Representación y Organización de Periódicos Digitales con el Lenguaje XML †

D. M. Llidó, R. Berlanga, M.J. Aramburu y I.Sanz*

Departamento de Informática, Universitat Jaume I, Castellón
{ dllidó, aramburu, berlanga } @inf.uji.es

* STARLab Vrije Universiteit Brussel
isanz@vub.ac.be

Resumen

En este trabajo mostramos como la edición de periódicos digitales en XML (*eXtended Mark-up Language*) supone una mejora para el desarrollo de herramientas de almacenamiento, localización y búsqueda de información periodística en la Web. Actualmente, los periódicos digitales consisten en una mera versión navegable de la edición impresa, sin proporcionar ningún valor añadido. Hasta ahora el lenguaje HTML (*Hypertext Mark-up Language*) ha servido para tal propósito, pero se muestra claramente insuficiente ante los retos que suponen los nuevos métodos de recuperación de la información y las futuras bibliotecas digitales de prensa.

Palabras Clave:

Internet, Prensa Electrónica, XML, Bibliotecas Digitales.

Abstract

In this work we show how the publication of digital newspapers in XML (*eXtended Mark-up Language*) allows to improve the development of tools to store, locate and search information in the Web. Nowadays, digital newspapers constitute a navigable version of the printed publication, without providing any added value. So far, the HTML language (*Hypertext Mark-up Language*) has accomplished this purpose, but it is clearly insufficient for achieving the requirements of the new information retrieval methods, and the future digital libraries of newspapers.

Keywords:

Internet, Electronic Press, XML, Digital Libraries.

INTRODUCCIÓN

La irrupción de la World Wide Web como medio de comunicación, junto con la avanzada informatización del proceso de edición de periódicos y revistas, ha permitido la masiva aparición de periódicos digitales en Internet. Actualmente podemos encontrar en este medio numerosos ejemplares de periódicos locales, nacionales e internacionales que replican total o parcialmente el contenido de sus homólogos impresos. Sin embargo, hasta la fecha, todos los servidores de periódicos digitales se han limitado a crear una versión navegable de la edición impresa, sin proporcionar ningún valor añadido nuevo. Este tipo de servicio implica un esfuerzo muy reducido ya que aprovecha gran parte del proceso informatizado de la edición impresa (ver figura 1).

Una cuestión de actualidad que surge de los distintos diarios con versión digital es precisamente qué valores añadidos puede aportar la edición digital frente a la impresa [1]. El más evidente de éstos es la posibilidad de seleccionar cierta información entre los ejemplares publicados en varios diarios. Otros valores añadidos pueden ser: la publicación de noticias "on-line", el relacionar las noticias actuales con otras acontecidas en el pasado, el proporcionar a los lectores información adicional sobre el contexto de las noticias, la subscripción de lectores a foros de debate (listas de mail), etc. Sin embargo, todos estos nuevos servicios requieren una

† Este trabajo ha sido subvencionado por la CICYT (TEL97-1119) y la Fundació Bancaixa de Castelló.

reestructuración profunda en el proceso de la edición digital actual, el cual debería separarse del de la impresa para así tomar su propia identidad y no ser una mera versión navegable de la misma.

A la hora de llevar a cabo esta reestructuración, la mayor dificultad que se plantea en los actuales servidores de periódicos digitales, es el tipo de formato utilizado. Como veremos en la sección 2, la inmensa mayoría de periódicos digitales se publican en forma de ficheros HTML, los cuales se obtienen automáticamente a partir de los mismos ficheros que conforman la edición impresa. El lenguaje HTML está especialmente orientado a la apariencia de los documentos, y no a su estructura y significado. Sin embargo, los documentos digitales manejados internamente en el diario sí contienen etiquetas sobre la estructura y contenido. El problema es que éstas se pierden en el proceso de conversión a HTML. Algo similar sucede en el proceso de archivar los periódicos para su futura consulta, donde los documentos se almacenan en un *Sistema de Recuperación de la Información* como ficheros de texto planos. En conclusión, podríamos decir que actualmente no se está aprovechando el esfuerzo realizado en la informatización de las redacciones de los periódicos, ya que mucha meta-información de los ejemplares, necesaria para los valores añadidos de la edición digital, se pierden irremediablemente después de ser publicados.

En este trabajo proponemos una nueva aproximación a la edición de periódicos basada en el lenguaje XML (*eXtended Mark-up Language*). Ésta consiste en introducir este lenguaje en todas las etapas de producción, desde los redactores hasta el servidor de periódicos y el propio archivo del periódico. El objetivo final es dotar a los periódicos digitales de cierta facultades para que puedan ser más útiles después de su publicación.

El Proyecto Chronology

La principal motivación del presente trabajo surge del proyecto Chronology [2], que está siendo desarrollado por el Grupo de Bases de Conocimiento de la Universitat Jaume I de Castellón junto con la Universitat de València. En términos generales, este proyecto pretende desarrollar una base documental de prensa junto con las herramientas necesarias para analizar y explotar toda la información histórica de dicha base.

En este contexto, la estructura lógica de los periódicos resulta muy útil a la hora de recuperar la información, ya que el contenido de cada documento está relacionado con su ubicación en el periódico [3]. Por ejemplo, una palabra será más representativa para describir el contenido del documento si se encuentra en el resumen. Asimismo, para explotar eficazmente la información de los periódicos es necesario especificar cierto tipo de meta-información para cada documento (ej. la fecha y el lugar donde acontecen los hechos descritos en cada noticia) [4]. Entre otras cosas, esta meta-información permitirá reconocer el tipo de documentos y su contenido.

De lo anterior se deduce que los documentos que conforman un periódico no deben ser un simple bloque de texto que se visualiza con un determinado formato, y que después de ser publicados dejan de tener valor (tal y como ocurre en la actualidad con todos los periódicos digitales). Por el contrario, estos documentos deben describir una estructura y un contenido que puedan hacerles útiles en el futuro para el estudio de expertos e investigadores, o como herramienta de consulta de los propios lectores.

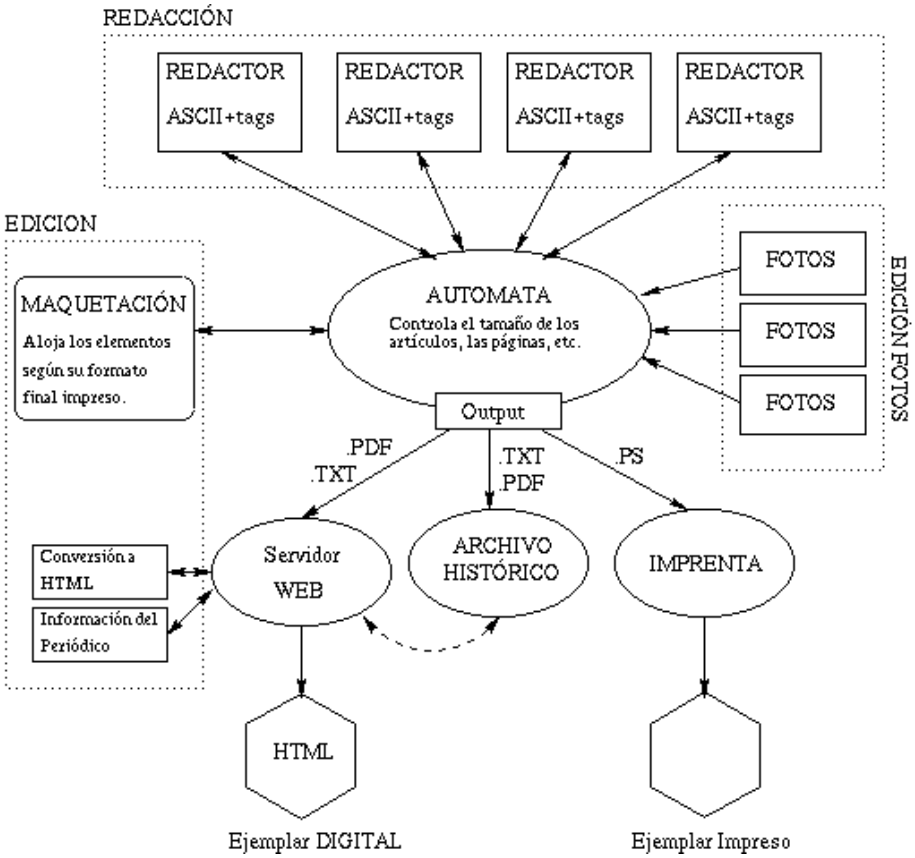

Figura 1: Proceso de producción de un periódico.

Actualmente, la base documental *Chronology* es alimentada automáticamente con ejemplares en formato HTML publicados diariamente en diversos servidores de prensa [5]. Para este proceso se han definido unos agentes Internet que son capaces de reconstruir la estructura implícita del periódico a partir de los estilos visuales de HTML [6]. Esta aproximación ha implicado dos problemas de difícil solución. Por un lado la extracción de la estructura lógica de los periódicos a partir de las etiquetas visuales resulta compleja y requiere la supervisión de alguna persona familiarizada con el libro de estilos empleado en cada diario. Por otro lado, errores tipográficos o pequeñas variaciones en las etiquetas de los documentos, pueden hacer fallar a los agentes en el reconocimiento de la estructura. Las causas de estos inconvenientes se encuentran en las propias limitaciones del lenguaje HTML para describir las unidades conceptuales de un documento.

En resumen, el futuro desarrollo de bibliotecas digitales de prensa pasa necesariamente por la definición de un lenguaje de descripción estándar para los periódicos y su meta-información (metadata). En este trabajo trataremos de aproximarnos a este lenguaje partiendo de otros trabajos previos sobre modelos de datos para documentos [4] y el lenguaje XML [7].

EDICIÓN DIGITAL DE UN PERIÓDICO CON XML

En esta sección describimos las pautas a seguir para definir en XML los tipos de documentos que representan a los periódicos digitales españoles. Además, analizaremos como organizar estos documentos XML para su difusión en Internet y su posterior almacenamiento en una base documental. El sistema de tipos propuesto permite incluir otras definiciones de tipo utilizadas por las agencias de noticias (ej. EFE o Reuters). De este modo, las noticias recibidas en la redacción de un periódico desde estas agencias podrían incorporarse directamente en el periódico conservando su meta-información.

Preliminares

La idea de publicar documentos estructurados para su intercambio y manipulación con un formato estándar y abierto, data de principios de los años 60. Una de estas líneas de investigación condujo a la definición de SGML (*Standard Generalized Markup Language*) [8],

que en 1986 se convirtió en un estándar ISO para el intercambio de documentos estructurados. Posteriormente, con la aparición del primer navegador Web (MOSAIC) en 1993 se desarrolló el lenguaje HTML (*Hypertext Markup Language*), que es una aplicación de SGML. Sin embargo como se ha dicho anteriormente, este lenguaje está especialmente orientado a describir el aspecto visual de un documento y sus enlaces con otros documentos.

Recientemente, se han multiplicado los esfuerzos para establecer algún tipo de orden sobre la información publicada en la Web, constatándose las notables deficiencias del lenguaje HTML para estos propósitos [9]. Así, ha surgido la necesidad de desarrollar un lenguaje más próximo al espíritu de SGML para describir la estructura y contenido de los documentos, además de su aspecto visual y enlaces. XML (*Extended Markup Language*) [7] es hoy por hoy la propuesta más seria y reconocida para superar las deficiencias de HTML.

Al igual que SGML, el lenguaje XML permite la definición de tipos de documentos ó DTDs (*Document Type Definition*). Todo documento XML debe asociarse a un DTD y ajustarse al mismo; concretamente, debe cumplir la gramática implicada por su DTD. Finalmente, para dar un formato visual a un documento XML, su DTD debe asociarse a una hoja de estilos [10].

La adopción de XML está especialmente justificada cuando se dispone de un conjunto de tipos de documento bien establecido y éstos son reutilizados con mucha frecuencia. Éste es precisamente el caso de la prensa electrónica. De hecho, las primeras aplicaciones de XML se han realizado en el campo de la industria de noticias. Así, el formato NITF (*News Industry Text Format* [11]), elaborado por la *International Press Telecomunications Council* y la *Newspaper Association of America*, está siendo utilizado en la actualidad por reconocidas agencias internacionales de noticias, como por ejemplo Reuters.

Sin embargo, el formato NITF está orientado a la difusión de noticias y no a la edición digital de periódicos. De hecho, los tipos definidos en este formato son demasiado simples para representar un periódico completo. Por ejemplo, no es posible la definición de elementos típicos de un periódico tales como las portadas, las secciones temáticas, las subnoticias, etc. Tampoco se distingue entre los diferentes elementos periodísticos, tales como los reportajes, columnas, crónicas, breves, etc. En cambio, el formato NITF proporciona un juego muy completo de etiquetas para describir el contenido semántico de los documentos (por ejemplo, existen etiquetas para indicar las ciudades, los eventos, las marcas temporales, los nombres propios, etc.). Este tipo de etiquetas es muy útil para tareas de filtrado y recuperación de la información en los archivos del periódico, y por tanto es interesante conservarlas en la definición de tipos que se proponga para un periódico digital.

Infoestructura de un periódico digital

La gran mayoría de diarios nacionales organizan la información de cada número de una forma similar y bastante regular. Así, los artículos periodísticos se agrupan en secciones temáticas fijas (por ejemplo: 'Internacional', 'Nacional', 'Deportes', etc.). Opcionalmente, dentro de estas secciones fijas se suelen crear secciones temáticas temporales referentes a algún evento de relevancia (por ejemplo: 'Olimpiadas', 'Elecciones Generales', etc.). Finalmente todas las secciones se agrupan en lo que conforma el número o ejemplar.

El gran volumen de información contenido en un solo ejemplar hace poco recomendable representar éste con un único fichero HTML o XML. Además, si se tiene en consideración el proceso de edición del periódico (figura 1), el ejemplar final es más bien un agregado de ficheros con los distintos artículos periodísticos. Así pues, la edición digital de un periódico requiere una infoestructura[1] donde los artículos periodísticos de un ejemplar se organicen de forma jerárquica según la estructura lógica del periódico (ver figura 2). De hecho, actualmente todos los periódicos digitales siguen estas pautas.

En la infoestructura de un periódico se deben distinguir los siguientes tipos de documentos:

- *Artículos periodísticos*, los cuales contienen toda la información de periódico en forma de datos multimedia (texto, imágenes, etc.). Generalmente existen varios tipos de estos documentos, entre otros: 'Noticia', 'Columna', 'Fotonoticia', 'Crónica', etc. Aunque cada diario emplea sus propios tipos según su libro de estilo (ej. [12]), suele haber una correspondencia clara entre ellos.
- *Índices*, que son documentos que contienen todos los enlaces necesarios para agrupar los artículos periodísticos por secciones temáticas (temporales o fijas) y en última instancia por ejemplares. Además de servir de "pegamento", estos ficheros deberían asegurar que todos los documentos contenidos existen y que además son del tipo apropiado.

[1] Conjunto de documentos de hipertexto enlazados entre sí que suelen ubicarse en un mismo servidor.

- *Portadas*, que son documentos que contienen una selección de enlaces a artículos de diferentes secciones junto con un resumen de los mismos. En realidad, este tipo de documentos no forma parte del periódico, si no que son herramientas de ayuda a la navegación para los lectores. Las portadas pueden obtenerse de forma automática a partir de los índices y los propios artículos periodísticos. Es importante mencionar que estos documentos son los que podrían dar realmente un valor añadido a los lectores, ya que podrían filtrar, relacionar o enriquecer los contenidos del periódico.

Lo que resta de sección se dedica a mostrar algunos ejemplos de tipos de documento para expresar la organización y contenido de un periódico digital. En las conclusiones discutiremos el impacto de utilizar XML en la edición de prensa electrónica.

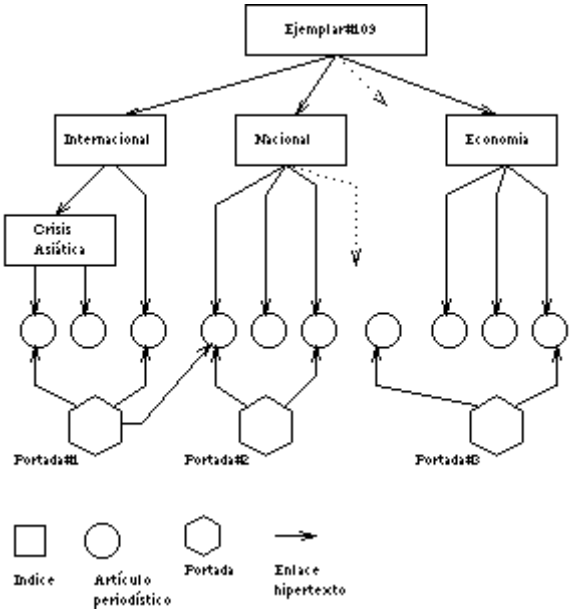

Figura 2: Ejemplo de Infoestructura de un periódico digital

Document Type Definition (DTD)	Ejemplo
<!ELEMENT **Noticia** (seccion?, cintillo?, Imagen?, antetitulo*, titular+, Imagen?, Data, subtitulo*, Contenido+, Despiece*)> <!ELEMENT **Despiece** (antetitulo*, titular+, Data?, Contenido+)> <!ELEMENT **Data** (autor,lugar?)> <!ELEMENT **Imagen** (foto,pie?)> <!ELEMENT **Contenido** ((parrafo|Imagen)+)> <!ELEMENT **seccion** (#PCDATA)> <!ELEMENT **cintillo** (#PCDATA)> <!ELEMENT **titular** (#PCDATA)> <!ELEMENT **subtitulo** (#PCDATA)> <!ELEMENT **parrafo** (#PCDATA)> <!ELEMENT **antetitulo** (#PCDATA)> <!ELEMENT **autor** (#PCDATA)> <!ELEMENT **lugar** (#PCDATA)> <!ELEMENT **foto** (#EMPTY)>	**<Noticia>** **<seccion>**INTERNACIONAL**</seccion>** **<cintillo>** Sección especial ... **</cintillo>** **<titular>** Alta Tensión</titular> **<Data>** **<autor>**W. Pérez**</autor>** **<lugar>**Bruselas **</lugar>** **</Data>** **<Contenido>** **<Imagen>** **<foto>**<img src="i.jpg">**</foto>** **<pie>** Imagen de ...**</pie>** **</Imagen>** **<parrafo>** La puesta en marcha ... **</parrafo>** ... **</Contenido>** **<Despiece>** ... **</Despiece>** **</Noticia>**

Tabla 1: DTD para el tipo de documento "Noticia".

ARTÍCULOS PERIODÍSTICOS

Como ejemplo de artículo periodístico mostraremos el DTD de una "Noticia", que es el tipo de documento más utilizado en los diarios (ver Tabla 1).

Los artículos periodísticos son los elementos más ricos que ofrece un periódico desde el punto de vista estructural. Por un lado contienen ciertos metadatos, como por ejemplo, el autor,

el lugar y la agencia que proporciona la noticia (además de los atributos semánticos que puedan ser etiquetados según el formato NITF). Por otro lado, este tipo de documento puede contener otros documentos anidados, como es el caso del tipo "Despiece" dentro del tipo "Noticia". Tanto los metadatos como la estructura de las noticias son muy importantes para localizar la información contenida en las mismas.

ÍNDICES

Los documentos de tipo índice tienen el papel de agregar los artículos periodísticos según su contenido. Adicionalmente, los índices permiten restringir el tipo de documentos que forman parte de estas agregaciones. Así por ejemplo, es posible definir que el índice de la sección Internacional puede contener "Artículos" en general (Nocitias, Crónicas o Columnas), mientras que el de la sección de Opinión sólo puede contener "Cartas".

Con objeto de representar la estructura general de un periódico y sus secciones, el DTD asociado deberá incluir una serie de declaraciones similares a las que se dan a continuación para el ejemplo anterior:

```
<!ENTITY %Ejemplar "fecha,Secciones+">
<!ENTITY %Secciones "(Seccion|Seccion-Opinion)">
<!ENTITY %Seccion "nombre-sec, fecha, T-Articulo+">
<!ENTITY %Seccion-Opinion "nombre-sec, fecha, T-Carta+">
```

De esta manera, las secciones se componen de un nombre, una fecha y un conjunto de referencias a los ficheros que contienen sus artículos o sus cartas en cada caso. Para diferenciar las referencias a artículos de las referencias a cartas, cada enlace tienen unos atributos que, entre otras cosas, indican el tipo de los documentos a los que pueden apuntar. Las siguientes declaraciones definen estos dos tipos de enlaces.

```
<!ENTITY %link-car "
        href CDATA #REQUIRED
        ident ID #IMPLIED
        titulo CDATA #IMPLIED
        tipo CDATA #FIXED 'carta'">
<!ELEMENT T-Carta #PCDATA>
<!ATTLIST T-Carta %link-car>


<!ENTITY %Articulo "(T-Noticia| T-Cronica| T-Columna)">
<!ELEMENT T-Articulo %Articulo >

<!ENTITY %link-not "
        href CDATA #REQUIRED
        ident ID #IMPLIED
        titulo CDATA #IMPLIED
        tipo CDATA #FIXED 'noticia'">
<!ELEMENT T_Noticia #PCDATA>
<!ATTLIST T-Noticia %link-not>


<!ENTITY %link-cro "
        href CDATA #REQUIRED
        ident ID #IMPLIED
        titulo CDATA #IMPLIED
        tipo CDATA #FIXED 'cronica'">
<!ELEMENT T-Cronica #PCDATA>
<!ATTLIST T-Cronica %link-cro>

<!ENTITY %link-col "
        href CDATA #REQUIRED
        ident ID #IMPLIED
        titulo CDATA #IMPLIED
        tipo CDATA #FIXED 'columna'">

<!ELEMENT T-Columna #PCDATA>
<!ATTLIST T-Columna %link-col>
```

Durante el proceso de publicación de un periódico se deberá comprobar que todas sus secciones están referenciadas desde el lugar correspondiente y enlazadas de la forma adecuada. Para ello, el procesador de documentos XML deberá analizar los enlaces para validar que cada referencia `href` apunta a un documento del tipo indicado por el atributo `tipo` asociado.

En cuanto a la realización de portadas, cabe decir que su codificación en XML se hará de una forma similar a la de los índices, pero definiendo una estructura paralela que permite acceder a los documentos desde una dimensión diferente. De esta manera se pueden definir organizaciones alternativas para un periódico de acuerdo a los distintos perfiles de usuarios.

CONCLUSIONES

La principal conclusión de este trabajo es una defensa del uso generalizado del lenguaje XML para la edición de prensa electrónica. Como se mencionó en la introducción, la adopción de este lenguaje permitiría el desarrollo de herramientas más potentes para la búsqueda, localización y filtrado de la información publicada, además de aportar nuevos valores añadidos a las ediciones digitales.

Las ventajas de XML no se reducen a mejorar los servicios de cara a los lectores, si no que además se mejora el proceso de producción del diario. Por un lado, los DTDs permiten comprobar la coherencia y consistencia de los ejemplares que se están editando con respecto a los estilos propios del periódico. Además, se podría reducir el número de formatos de documentos utilizados actualmente en la edición de un periódico (ej. ASCII, PDF, HTML) a solamente dos: XML para los documentos digitales, y el formato PostScript para la edición impresa. Por último, al poder prescindir de los programas de conversión entre estos formatos, los procesos de edición digital y archivo se simplifican notablemente.

En cuanto a los valores añadidos de la edición digital, XML posibilita, entre otras, las siguientes tareas: filtrar la información requerida por el usuario a nivel del navegador, mejorar las consultas en los índices de búsqueda, distribuir las noticias a determinados lectores suscritos al diario según sus perfiles y desarrollar herramientas más sofisticadas como las que queremos llevar a cabo en el proyecto Chronology.

REFERENCIAS

[1] Fuentes, E. y Gonzalez A. La Prensa Española en Internet: Análisis de los Servicios de Valor Añadido. En: FESABID'98, *VI Jornadas Españolas de Documentación*, Valencia, 1998.

[2] *Proyecto Chronology* (CICYT TEL97-1119): http://nuvol.uji.es/~berlanga

[3] Aramburu M. J. y Berlanga, R. Modelling Periodicals with an Extended Object Oriented Model. En: *7th Workshop on Database and Expert Systems Applications*, Zurich, 1996, 9-14.

[4] Aramburu M.J. y Berlanga, R. An Approach to a Digital Library of Newspapers. *Information Processing & Management, Special Issue on Electronic News. Pergamon Press Publishers*, 33(5), 1997.

[5] Aramburu, M.J., Sanz, I., García, S. y Berlanga, R. Procesamiento de Periódicos Electrónicos para su Almacenamiento con Oracle 8. En: *IV Congreso sobre Tecnología de Objetos*, Universidad de Deusto, Bilbao, 1998, 19-28.

[6] Sanz, I., Berlanga, R. y Aramburu, M.J. Gathering Metadata from Web-based Repositories of Historical Publications. En: *9th Workshop on Database and Expert Systems Applications*, Viena, 1998, 473-478.

[7] Holzner, S. *XML Complete*. McGraw-Hill.1998.

[8] ISO 8879. *Information Processing- Text and Office Systems, Standard Generalized Markup Language*, 1996.

[9] Megginson, D.*Structuring XML Documents*. Prentice-Hall. 1998.

[10] *What are style sheets?*: http://www.w3.org/Style/

[11] *News Industry Text Format*: http://www.iptc.org/iptc/nitf.dec

[12] *Libro de Estilo El País* 14ª edición. Madrid, El País, 1998.