

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/315210943>

Análisis de metadatos de noticias para la extracción de información del código fuente. El software MetadatosHTML

Article in Information Research an international electronic journal · March 2017

CITATIONS

2

READS

638

5 authors, including:



María-José Baños-Moreno
University of Murcia

30 PUBLICATIONS 57 CITATIONS

SEE PROFILE



Eduardo Ribeiro Felipe
Federal University of Minas Gerais

25 PUBLICATIONS 29 CITATIONS

SEE PROFILE



Juan Antonio Pastor-Sánchez
University of Murcia

72 PUBLICATIONS 297 CITATIONS

SEE PROFILE



Gercina Angela de Lima
Federal University of Minas Gerais

142 PUBLICATIONS 558 CITATIONS

SEE PROFILE

ANÁLISIS DE METADATOS DE NOTICIAS PARA LA EXTRACCIÓN DE INFORMACIÓN DE CÓDIGO FUENTE. EL SOFTWARE METADADOSHTML

RESUMEN

Introducción. Los objetivos de este trabajo son 1) determinar qué esquemas se utilizan para título, resumen, palabras clave, autoría y periódico en prensa; 2), conocer qué pautas siguen los periódicos en la implementación de dichos esquemas; 3) averiguar cómo esto afecta a la extracción de valores de datos.

Metodología. Para ello, se define una muestra de diarios y se analiza su código fuente, identificando esquemas utilizados y patrones de uso. Esto permite extraer valores de dato utilizando la aplicación MetadadosHTML.

Resultados. Se han detectado esquemas estándar, ad hoc y propios de los periódicos. Se han hallado diversas prácticas, como valores agrupados en una misma línea de código o por separado; ruido en un valor y errores al referir los nombres de los atributos de esquemas estándar.

Conclusiones. Ello dificulta la extracción de información, por cuanto es necesario no sólo conocer esquemas y atributos, hay que saber qué prácticas sigue cada medio. Los errores al referir los nombres de los atributos impediría la extracción de valores de dato siguiendo los esquemas. Es necesario avanzar en el uso de estándares. También resulta imprescindible la adopción de buenas prácticas en el uso de esquemas estándar, ad hoc y propietarios.

PALABRAS CLAVE

Metadatos, esquemas de metadatos, periódicos, palabras clave, extracción de información, MetadadosHTML, HTML, Open Graph Protocol, Twitter Cards, Schema.org, The New York Times

INTRODUCCIÓN: METADATOS Y ESQUEMAS EN PRENSA

Los metadatos son el conjunto de atributos y valores de dato que representan un recurso, facilitando su identificación (Abbud Grácio & Fadel, 2010, pp. 10-11; AENOR, 2012; Hillman, 2005; Kallipolitis, Karpis, & Karali, 2012; Yaginuma, Pereira, & Baptista, 2003a). Más específicamente: *‘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, 2011, p. 22). Cuando esos metadatos se organizan bajo una estructura de elementos comunes (Pereira & Baptista, 2004; Yaginuma, Pereira, & Baptista, 2003b) podemos hablar de esquemas de metadatos. En (Baños-Moreno, Felipe, Pastor-Sánchez, Martínez-Béjar, Rodrigo, & Lima, 2015) se analizan qué esquemas existen en el ámbito periodístico desde dos perspectivas: 1) bibliografía especializada; y 2) códigos fuente de una muestra de periódicos. En esta publicación se destaca, entre otras cosas: 1) La existencia de diversos esquemas; 2) La ausencia de esquemas especializados en prensa en los códigos analizados; 3) Escasa referencia en la bibliografía a los esquemas mostrados en los códigos fuente.

Si bien *‘ya no se piensa tanto en las soluciones tecnológicas de edición y publicación como en el consumo de datos publicados’* (Pastor Sánchez, 2016), la diversidad de estándares y los diferentes grados de implantación dificultan la definición unificada de recursos accesibles (González Cristóbal et al., 2002) y, con ello, la reutilización de información disponible en web. Este trabajo profundiza en la investigación descrita en (Baños-Moreno et al., 2015), poniendo el foco de atención en: título, resumen, palabras clave, autor/a y nombre del medio. Los objetivos son 1) determinar qué esquemas se utilizan para estos datos; 2), conocer qué pautas siguen los periódicos en la implementación de dichos esquemas, a través de etiquetas meta y atributos en otras etiquetas para; y 3) averiguar cómo esto afecta a la extracción de valores de datos (*data values*). Para ello, se analizan los códigos fuente de un conjunto de noticias, identificando esquemas de metadatos y sus patrones de uso. Para la extracción, se emplea una aplicación

desarrollada *ad hoc* con el propósito de facilitar la reutilización de esta información declarada en el código fuente de noticias para el modelado de una ontología de dominio. El motivo por el que se utiliza información de actualidad es su abundancia en Internet y permanente actualización, lo que asegura una continua renovación de términos, conceptos y relaciones entre estos (Baños-Moreno, Pastor-Sánchez, & Martínez-Béjar, 2013).

METODOLOGÍA

Se definió una muestra de periódicos (ver Tabla 1) utilizada en trabajos previos (Baños-Moreno et al., 2015; Baños-Moreno, 2013; Baños-Moreno et al., 2013), para el cual se dividió el mundo en 9 zonas, de acuerdo a sus características 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 en cada país, de acuerdo a datos de accesos Web y compra de edición impresa de 4International Media & Newspaper (4IMN).

Zona del mundo	País	Periódico	Código
Europa	Alemania	Süddeutsche Zeitung	101
	Francia	Le Monde	111
	Reino Unido	The Daily Telegraph; Financial Times; The Economist	121 122 123
Norteamérica	Estados Unidos	The New York Times; The Wall Street Journal	201 202
Latinoamérica	Brasil	O Globo	301
	México	El Universal	311
Ex-Repúblicas rusas	Rusia	Pravda	401
África Subsahariana	Nigeria	Nigerian Tribune	501
	Sudáfrica	Independent Online	511
Asia	China	China Daily	601
	India	The Times of India	611
	Japón	Asahi Shimbun	621
Oriente	Arabia Saudí	Arab News	701
	Emiratos Árabes Unidos	Gulf News	711
	Israel	Yedioth Aharonot	721
	Turquía	Today's Zaman	731
Oceanía	Australia	The Australian Financial Review	801
Norte de África	Egipto	The Daily News Egypt	901

Tabla 1. Periódicos utilizados a los que se ha accedido en este trabajo

Los códigos de la columna *Código* de la Tabla 1 identifican los periódicos por: zona del mundo (primera cifra), país (segunda cifra) y periódico (tercera cifra). Estos códigos son utilizados a lo largo de este documento para simplificar el texto.

Para el análisis, se accedió a tres noticias por periódico en días diferentes durante enero y febrero de 2016. Para la identificación de metadatos, se revisó el código fuente de los diarios, tanto de forma manual como utilizando el servicio de validación de marcado de contenidos web del W3C (*Markup Validation Service*).

DATO	VALOR DE DATO
TÍTULO	<i>Jeb Bush Adds a Weapon (His Brother) Despite Worries It Could Misfire</i>
RESUMEN	<i>Former President George W. Bush is set to take a stage in South Carolina on Monday with his brother, in an appearance that offers a glimpse into the complicated dynamics of a family dynasty</i>
PALABRAS CLAVE	<i>Presidential Election of 2016, Primaries and Caucuses, Republican Party, Bush Jeb, Bush George W, South Carolina</i>
AUTOR/A	<i>Ashley Parker</i>
NOMBRE DEL MEDIO	<i>The New York Times</i>

Figure 1. Datos y valores de datos para noticia redactada por (Parker, 2016)

Se tuvo en cuenta para ello tanto las etiquetas meta como los atributos que aparecen en el cuerpo de la página web (<body>), en función de las diferentes formas que pueden adoptar. En este sentido, un metadato está compuesto por pares de datos y valores de datos, de la forma en que se muestra en la Figura 1.

A partir del análisis, se hallaron tres tipos de esquemas: estándares, ad hoc y propietarios, descritos en el apartado Resultados. Además, se identificaron una serie de características e incidencias para título, resumen, palabras clave, autor/a y nombre del medio.

Teniendo en cuenta estos aspectos, se diseñó un algoritmo para la extracción de información por parte de la aplicación desarrollada al efecto, MetadadosHTML. Para ello, se utilizó una muestra de más de 5500 noticias extraídas anteriormente, cuya URLs fueron almacenadas en una hoja de cálculo.

No se tiene en cuenta en esta categorización aquellos atributos que tienen como único fin la separación en secciones de una página web. Por ejemplo, <div id="Title_e"> ó <div id="article_head">, utilizados por China Daily y Asahi Shimbun, respectivamente. Tampoco la información en scripts, aunque algunos medios aplican algunos de los esquemas nombrados anteriormente, como schema.org (The Wall Street Journal o Pravda.ru, entre otros).

RESULTADOS

Los esquemas identificados se pueden clasificar en tres categorías: estándares, *ad hoc* y propietarios:

1. **Esquemas externos y reconocidos internacionalmente (esquemas estándar)**, como Open Graph Protocol (en adelante OG), Dublin Core (DC), Twitter Cards (TC), hAtom o schema.org (SC). En este grupo, también se incluyen algunas etiquetas meta de HTML, donde el atributo *name* indica el metadato y *content* el valor de dicho metadato. Existen etiquetas meta que se refieren al contenido del documento y que pueden considerarse ‘*nombres de metadatos estándar*’ (Hickson et al., 2014): <meta name="author" content="...">, <meta name="description" content="..."> y <meta name="keywords" content="...">. No incluye <meta name="title" content="...">.

En el caso de utilizar XHTML, uno de los usos de los atributos de metadatos, como *property*, ‘es con elementos que representan el contenido del documento, ya que la misma cadena literal puede utilizarse para especificar tanto el contenido del documento como los

metadatos’ (Axelsson et al., 2006). Así, nos encontramos con `<span property="dc:title" content="..." class="...">`, de Nigerian Tribune (501), que aquí utiliza el esquema de Dublin Core para referirse al título de una noticia.

Por otro lado, schema.org (SC) es un esquema que usa microdatos para, mediante el atributo *itemprop*, referirse a una meta-información (Baños-Moreno et al., 2015). Este atributo se inserta habitualmente en las etiquetas del cuerpo (`<body>`), por ejemplo: `<h1 itemprop="headline">...</h1>`, aunque también se ha aplicado como atributo de etiquetas meta, como se muestra en algunos ejemplos en (Leithead, Pfeiffer, Navara, & O’Connor, 2012) y en las siguientes etiquetas, de una noticia de The New York Times (201), de la que hemos suprimido los valores de dato:

```
<meta itemprop="genre" content="..." />
<meta itemprop="identifier" name="articleid" content="..." />
<meta itemprop="usageTerms" name="usageTerms" content="..." />
<meta itemprop="inLanguage" content="..." />
```

Figure 2. Etiquetas meta de The New York Times (201) usando el atributo *itemprop*, de schema.org

Algo similar ocurre con el microformato hAtom 0.1. (Janes, 2013), utilizado por The Daily News Egypt (901), por ejemplo: `<h1 class="entry-title">...</h1>`, que inserta meta información en las etiquetas de `<body>`.

Para algunos de estos esquemas, se han desarrollado validadores que analizan la corrección de las etiquetas y el uso de atributos, como Markup Validation Service, de la W3C, Card validator, de Twitter Cards, ó Structured Data Testing Tool, de Google.

El siguiente gráfico resume los esquemas estándar utilizados para los datos analizados: título, resumen, palabras clave, autor y nombre del medio.

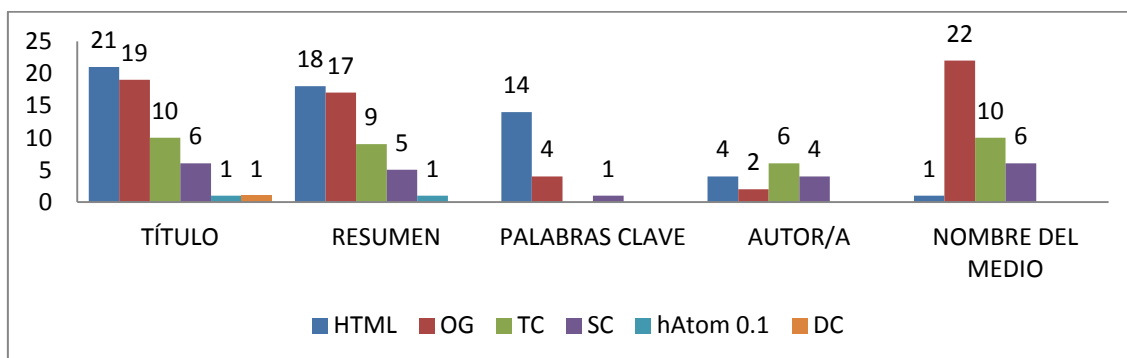


Gráfico 1. Frecuencia de esquemas estándar en la muestra de periódicos por tipos de metadato

Como puede apreciarse, Open Graph Protocol (OG) y HTML son los estándares más extendidos, especialmente en cuanto a título, resumen y palabras clave. Como se indica más adelante (ver Tabla 11), la cifra más elevada (22) para el nombre del medio, en OG, se debe al uso de este esquema en dos atributos: *site_name* y *article:publisher*, como también ocurre con schema.org (SC).

2. **Esquemas externos y desarrollados para un propósito concreto (esquemas *ad hoc*)**, utilizados por unos pocos periódicos, por lo que su uso no es extendido. Es el caso de Sailthru, y su tecnología propietaria Horizon, que ‘realiza un seguimiento del comportamiento en el sitio, móvil y del correo electrónico para desarrollar un perfil de interés único para cada usuario final’ (Sailthru, 2013). Es utilizado, al menos, por The Economist. La Figura 2 muestra tres líneas de código, de tres noticias diferentes, para los tres esquemas propietarios identificados en la muestra.

Por otro lado, existe un esquema desarrollado por la empresa **Bi3d** (Believe in 3D), que ‘desarrolla la próxima generación de interfaces humano-ordenador interactivos para la industria del marketing y el entretenimiento’ (BI3D, 2013), aunque parece que en este momento no está operativa. Su esquema es utilizado por El Universal (311).

```
<meta name="sailthru.flytitle" content="..." />
<meta name="bi3dArtTitle" content="..." />
<meta name="parsely-page" content="{"title": "...", "authors": ["..."], "tags": ["..."]}"/>
```

Figure 3. Etiquetas ad hoc de los esquemas de metadatos de Sailthru, Bi3d y Parse.ly

Parse.ly es otra plataforma de análisis diseñada específicamente para los editores en línea y ofrece tres formas de incorporar estos datos a la página web: *JSON-LD*; *repeated metadata*; *parsely-page*. Parse.ly sigue el esquema NewsArticle de schema.org. The Daily Telegraph (121) utiliza la opción *parsely-page*, ya obsoleta y reemplazada por JSON-LD pero que aún puede ser rastreada (Parse.ly, s.f.). La siguiente figura muestra una previsualización de los metadatos definidos en la etiqueta `<meta name="parsely-page" content="...">` a partir de la validación de la noticia (Kirkup, 2016).



How EU referendum polls are getting it wrong amid a 'cloud of uncertainty'

PUBLISHED ON	AUTHOR	SECTION
Feb 1, 2016 3:42:40 PM	James Kirkup	News

TAGS: Story, Europe, EU Referendum, Comment, James Kirkup, Comment

EXCERPT:
Polls of public opinion on European Union membership may be seriously flawed and should be read with great caution. Errors and uncertainties in the way pollsters attempt to capture the views of the electorate mean their surveys could badly miss the referendum result. Despite ad...

Figure 4. Datos extraídos por el validador de Parse.ly a partir de (Kirkup, 2016)

Webtrends, otra empresa de análisis web, escanea el código fuente buscando conocidos parámetros de consulta Webtrends para añadir a los datos. En la muestra, The Daily Telegraph recurre al espacio de nombre **DCSext**, seguido por punto y el nombre del atributo del que coleccionar datos. Esto se observa, por ejemplo, en la etiqueta `<meta name="DCSext.author" content="..." />` ó `<meta name="DCSext.cmsSect" content="..." />`, donde “cmsSect” hace referencia a las palabras clave y coincide con los valores de dato que este periódico asigna a cualquier noticia en `<meta property="article:tag" content="..." />`.

En este grupo también podría incluirse la etiqueta meta definida por **Google**, `<meta name="news_keywords" content="...">` para que los periódicos pudieran indicar cuáles son las palabras más relevantes de una noticia (Google Inc., 2016).

El siguiente gráfico muestra los esquemas *ad hoc* utilizados para los datos analizados: título, resumen, palabras clave, autor y nombre del medio.

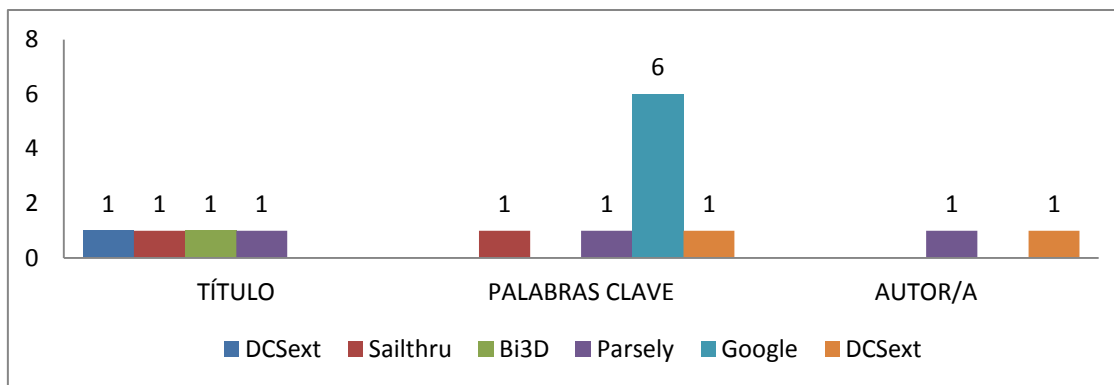


Gráfico 2. Frecuencia de esquemas ad hoc en la muestra de periódicos por tipos de metadato

Este tipo de esquemas es mucho menos frecuente que el tipo anterior, destacando especialmente la etiqueta meta de Google, *news_keywords*.

3. **Esquemas propios, desarrollados por el periódico (esquemas propietarios)** para la estructuración y categorización de sus artículos. El caso más representativo es The New York Times (ver Tabla 2).

Campo	Significado	Correspondencia
hdl	Headline	Titular
byl	Byline	Autor del artículo
lp	Lead Paragraph	Resumen
des	Subjects / Descriptive Terms	Palabras clave
per	People	Identificador para personas
org	Organizations	Identificador para organizaciones
geo	Geographic Locations	Identificador para lugares geográficos

Tabla 2. Algunos elementos del esquema utilizado por The New York Times. Fuente: (Harris, 2007)

En ocasiones, se pueden incrustar datos no visibles y personalizados (*embedded custom non-visible data*) gracias a HTML5. Se trata de atributos sin espacio de nombre (*namespace*) cuyo *name* comienza con la cadena *data-* y tienen, al menos un carácter después del guión (Hickson et al., 2014), como en The New York Times (201) y The Wall Street Journal (202) (ver Figura 4). Puesto que son elementos personalizables, en este trabajo se clasifican como esquemas propietarios.

<pre><div id="..." aria-label="..." role="..." class="..." data-shares="..." data-url="..." data-title="..." data-author="By ..." data-media="..." data-description="..." data-publish-date="..."></pre>	<pre><li class="..." data-id="..." data-headline="..." data-url="..." data-authors="..." data-date="..." data-summary="..." title="..." aria-label="..." ></pre>
--	---

Figure 5. Atributos *data-* para artículos en The New York Times y The Wall Street Journal, respectivamente

Este tipo de esquema es el menos frecuente, con presencia solamente en dos medios, The New York Times (201) y The Wall Street Journal (202), con dos esquemas propios por cada medio.

A continuación, se muestran las particularidades de las líneas de código fuente para cada uno de los datos objeto de este análisis.

TITULAR DE LA NOTICIA

Para el titular, se utiliza, en todos los casos, la etiqueta <title>, que se refiere al título de una página web cualquiera. Por el contrario, <meta name="title" content="..." /> está menos extendida. Además, se han identificado esquemas de los tres tipos indicados, aunque con un marcado incorrecto. Por ejemplo, en esta línea: <h1 itemprop="headline name">, de The Economist, aparece el atributo *headline name*, que no reconoce schema.org. También se han hallado atributos con diferente denominación para referirse a este dato, como *hdl*, *headline* o *article.origheadline* (ver Tabla 3).

	Línea de código	Tipo de esquema	N
1	<title>...</title>	Estándar (HTML)	21
2	<meta property="og:title" content="..." />	Estándar (OG)	19
3	<meta name="twitter:title" content="..." />	Estándar (TC)	10
4	<h1 itemprop="headline">...</h1> <h1 itemprop="headline name">...</h1> <h3 itemprop="headline" class="headline">...</h3> <meta itemprop="name" content="..." />	Estándar (SC)	6
5	<h1 class="entry-title">...</h1>	Estándar (hAtom 0.1)	1
6		Estándar (DC)	1
7	<meta name="sailthru:flytitle" content="..." />	Ad hoc (Sailthru)	1
8	<meta name="bi3dArtTitle" content="..." />	Ad hoc (Bi3D)	1
9	<meta name="parsely-page" content="{\"title\": \"...\" ...}\" />	Ad hoc (Parse.ly)	1
10	<meta name="hdl" content="..." />	Propietario (201)	1
11	<div id="..." aria-label="..." role="..." class="..." data-title="...">	Propietario (201)	1
12	<meta name="article.origheadline" content="..." /> <meta name="article.headline" content="..." />	Propietario (202)	1
13	<li class="..." data-headline="...">...	Propietario (202)	1

Tabla 3. Formas de identificar el titular en los medios de la muestra. Fuente: Elaboración propia a partir de los códigos fuente de los diarios de la muestra

En ocasiones, además, el texto incluye ruido, información que no es propiamente el titular. Por ejemplo: <title>Jeremy Corbyn tries to force public schools to open up music, arts, sport facilities to state schoolchildren - Telegraph</title>.

RESUMEN O DESCRIPCIÓN

Este dato suele recoger un resumen del contenido de un documento. Cuando ese documento es una noticia, habitualmente la información aquí contenida es un subtítulo o bien la entradilla, ‘un resumen lo suficientemente completo y autónomo como para que el lector conozca lo fundamental de la noticia sólo con leerlo’ (Lajusticia, 2000).

Se han identificado esquemas de los tres tipos indicados (ver Tabla 4) pero, al igual que en el caso del titular, con un marcado incorrecto. Así, en <p class="article entry-summary">, de Süddeutsche Zeitung (101), el atributo *class="article entry-summary"* no se corresponde con el definido por el esquema de hAtom, del que parece provenir, siendo el correcto *class="entry-summary"*.

	Línea de código	Tipo de esquema	N
1	<meta name="description" content="..." />	Estándar (HTML)	18
2	<meta property="og:description" content="..." />	Estándar (OG)	17
3	<meta name="twitter:description" content="..." />	Estándar (TC)	9
4	<h2 class="..." itemprop="description">...</h2> <meta itemprop="description" name="..." content="..." />	Estándar (SC)	5
8	<p class="article entry-summary">...</p>	Estándar (hAtom 0.1)	1
5	<meta name="lp" content="..." />	Propietario (201)	1
6	<div id="..." aria-label="..." role="..." class="..." data-description="...">	Propietario (201)	1

7	<meta name="article.summary" content="..." />	Propietario (202)	1
8	<li class="..." data-summary="...">...	Propietario (202)	1

Tabla 4. Formas de identificar el resumen en los medios de la muestra. Fuente: Elaboración propia a partir de los códigos fuente de los diarios de la muestra

A veces, también el texto incluye ruido, información que no es propiamente el resumen. Por ejemplo, esta línea es utilizada para mostrar un resumen en el periódico El Universal. Sin embargo, lo primero que aparece es el perfil de Twitter del medio: <meta name="description" content="metropoli@eluniversal.com.mx ..." />.

PALABRAS CLAVE

Las palabras clave son términos que, en su conjunto, representan el contenido de un documento. En general, se utilizan dos formas para referirse a este elemento: *keywords* y *tag*.

En la siguiente tabla, se muestra el metadato a extraer, de qué esquema, en qué periódicos aparece y algunas notas relevantes.

	Línea de código	Tipo de esquema	N
1	<meta name="keywords" content="..." />	Estándar (HTML)	14
2	<meta property="article:tag" content="..." />	Estándar (OG)	4
3	<meta itemprop="keywords" content="..." />	Estándar (SC)	1
4	<meta name="sailthru.tags" content="..." />	Ad hoc (Sailthru)	1
5	<meta name="parsely-page" content="{..., "tags": ["..."]}"/>	Ad hoc (Parse.ly)	1
6	<meta name="news_keywords" content="..." />	Ad hoc (Google)	7
7	<meta name="DCSext.cmsSect" content="..." />	Ad hoc (DCSext)	1
8	<meta name="org" content="..." /> <meta name="des" content="..." /> <meta name="geo" content="..." /> <meta name="per" content="..." />	Propietario (201)	1

Tabla 5. Formas de identificar las palabras clave en los medios de la muestra. Fuente: Elaboración propia a partir de los códigos fuente de los diarios de la muestra

Para este dato, destacan algunos aspectos que influyen en la extracción de información a partir del código fuente:

1. **Campos repetibles:** Casi todos los medios incluyen en una misma línea el conjunto de palabras clave que representa el contenido de un artículo. Pero hay excepciones: algunos periódicos repiten la meta etiqueta con valores de dato diferentes, tantas veces como palabras clave refiera, como se muestra en las Tablas 6 y 7.

Línea de código	Repiten etiqueta	No la repiten	Total
<meta name="keywords" content="..." />	1	13	14
<meta property="article:tag" content="..." />	3	1	4
<meta name="org" content="..." /> <meta name="des" content="..." /> <meta name="geo" content="..." /> <meta name="per" content="..." />	201	---	1

Tabla 6. Periódicos que repiten las etiquetas meta para las palabras clave

2. **Mismo propósito, información diferente.** The New York Times (201) muestra las palabras clave en distintas etiquetas meta (ver Tabla 7, a partir de (Martin, 2016)). Como puede observarse, <meta name="news_keywords" content="Iowa" /> rompe el patrón de otras líneas que, de forma agrupada en una única línea o separada en varias, muestra la misma información.
3. **Escasa categorización:** The New York Times es el único diario que indica, utilizando un esquema propio, el tipo de palabras clave (ver Tabla 7): *des* → descriptores; *org* → organización; *per* → persona; *geo* → lugar.

Tipo de esquema	Línea de código fuente
Estándar (OG)	<pre> <meta property="article:tag" content="Presidential Election of 2016" /> <meta property="article:tag" content="Republican Party" /> <meta property="article:tag" content="Cruz, Ted" /> <meta property="article:tag" content="Primaries and Caucuses" /> <meta property="article:tag" content="Iowa" /> </pre>
Propietario	<pre> <meta name="des" content="Presidential Election of 2016" /> <meta name="org" content="Republican Party" /> <meta name="per" content="Cruz, Ted" /> <meta name="des" content="Primaries and Caucuses" /> <meta name="geo" content="Iowa" /> </pre>
Estándar (HTML)	<pre> <meta name="keywords" content="Presidential Election of 2016,Republican Party,Cruz Ted,Primaries and Caucuses,Iowa" /> </pre>
Ad hoc (Google)	<pre> <meta name="news_keywords" content="Iowa" /> </pre>

Tabla 7. Etiquetas meta de The New York Times (201) para mostrar las palabras clave

4. **Ruido en el valor de dato:** Como en los casos de titular y resumen, a veces se incluye en esta etiqueta meta información que no corresponde a este espacio. Por ejemplo:

- En `<meta name="news_keywords" content="VW, Matthias Müller, Volkswagen, Wirtschaft">`, Süddeutsche Zeitung (101) incluye el nombre del autor del artículo. También lo hace The Daily Telegraph (121)
- En `<meta name="keywords" content="israel, news, yedioth, ahronoth, english, ..., breaking news, jew, jews, www.ynetnews.com">` Yedioth Aharonot (721) ha incluido el nombre del medio, la sección de la noticia y la URL del medio, que no son palabras clave
- Arab News (701) repite la etiqueta `<meta name="keywords" content="...">`: la primera vez hace referencia a la sección en la que se ha clasificado la noticia; la segunda recoge las palabras clave (ver Figura 5)

```

<meta name="keywords" content="Middle-East"/>
<meta name="keywords" content="Jubail,Al-Ahsa,Al-Kharj,Taif,Qatif,oil,..., jobs in Saudi,work in saudi arabia, umrah,Hajj,نور عرب">

```

Figure 6. Etiquetas meta estándar (HTML) usadas por Arab News (7011) para sección y las palabras clave

5. **Código fuente con fines publicitarios.** Finalmente, se puede destacar la orientación a cuestiones publicitarias de algunos elementos de descripción de noticias, habitualmente mediante Javascripts. Este aspecto no se ha tenido en cuenta en este trabajo.

Todos los periódicos recogen en su código fuente líneas para palabras clave, con tres excepciones: Le Monde (111), Nigerian Tribune (501) e Independent Online (502).

AUTOR/A DE LA NOTICIA

Autor/a es la persona o personas que han escrito una noticia y en los códigos fuente suele aparecer con dos denominaciones: *author* ó *creator*, como puede observarse en la Tabla 8.

Para este dato, destacan estos aspectos:

- Mismo propósito, información diferente:** el valor de dato de la etiqueta meta no siempre se corresponde con el nombre de quien ha redactado el artículo. Por ejemplo, `<meta name="author" content="...">`, de Süddeutsche Zeitung (101), tiene siempre el mismo valor de dato: "Süddeutsche.de GmbH, Munich, Germany", esto es, nombre del medio. Algo similar sucede con `<meta name="twitter:creator">`, de The Economist (122), que indica como valor la sección en que se publica la noticia. Otros utilizan esta misma línea para hacer referencia al perfil de Twitter del medio
- Errores de marcado:** Como ocurría Süddeutsche Zeitung (101) y el atributo `class="article entry-summary"` mal marcado, The Daily Telegraph (121) indica mal el

atributo de OG para el objeto *article* en `<meta property="article:authorName" content="..." />`, en lugar de *article:author*.

3. **Dos atributos, mismo valor:** The New York Times (201) hace referencia a autor y creador en la misma línea, con el mismo valor ([Web Hypertext Application Technology Working Group (WHATWG)], 2016a): `<span ... itemprop="author creator" ...>`.
4. **Campos repetibles:** Cuando una noticia ha sido escrita por varios autores, habitualmente se repiten las líneas de código
5. Otros medios recogen este dato como una palabra clave de la noticia o bien lo que aparece es un enlace al conjunto de noticias redactadas por la misma persona (como The New York Times).

Línea de código	Tipo de esquema	N
1 <code><meta name="author" content="..."></code>	Estándar (HTML)	4
2 <code><meta property="article:authorName" content="..." /></code> <code><meta property="article:author" content="..." /></code>	Estándar (OG)	2
3 <code>...</code> <code><div class="..." itemprop="author" ...>...</div></code> <code><b itemprop="author" ...>...</code> <code><span class="byline" itemprop="author creator" itemscope</code> <code>itemtype="http://schema.org/Person" itemid="..."></code>	Estándar (SC)	4
4 <code><meta name="twitter:creator" content="..." /></code>	Estándar (TC)	6
5 <code><meta name="parsely-page" content="{..., "authors": ["..."]}" /></code>	Ad hoc (Parse.ly)	1
6 <code><meta name="DCSext.author" content="..." /></code>	Ad hoc (DCSext)	1
7 <code><meta name="byl" content="..." /></code>	Propietario (201)	1
8 <code><div class="..."><div id="..." aria-label="..." role="..." class="..." data-</code> <code>shares="..." ... data-author="..."></code>	Propietario (201)	1
9 <code><li class="..." data-authors="..." ...></code>	Propietario (202)	1

Tabla 8. Formas de identificar autor/a en los medios de la muestra. Fuente: Elaboración propia a partir de los códigos fuente de los diarios de la muestra

Para este dato no se identificó ningún metadato en Pravda (401), Nigerian Tribune (501), China Daily (601), The Times of India (602), Asahi Shimbun (603), Arab News (701) ni en Yedioth Aharonot (721).

HTML5 permite expresar la autoría mediante la etiqueta *rel="author"*. Este atributo, escaso en la muestra, no se ha tenido en cuenta en este trabajo, ya que HTML5 aún no se ha asentado.

NOMBRE DEL PERIÓDICO

El medio es el que canaliza y publica los artículos de sus periodistas, agencias e invitado/as. Esta información es susceptible de ser expresada en diversos campos, pues el periódico es, normalmente, el propietario del sitio web que recoge los artículos, dueño del copyright de la información que publica, tiene uno o varios perfiles en redes sociales, etc.

Para este dato, destacan algunos aspectos que influyen en la extracción de información a partir del código fuente:

1. **Diferentes atributos para un mismo tipo de información.** Para el nombre del periódico existen varios atributos que, aunque en principio están destinados a fines diferentes, suelen mostrar el nombre del medio: publicador (*publisher*), copyright (“copyright”), nombre del sitio web (*site name*), etc.
Por otro lado, como se ha indicado en el bloque anterior, Süddeutsche Zeitung (101) recoge en la etiqueta meta de autor el nombre del medio, contabilizándose en la Tabla 10, por lo que no se tiene en cuenta en este punto. Lo mismo sucede con `<meta name="twitter:creator" content="..." />`, usada por O Globo (301).

2. **Mismo atributo, diferentes valores de dato.** Un aspecto a destacar es el uso de códigos para identificar el medio, como hacen O Globo (301) o The Daily Telegraph (121), de quien es esta etiqueta meta: `<meta property="article:publisher" content="143666524748" />`; Otros, como The Economist (122), The Wall Street Journal (202), The Australian Financial Review (801) o Daily News Egypt (901), lo rellenan con la URL de su perfil en Facebook
3. **Otros datos como ruido de la etiqueta meta o atributo referentes al nombre del medio.** En ocasiones, este metadato es declarado en las etiquetas donde debería aparecer como valor de dato el nombre del autor de un artículo, aún cuando en la noticia aparece una persona con nombre y apellidos
4. **Nombre del medio como ruido de etiquetas de otros datos.** Por otro lado, también es frecuente encontrar el nombre del medio entre las etiquetas `<Title>` y `</Title>`, complementando al título, e incluso entre las palabras clave de la noticia. No se han tenido en cuenta en la siguiente Tabla.
5. **Etiquetas que no son estándar o han dejado de ser utilizadas.** Entre las etiquetas meta, las cuatro vistas anteriormente son consideradas estándar (Hickson et al., 2014; [Web Hypertext Application Technology Working Group (WHATWG)], 2016b). No ocurre así con `<meta name="copyright" content="...">`, que podría ser sustituida por `dcterms:rights`, entre otras posibilidades. Sólo es utilizada por Süddeutsche Zeitung (101) y no se recoge en la Tabla que sigue.
6. También se ha encontrado un caso de uso de la etiqueta `<acronym title="...">...</acronym>`, por The Financial Times (123), dónde el valor del atributo "title" es el nombre del medio. Esta etiqueta ya no es soportada en HTML5 y ha sido reemplazada por `<abbr>`, sin uso en la muestra.

Finalmente, The New York Times (201), como con el autor de la noticia, usa una misma línea para hacer referencia a tres atributos, poseedor del copyright de la noticia, proveedor de la noticia y organización fuente, con el mismo valor ([Web Hypertext Application Technology Working Group (WHATWG)], 2016a): `<span itemprop="copyrightHolder provider sourceOrganization" itemscope itemtype="http://schema.org/Organization" itemid="...">`.

La siguiente Tabla resume los esquemas identificados para el nombre del medio.

Línea de código	Tipo de esquema	N	
1 <code><acronym title="...">...</acronym></code>	Estándar (HTML)	1	22
2 <code><meta property="og:site_name" content="..."></code>	Estándar (OG)	16	
3 <code><meta property="article:publisher" content="..." /></code>		6	
4 <code>...</code> <code><meta itemprop="publisher" content="..."></code> <code><div class="logo" itemscope itemprop="publisher" itemtype="http://schema.org/Organization">...</code>	Estándar (SC)	4	6
5 <code>...</code> <code><div itemscope itemprop="copyrightHolder" itemtype="http://schema.org/Organization"></code> <code><meta itemprop="name" content="..."></code>		2	
6 <code><meta name="twitter:site" content="..." /></code> <code><meta name="twitter:site" value="..." /></code>	Estándar (TC)	10	
7 <code><meta name="DC:publisher" content="..."></code>	Estándar (DC)	1	
8 <code><meta name="cre" content="..." /></code>	Propietario (201)	1	
9 <code><meta name="page.site" content="wsj" /></code>	Propietario (202)	1	

Tabla 9. Formas de identificar autor/a en los medios de la muestra. Fuente: Elaboración propia a partir de los códigos fuente de los diarios de la muestra

Igual que ocurre con el autor, existe un atributo `rel="publisher"`, que indica que el dato que aparece en la etiqueta, habitualmente una página web o el perfil de alguna red social del medio,

se refiere a un publicador. Véase, por ejemplo, `<a href="https://plus.google.com/+arabnews" rel="publisher" ...>`, de Arab News (701). No se ha tenido en cuenta en la tabla anterior.

Para este metadato no se identificó ningún metadato en Nigerian Tribune (501), China Daily (601), Arab News (701) ni en Gulf News (711).

EL SOFTWARE METADATOSHTML

Dada la diversidad de notaciones identificadas para los distintos datos, se desarrolló un software para la extracción de información en las noticias que componen la muestra (documentos HTML), almacenadas inicialmente en una hoja de cálculo. Para la base de datos se utilizó el software Microsoft Access. De esta forma, es posible mantener disponible en el tiempo la información extraída para su consulta. Para la construcción del algoritmo se utilizó el lenguaje Object Pascal.

La primera etapa consistió en la construcción de una base de datos, compuesta por tres Tablas (ver Tabla 11), que facilitara la relación entre sus elementos. En esta etapa inicial también se importaron los datos de la hoja de cálculo a la Tabla URL.

Nombre de la tabla	Objetivo
URL	Almacenar las direcciones (URLs) de las noticias
Tags	Almacenar las etiquetas meta y atributos que se desean extraer
Metadados	Almacena el contenido de las etiquetas meta y atributos de la Tabla Tags

Tabla 10. Tablas que componen la base de datos de MetadatosHTML

Posteriormente y, a partir del análisis realizado anteriormente, se registraron etiquetas meta y atributos en la Tabla Tag. Este proceso conlleva un importante esfuerzo pues, como ya se ha indicado en el este apartado, diversos periódicos optaron por utilizar notaciones diferentes, sin seguir un patrón y, a veces, en formato propietario. Con esta infraestructura informacional, se inició el desarrollo propiamente dicho del software.

MetadadosHTML tiene tres funcionalidades:

1. Acceso al código fuente de las URLs contenidas en la hoja de cálculo inicial. Debido al gran número de URLs objeto de análisis, se descargaron todos los códigos HTML de las noticias en local, de forma que su análisis fuera realizado sin necesidad de acceder de nuevo a las páginas web. De esta forma se favorece la independencia de la conexión de datos y mayor velocidad de procesamiento posterior
2. Definición de las etiquetas meta y atributos de cada medio, teniendo en cuenta sus particularidades y formatos. Esta realiza el análisis del documento para la extracción de los valores de dato deseados. Debido a la ausencia de patrones de notación en parte del corpus seleccionado, hubo una dificultad adicional para la extracción de información. La estrategia adoptada para resolver este problema fue, junto a las etiquetas meta y atributos de la Tabla Tags, la inclusión de los marcadores de inicio y fin de cada elemento. El objeto es, simplemente, relacionarlos en el algoritmo por el nombre. Esto sería suficiente para documentos bien formados y válidos según la notación W3C
3. Extracción del contenido de los códigos fuente, exportación a una hoja de cálculo y almacenamiento para su consulta y análisis posterior.

La siguiente figura muestra una captura de pantalla de la aplicación MetadadosHTML.

analizados. También se ha observado que algunos periódicos cometen errores al referir los nombres de los atributos de esquemas estandarizados, como Open Graph Protocol (OG) o Schema.org, (SC) lo que, sin duda, impediría la extracción de valores de dato siguiendo los esquemas indicados.

Entre los esquemas estándar, las etiquetas HTML son las más utilizadas (100%), seguidas de Open Graph Protocol (90%), Twitter Cards (71%) y Schema.org (52%), distintos grados de aplicación, lo que se traduce en mayores o menores posibilidades de interoperabilidad entre sistemas. El resto de esquemas son utilizados por un sólo medio en cada caso (5%), a excepción de la etiqueta `<meta name="news_keywords">`, de Google, que aparece en un 30% de la muestra de periódicos. Se puede considerar que estas frecuencias están ligadas a los propósitos con que se desarrollan los esquemas, entre los que destacan:

- 1) Hacer posible la interoperabilidad entre sistemas mediante elementos puestos en común, como Schema.org (SC) o Dublin Core (DC)
- 2) Facilitar la búsqueda y recuperación de información en motores, como la etiqueta meta news_keywords, o a nivel interno del medio de comunicación (The New York Times, The Wall Street Journal)
- 3) Mejorar la visualización de la información a través de redes sociales (esencialmente Facebook y Twitter)
- 4) Realizar acciones de seguimiento y analítica web, como Sailthru o Webtrends. Sobre estas últimas, no se han encontrado etiquetas meta para el resumen, probablemente porque no es un elemento de búsqueda y de seguimiento.

Esto evidencia la necesidad de avanzar en el uso de *‘estándares para que los datos estén disponibles en distintos formatos, que hagan más versátil su reutilización’* (Pastor Sánchez, 2016), especialmente en aquellos que tienen como objetivo esencial la interoperabilidad entre sistemas. En ese sentido, resulta imprescindible la adopción de buenas prácticas, marcando las noticias en las formas en que los estándares lo indican, así como en el diseño y aplicación de esquemas ad hoc y propietarios. Una buena forma de hacerlo sería explicitando qué esquemas y metadatos utilizan los periódicos, a través de documentos disponibles en la Web.

Si nos centramos en el tipo de dato, los más asentados son título y resumen. Estos aparecen, mediante etiquetas meta o atributos, en todos los diarios de la muestra. Destaca la ausencia de metadatos, en cualquiera de las formas identificadas, en palabras clave (16%), nombre del medio (19%) y autoría (33%). Todos son elementos básicos del análisis de contenido (palabras clave) y formal (autor y publicación) de un documento y, como tal, aparecen definidos por Dublin Core (DC). En el ámbito específico de medios de comunicación impresa, existen trabajos, como (Yaginuma et al., 2003b), en el marco del proyecto Omnipaper, que también los considera imprescindibles para la correcta representación de una noticia.

En cuanto a MetadatosHTML, destaca la sujeción a los cambios en los códigos fuente de los medios. Si un periódico redefine alguna de sus etiquetas meta o comienza a utilizar un nuevo esquema, el software no puede aprovecharlas, lo que obliga a una continua revisión. Además, destaca la dificultad de recoger todos los atributos mostrados en los códigos fuente, pues no siempre están disponibles en la Web ni se dan a conocer los esquemas utilizados. De acuerdo con esto, se percibe la necesidad de definir una política de indización entre los periódicos. Para que los metadatos de las noticias sean reconocidos por diversos analizadores (*parsers*), esta política debe estar alineada con los principales estándares de metadatos. Es evidente que la simple adopción de metadatos no es suficiente para la divulgación y accesibilidad de los documentos. En muchos sistemas de recuperación de información dichos documentos no serán objeto de análisis ni de indización.

Los resultados iniciales de MetadatosHTML evidencian la necesidad de establecer un orden de preferencia en la extracción de valores de datos a partir de unos esquemas u otros. Para ello,

debe modificarse el software, teniendo en cuenta frecuencia de uso de los esquemas y el grado de normalización de las prácticas de cada medio. El propósito es asegurar la correcta extracción de valores de datos con el menor número de procesos. La información extraída será utilizada para la construcción de una ontología de dominio, facilitando la definición de clases, subclases y relaciones.

BIBLIOGRAFÍA

- Abbud Grácio, J. C., & Fadel, B. (2010). Estratégias de preservação digital. En *Gestão, mediação e uso da informação* (pp. 58-83). São Paulo, Brasil: Editora UNESP; Cultura Acadêmica
- AENOR. (2012). Información y documentación. Metadatos para la gestión de documentos. Parte 3: Método de auto-evaluación. UNE-ISO/TR 23081-3. AENOR.
- Axelsson, J., Birbeck, M., Dubinko, M., Epperson, B., Ishikawa, M., McCarron, S., ... Pemberton, S. (Eds.). (2006). XHTML™ 2.0 (W3C Working Draft 26 July 2006). W3C
- Baños-Moreno, María-José, Felipe, E. R., Pastor-Sánchez, J. A., Martínez-Béjar, Rodrigo, & Lima, G. (2015). Metadatos en noticias: un análisis internacional para la representación de contenidos en periódicos. Presentado en II Congreso ISKO España y Portugal / XII Congreso ISKO, Facultad de Comunicación y Documentación, Universidad de Murcia, Murcia, España
- Baños-Moreno, M.-J. (2013). Fuentes para la actualización de macro-tesauros: Noticias de divulgación científica. *Cuadernos de Gestión de Información*, 3(1), 13–24.
- 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
- BI3D. (2013). BI3D (Believe in 3D!) [Perfil público en LinkedIn]
- 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
- Google Inc. (2016). Ayuda para editores: Palabras clave y consultas de búsqueda
- Harris, J. (2007, octubre 23). Messing Around With Metadata
- Hickson, I., Berjon, R., Faulkner, S., Leithead, T., Navara, E. D., O'Connor, E., & Pfeiffer, S. (Eds.). (2014, octubre 28). HTML5. A vocabulary and associated APIs for HTML and XHTML (W3C Recommendation 28 October 2014). W3C
- Hillman, D. (2005, noviembre 7). Using Dublin Core (DCMI Recommended Resource). Recuperado a partir de <http://dublincore.org/documents/usageguide/> (Archivado por WebCite® en <http://www.webcitation.org/5bZjF2a6c>)
- Janes, D. (2013, agosto). hAtom 0.1 [Draft Specification]
- Kallipolitis, L., Karpis, V., & Karali, I. (2012). Semantic search in the World News domain using automatically extracted metadata files. *Knowledge-Based Systems*, 27, 38-50
- Kirkup, J. (2016, enero 2). How EU referendum polls are getting it wrong amid a «cloud of uncertainty». *The Daily Telegraph*
- Lajusticia, M. R. B. (2000). Estructura textual, macroestructura semántica y superestructura formal de la noticia. *Estudios sobre el mensaje periodístico*, (6), 239-258.
- Leithead, T., Pfeiffer, S., Navara, & O'Connor, E. (Eds.). (2012, agosto 22). HTML5. A vocabulary and associated APIs for HTML and XHTML (Editor's Draft 22 August 2012). W3C
- Martin, J. (2016, febrero 1). Ted Cruz Wins Republican Caucuses in Iowa. *The New York Times*
- Parker, A. (2016, febrero 14). Jeb Bush Adds a Weapon (His Brother) Despite Worries It Could Misfire. *The New York Times*. Charleston, S.C.
- Parse.ly. (s.f.). Technical documentation. The metadata tag
- Pastor-Sánchez, J. A. (2011). *Tecnologías de la web semántica*. Barcelona, España: Editorial UOC, S.L.

- Pastor Sánchez, J.-A. (2016, enero 29). [ThinkEPI] Quince años de web semántica: de las tecnologías a las buenas prácticas. *IWETEL*
- Pereira, T., & Baptista, A. A. (2004). Incorporating a semantically enriched navigation layer onto an RDF metadatabase. En J. Engelen, S. M. S. Costa, & A. C. S. Moreira (Eds.), *Building digital bridges: linking cultures, commerce and science*. Brasília
- Sailthru. (2013, octubre 1). Custom Horizon Tags
- [Web Hypertext Application Technology Working Group (WHATWG)]. (2016a, febrero 4). Microdata
- [Web Hypertext Application Technology Working Group (WHATWG)]. (2016b, febrero 4). Standard metadata names
- Yaginuma, T., Pereira, T., & Baptista, A. A. (2003a). Design of metadata elements for digital news articles in 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 ICC/IFIP International Conference on Electronic Publishing* (pp. 132-139). Minho, Portugal: Universidade do Minho
- 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