A mis padres, por enseñarme a tener paciencia, perseverancia y tesón.

A mi hermano, por hacerme entrar al trapo quiera o no quiera y porque siempre aprendo algo nuevo de él.

A mi primo Roberto porque se merece estar en esta dedicatoria.

Y a mis compañeros de trabajo porque ya son para mí como mi segunda familia después de todo lo que hemos vivido.

Javier S. Zurdo

A Beatriz, mis padres y mis hermanos, por todo el apoyo que me dan siempre.

Pablo Toharia

Introducción

Este libro surge con el propósito de acercar al lector a los aspectos más importantes que encierran los lenguajes de marcas ante la creciente demanda de personal cualificado para su utilización. Con tal propósito, puede servir de apoyo también para estudiantes de los Ciclos Formativos de Grado Superior de Administración de Sistemas Informáticos en Red y Desarrollo de Aplicaciones Multiplataforma, así como para profesionales de distinto rango.

Para todo aquél que use este libro en el entorno de la enseñanza (Ciclos Formativos, Profesionales o Universidad), se ofrecen varias posibilidades: utilizar los conocimientos aquí expuestos para inculcar aspectos genéricos de los lenguajes de marcas o simplemente centrarse en trabajar a fondo alguno de ellos. La extensión de los contenidos aquí incluidos hace imposible su desarrollo completo en la mayoría de los casos.

Ra-Ma pone a disposición de los profesores una guía didáctica para el desarrollo del tema que incluye las soluciones a los ejercicios expuestos en el texto. Puede solicitarlo a *editorial@ra-ma.com*, acreditándose como docente y siempre que el libro sea utilizado como texto base para impartir las clases.

1 Reconocimiento de las características de lenguajes de marcas

OBJETIVOS DEL CAPÍTULO

- ✓ Conocer qué es un lenguaje de marcas.
- ✓ Conocer los orígenes y evolución de los lenguajes de marcas.
- ✓ Conocer las organizaciones desarrolladoras de los lenguajes de marcas.
- ✓ Distinguir la clasificación de los lenguajes de marcas.
- ✓ Conocer las gramáticas de los lenguajes de marcas.

1.1 DEFINICIÓN Y CLASIFICACIÓN DE LENGUAJES DE MARCAS

Los **lenguajes de marcas** (también llamados **lenguajes de marcado**) son aquellos que combinan la información, generalmente textual, que contiene un documento con marcas o anotaciones relativas a la estructura del texto o a la forma de representarlo. El lenguaje de marcas es el que especifica cuáles serán las etiquetas posibles, dónde deben colocarse y el significado que tendrá cada una de ellas. Así mismo, la presencia de etiquetas o marcas intercaladas en el contenido hace explícita la estructura del documento o cualquier información adicional que se quiera resaltar. Por otro lado, hay que tener en cuenta que las propias etiquetas o marcas generalmente no se suelen presentar al usuario final, ya que este suele estar interesado en el propio contenido del documento.

A continuación, se muestra un ejemplo en el que mediante una serie de marcas o etiquetas se ha representado una información relativa a una noticia:

EJEMPLO 1.1

```
<noticia>
   <lugar>Madrid</lugar>
   <fecha>27/08/2010</fecha>
   <desc>Se ha inaugurado una estación de tren</desc>
</noticia>
```

¿SABÍAS QUE...?

Los lenguajes de marcas han de diferenciarse de los lenguajes de programación. El lenguaje de marcas no tiene funciones aritméticas o variables, como sí poseen los lenguajes de programación.

1.2 TIPOS DE LENGUAJES DE MARCAS

Los lenguajes de marcas se suelen dividir en tres grupos si bien hay que tener en cuenta que existen lenguajes que combinan características de más de un grupo:

- **Lenguajes orientados a presentación**. Este tipo de leguajes son los usados tradicionalmente por los procesadores de texto como puede ser Microsoft Word® y codifican cómo ha de presentarse el documento, por ejemplo, indicando que una determinada palabra debe presentarse en fuente itálica o que se debe dejar un espacio de 10 puntos al terminar el párrafo. Generalmente las marcas de los lenguajes orientados a

presentación se ocultan al usuario lo que permite obtener un efecto WYSIWYG[1]. Este tipo de lenguajes de marcas no suelen ser flexibles ni reusables.

¿SABÍAS QUE...?

En Word puedes ver las marcas pulsando el icono ¶ de la interfaz de Microsoft Word.

- **Lenguajes procedurales**. En este tipo de lenguajes las etiquetas son también orientadas a presentación pero se integran dentro de un marco procedural que permite definir macros (secuencias de acciones) y subrutinas. Entre los ejemplos más comunes de lenguajes procedurales podemos encontrar TeX, LaTeX y Postcript.

¿SABÍAS QUE...?

La mayoría de los documentos científicos, artículos de investigación o libros técnicos que contienen fórmulas matemáticas se escriben con Latex.

¿SABÍAS QUE...?

PostScript es un lenguaje de descripción de páginas (en inglés PDL, *Page Description Language*), utilizado en muchas impresoras y, de manera usual, como formato de transporte de archivos gráficos en talleres de impresión profesional.

- **Lenguajes descriptivos**. Este tipo de lenguajes no definen qué se debe hacer con un trozo o sección del documento sino que por el contrario las marcas sirven para indicar qué es esa información, esto es, describen que es lo que se está representando. La mayoría de los lenguajes de marcas que se usan hoy en día se encuentran dentro de este grupo como por ejemplo, el SGML y sus derivados (HTML, XML, etc.) que se verán a continuación.

¿SABÍAS QUE...?

El formato COLLADA está basado en XML y se utiliza para definir escenas de modelos tridimensionales, como el de los videojuegos.

1 *What You See Is What You Get* (Lo que ves es lo que obtienes)

1.3 EVOLUCIÓN DE LOS LENGUAJES DE MARCAS

Los lenguajes de marcas comenzaron a usarse a finales de la década de los 60 para poder introducir anotaciones dentro de documentos electrónicos, de la misma forma que se hacía cuando la documentación estaba en papel. De esta posibilidad de incorporar marcas es de donde reciben su nombre. Es en esas fechas cuando se estandariza el lenguaje **SGML** (*Standard Generalized Markup Language*), que es un descendiente directo del lenguaje **GML** propuesto por IBM. Este lenguaje surgió para permitir compartir información por parte de sistemas informáticos. Este estándar tuvo una gran aceptación pero no consiguió asentarse del todo debido principalmente a su complejidad lo que provocaba que el software que usará SGML terminaba siendo excesivamente extenso y complejo.

A finales de los 80 dentro del CERN (*Conseil Européen pour la Recherche Nucléaire*) se creó un lenguaje de marcado pensado para compartir información usando las redes de computadores y, de forma más general, a través de Internet. Este lenguaje se basaba en algunos principios de de SGML y lo denominaron HTML (*Hyper-text Markup Language*). La aparición de este lenguaje supuso de alguna manera una revolución en la forma de compartir información, gracias principalmente a la sencillez de sus sintaxis y del software necesario para interpretarlo. En poco tiempo el lenguaje HTML se extendió y empezó a crecer de forma en ocasiones descontrolada y casi siempre influenciado por razones meramente comerciales.

A mediados de los años 90 el consorcio **W3C** (*World Wide Web Consortium)* comenzó una iniciativa para intentar dotar a la *web* de un lenguaje más potente y que pudiera dar una estructurar semántica a la misma. Para ello se marcaron el objetivo de crear un nuevo lenguaje de marcas basado en SGML y que fuera sencillo como HTML. Finalmente, en el 1998, W3C hizo público un nuevo estándar que denominaron XML (*eXtended Markup Language*), más sencillo que SGML y más potente que HTML.

¿SABÍAS QUE...?

HTML es el lenguaje de marcas predominante para la elaboración de páginas web. Es usado para describir la estructura y el contenido en forma de texto, así como para complementar el texto con objetos tales como imágenes.

ACTIVIDADES 1.1

- Abra la página *www.google.es* en su explorador. Pulse el botón derecho del ratón y marque la opción "Ver código fuente".
- Busque información sobre el lenguaje de marcas XHTML.
- ¿En qué se diferencia un lenguaje de marcas a un lenguaje de programación?
- Visite la web *http://es.wikipedia.org/wiki/Categoría:Lenguajes_de_descripción.* Eche una ojeada a los diferentes lenguajes de marcas que existen.

1.4 ETIQUETAS, ELEMENTOS Y ATRIBUTOS

Existen tres términos comúnmente usados para describir las partes de un documento de lenguajes de marcas: etiquetas, elementos y atributos.

Una **etiqueta** (*tag*) es un texto que va entre el símbolo menor que (<) y el símbolo mayor que (>). Existen etiquetas de inicio (como <nombre>) y etiquetas de fin (como </nombre>).

Los **elementos** representan estructuras mediante las que se organizará el contenido del documento o acciones que se desencadenan cuando el programa navegador interpreta el documento. Constan de la etiqueta de inicio, la etiqueta de fin y de todo aquello que se encuentra entre ambas.

Algunos elementos no tienen contenido. Se les denomina elementos vacíos y no deben llevar etiqueta de fin.

Un **atributo** es un par nombre-valor que se encuentra dentro de la etiqueta de inicio de un elemento e indican las propiedades que pueden llevar asociadas los elementos.

EJEMPLO 1.2

Fíjese en el texto siguiente:

```
<direccion>
    <nombre>
        <titulo>Mrs.</titulo>
        <nombre> Mary </nombre>
        <apellidos> McGoon </apellidos>
    </nombre>
    <calle> 1401 Main Street </calle>
    <ciudad estado="NC"> Anytown</ciudad>
    <codigo-postal> 34829 </codigo-postal>
</direccion>
```

En el ejemplo anterior, el elemento <*nombre*> contiene tres elementos hijos: <*titulo*>, <*nombre*> y <*apellidos*> y *estado* es un atributo del elemento <*ciudad*>.

En el capítulo 2 se profundizará más sobre estos tres conceptos.

ACTIVIDADES 1.2

- Fíjese en el siguiente texto.

```
<noticia>
    <lugar>Madrid</lugar>
    <fecha>27/08/2010</fecha>
    <desc>Se ha inaugurado una estación de tren</desc>
</noticia>
```

Indique el nombre de una etiqueta, de dos elementos y de algún atributo.

- Escriba un texto que contenga etiquetas, elementos y atributos.

1.5 ORGANIZACIONES DESARROLLADORAS

Dentro de las organizaciones que se han encargado de desarrollar los lenguajes de marcas se encuentran:

Organización Internacional para la Estandarización (**ISO**, *International Organization for Standardization*). Se formó después de la Segunda Guerra Mundial (23 de febrero de 1947) y es el organismo encargado de promover el desarrollo de normas internacionales de fabricación, comercio y comunicación para todas las ramas industriales a excepción de la eléctrica y la electrónica. Su función principal es la de buscar la estandarización de normas de productos y seguridad para las empresas u organizaciones a nivel internacional.

Es una red de los institutos de normas nacionales de 163 países, sobre la base de un miembro por país, con una Secretaría Central en Ginebra (Suiza) que coordina el sistema.

Las normas desarrolladas por ISO son voluntarias, ya que es un organismo no gubernamental y no depende de ningún otro organismo internacional, por tanto, no tiene autoridad para imponer sus normas a ningún país. El contenido de los estándares está protegido por derechos de copyright y para acceder a ellos el público en general ha de comprar cada documento.

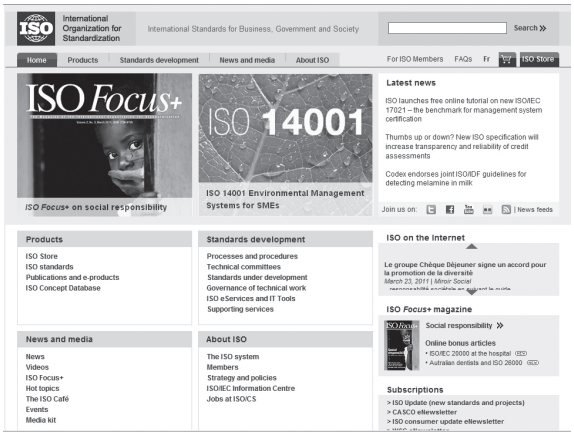

Figura 1.1. *Página web de la ISO*

Esta organización después del éxito que tuvo **GML** y, después de un largo proceso, publicó en 1986 el *Standard Generalized Markup Languaje* (**SGML**) con rango de Estándar Internacional con el código **ISO 8879**.

- **World Wide Web Consortium** (**W3C**). El W3C se creó en 1994 por Tim Berners-Lee en el MIT, actual sede central del consorcio. Posteriormente se unió, en abril de 1995, el INRIA en Francia, reemplazado por el ERCIM en el 2003 como el huésped europeo del consorcio y la Universidad de Kei (Shonan Fujisawa Campus) en Japón en septiembre de 1996 como huésped asiático. Su función principal es tutelar el crecimiento y organización de la web.

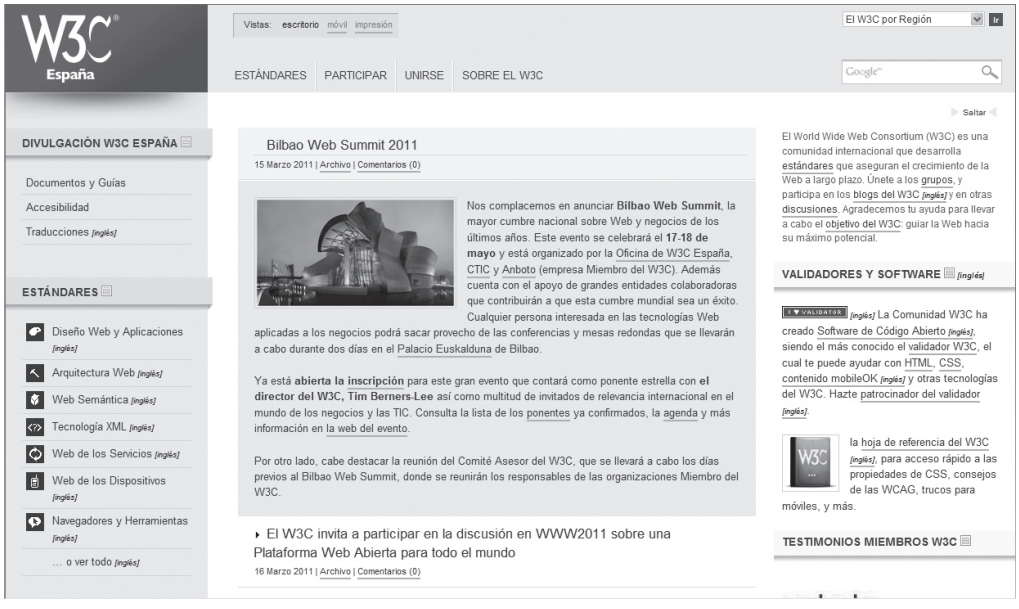

Figura 1.2. *Página web de la W3C*

Su primer trabajo fue normalizar el lenguaje HTML, el lenguaje de marcas con el que se escriben las páginas web. Al crecer el uso de la web, crecieron las presiones para ampliar el HTML. El W3C decidió que la solución no era ampliar el HTML, sino crear unas reglas para que cualquiera pudiera crear lenguajes de marcas adecuados a sus necesidades, pero manteniendo unas estructuras y sintaxis comunes que permitieran compatibilizarlos y tratarlos con las mismas herramientas. Ese conjunto de reglas es el XML, cuya primera versión se publicó en 1998.

ACTIVIDADES 1.3

- Busca información en Internet sobre las organizaciones ISO y W3C.

1.6 UTILIZACIÓN DE LENGUAJES DE MARCAS EN ENTORNOS WEB

Una **página web** es un documento electrónico adaptado para la *World Wide Web* que, normalmente, forma parte de un sitio web.

Está compuesta, principalmente, por información (solo texto o módulos multimedia) así como por hiperenlaces; además, puede contener o asociar datos de estilo para especificar cómo debe visualizarse, y también aplicaciones embebidas para hacerla interactiva.

Las páginas web están escritas en un lenguaje de marcas que proporciona la capacidad de manejar e insertar hiperenlaces, generalmente, HTML.

El contenido de la página puede ser predeterminado (**página web estática**) o generado en el momento de su visualización o al solicitarla a un servidor web (**página web dinámica**).

Respecto a la estructura de las páginas web, algunos organismos, en especial el W3C, suelen establecer directivas con la intención de normalizar el diseño, para así facilitar y simplificar la visualización e interpretación del contenido.

En el Capítulo 2 se ampliará la información sobre los lenguajes de marcas que se utilizan en entorno web.

1.7 GRAMÁTICAS

Todo documento de un lenguaje de marcas tiene en común una gramática que define el marcado permitido en esa clase, el marcado requerido y cómo debe ser utilizado dicho marcado en la instancia del documento.

1.7.1 DTD

El estándar define esta gramática mediante la **DTD** (**Definición de Tipo de Documento**) que establece las reglas de formación del lenguaje formal, es decir, qué combinaciones de símbolos elementales son sintácticamente correctas.

En la DTD se identifica la estructura del documento, es decir, aquellos elementos que son necesarios en la elaboración de un documento o un grupo de documentos estructurados de manera similar. Contiene las reglas de dichos elementos: el nombre, su significado, dónde pueden ser utilizados y qué pueden contener.

La especificación del W3C para HTML 4.0 contempla tres DTD:

- **DTD estricta** (*HTML 4.0 Strict DTD*): incluye todos los elementos y atributos que no han sido declarados "desaprobados" (*deprecated*), interpretando la expresión en el sentido de que no se recomienda ya su uso proponiéndose nuevos y mejores recursos para hacer lo mismo.
- **DTD transicional o flexible** *-loose-* (*HTML 4.0 Transitional DTD*): incluye todo lo que la anterior más los elementos y atributos desaprobados (*deprecated*).
- **DTD para documentos con marcos** (*HTML 4.0 Frameset DTD*): engloba todo lo incluido en la transicional más lo relativo a la creación de documentos con marcos (*frames*).

Recuerde que aunque la especificación recomienda ceñirse a los recursos de la DTD estricta, utilizar el resto de los elementos y atributos no es incorrecto.

La DTD es el formato de esquema nativo (y el más antiguo) para validar documentos XML, heredado de SGML. Utiliza una sintaxis no-XML para definir la estructura o modelo de contenido de un documento XML válido:

- ✓ Define todos los elementos.
- ✓ Define las relaciones entre los distintos elementos.
- ✓ Proporciona información adicional que puede ser incluida en el documento (atributos, entidades, notaciones).
- ✓ Aporta comentarios e instrucciones para su procesamiento y representación de los formatos de datos.

Es el método más sencillo usado para validar, y por esta razón presenta varias limitaciones, ya que no soporta nuevas ampliaciones de XML y no es capaz de describir ciertos aspectos formales de un documento a nivel expresivo.

Las DTD pueden ser internas o externas a un documento, o ambas cosas a la vez.

En el Capítulo 4 se ampliará la información sobre los DTD.

1.7.2 ESQUEMA XML

XML Schema es la evolución de la DTD descrita por el W3C, también denominado XSD (*XML Schema Definition*). Es un lenguaje de esquema más complejo y más potente, basado en la gramática para proporcionar una potencia expresiva mayor que la DTD. Utiliza sintaxis XML, cosa que le permite especificar de forma más detallada un extenso sistema de tipos de datos. A diferencia de las DTD, soporta la extensión del documento sin problemas.

A la hora de la validación del documento, la utilización de XSD supone un gran consumo en recursos y tiempo debido a su gran especificación y complejidad en la sintaxis (los esquemas son más difíciles de leer y de escribir).

Después de validar el documento con XML Schema, es posible expresar su estructura y contenido en términos del modelo de datos usado por el esquema de validación. Esta funcionalidad, conocida como *Post-Schema-Validation Infoset* (PSVI), se puede utilizar para transformar el documento en una jerarquía de objetos, a los cuales se puede acceder a través de un lenguaje de programación orientada a objetos (OOP).

El modelo de datos de XML Schema incluye:

- ✓ El vocabulario (nombres de elemento y atributo).
- ✓ El contenido modelo (relaciones y estructura).
- ✓ Los tipos de datos.

En el Capítulo 4 se ampliará la información sobre los esquemas XML.

1.7.3 RELAX NG

RELAX NG es un lenguaje de esquema basado en la gramática, muy intuitivo y más fácil de entender que el XML Schema. Tiene un alto poder expresivo, ya que, por ejemplo, permite validar elementos intercalados que pueden aparecer en cualquier orden.

Las aplicaciones de definición de documentos y validación para *RELAX NG* son más sencillas que las de *XML Schema*, haciéndolo más fácil de utilizar e implementar.

RELAX NG se ha convertido recientemente en un estándar ISO como la parte 2 de **DSDL** (*Document Schema Definition Language*).

RESUMEN DEL CAPÍTULO

En este capítulo se ha llevado a cabo una breve descripción de lo que es un lenguaje de marcas.

Se han indicado los orígenes y la evolución de los distintos lenguajes de marcas.

Se ha descrito de forma genérica lo que es una etiqueta y el concepto de elemento y de atributo.

Se ha hablado sobre las dos organizaciones desarrolladoras de los lenguajes de marcas (ISO y W3C).

Se ha descrito una sencilla clasificación de los lenguajes de marcas.

Se ha indicado lo que es la gramática de los lenguajes de marcas, indicando lo que son los DTD, XML Esquema y Relax NG.

TEST DE CONOCIMIENTOS

1 Indicar cuál de las siguientes afirmaciones es cierta:
- a) LaTeX es un lenguaje orientado a presentación.
- b) LaTeX es un lenguaje procedural.
- c) LaTeX es un lenguaje descriptivo.

2 Indicar cuál de las siguientes afirmaciones es cierta:
- a) El consorcio W3C comenzó una iniciativa para intentar dotar a la *web* de un lenguaje más potente y que pudiera dar una estructurar semántica a la misma.
- b) ISO comenzó una iniciativa para intentar dotar a la *web* de un lenguaje más potente y que pudiera dar una estructurar semántica a la misma.
- c) El lenguaje **GML** fue propuesto por el consorcio W3C.

3 Indicar cuál de las siguientes afirmaciones es cierta:
- a) Algunos elementos no tienen contenido. Se les denomina elementos vacíos y no deben llevar etiqueta de fin.
- b) Todos los elementos tienen que llevar obligatoriamente etiqueta de inicio y de final.
- c) Un **atributo** es un nombre que se encuentra dentro de la etiqueta de inicio de un elemento.

4 Indicar cuál de las siguientes afirmaciones es falsa:
- a) En la DTD se identifica la estructura del documento.
- b) Las DTD pueden ser internas o externas a un documento, o ambas cosas a la vez.
- c) En la DTD no se definen todos los elementos.

5 Indicar cuál de las siguientes afirmaciones es falsa:
- a) XML Schema es la evolución de la DTD descrita por el W3C.
- b) A XML Schema también se le denomina XSD.
- c) A XML Schema también se le denomina RELAX NG.

6 Indicar cuál de las siguientes afirmaciones es falsa:
- a) El modelo de datos de XML Schema incluye el vocabulario.
- b) El modelo de datos de XML Schema no incluye los tipos de datos.
- c) El modelo de datos de XML Schema incluye los elementos.