche Zeitung; Le Monde; The Daily Telegraph; Financial Times; The Economist; The New York Times; The Wall Street Journal; O Globo; El Universal; Pravda; Nigerian Tribune; The Times of India; Asahi Shimbun; Arab News; Gulf News; Yedioth Aharonot; Today's Zaman; The Australian Financial Review; The Daily News Egypt	19
Twitter Card	Süddeutsche Zeitung; Le Monde; The Daily Telegraph; The Economist; The New York Times; The Wall Street Journal; O Globo; El Universal; The Times of India; Gulf News; Today's Zaman; The Australian Financial Review; The Daily News Egypt	13

Tabla 4. Especificaciones de metadatos aparecidas en los códigos fuente de los periódicos de la muestra

Anexo 2

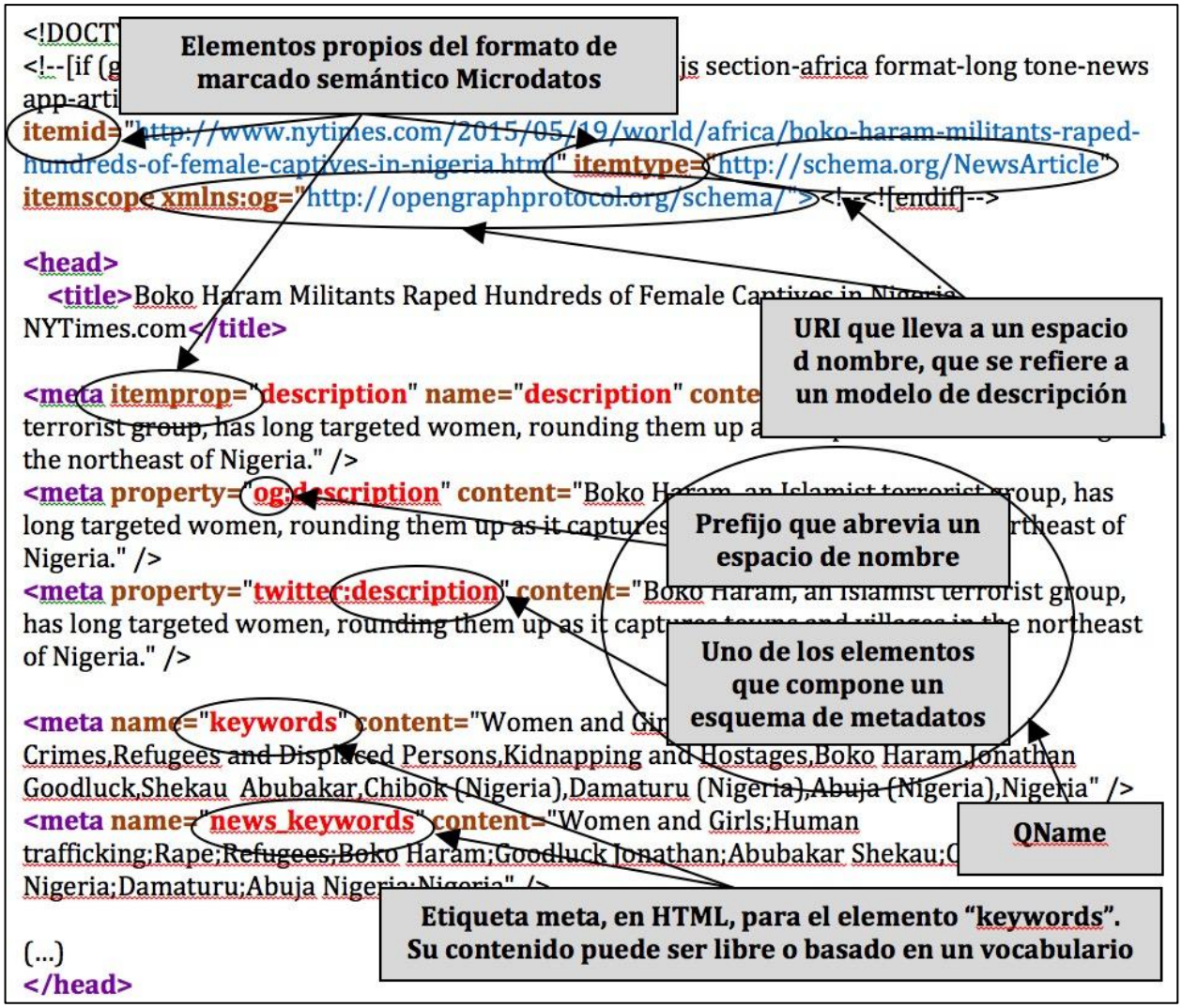

Figura 1. Parte del código fuente de una noticia de The New York Times, de 18 de mayo de 2015, Boko Haram Militants Raped Hundreds of Female Captives En Nigeria