1. Introducción al etiquetado de documentos

Introducción a los lenguajes de marcado

1.1. Fundamentos de los lenguajes de marcado

Los orígenes de los lenguajes de marcado hay que buscarlos en los primeros procesadores de texto y en los procesos de impresión física relacionados. Estas marcas indicaban a las impresoras cómo debían imprimir determinadas partes del texto. Con la generalización de las pantallas, **el uso de estas marcas se amplió para facilitar la presentación visual de la información**. La utilidad de las marcas en múltiples contextos no pasó desapercibida para la industria informática, y en la década de 1960 C. Goldfarb crea, a instancias de IBM, el *GML* (*Generalized Markup Language*), como lenguaje de marcas de formato estandarizado para cualquier clase de documento, independientemente de su contenido. En 1986 se convirtió en el estándar ISO 8879, *SGML* (*Standard Generalized Markup Language*). SGML es muy complejo, por lo que su uso es limitado. La aparición del *HTML*, que usaba un subconjunto limitado de SGML, a comienzos de la década de 1990 y su expansión en el web puso las bases para la aparición, durante esa década, de otros lenguajes de marcado de documentos, como *XML*, *DockBook* o *DITA*, y de un buen número de lenguajes especializados, que hacen uso del marcado.

El conocimiento y la difusión generalizada de los lenguajes de marcado se ha producido como consecuencia del desarrollo del world wide web, desde la década de 1990. Hasta ese momento, su utilización se limitaba al campo de la generación de documentación técnica especializada. Sin embargo, cuando Tim Bernes-Lee busco y desarrolló la manera de preparar documentos para distribuirlos y visualizarlos a través de internet, usando un navegador web, recurrió a los fundamentos del *SGML*, creó el primer *HTML*, y dio paso al web tal y como lo conocemos en la actualidad.

Las páginas o documentos web son, en realidad, documentos etiquetados. Su contenido es textual, y el texto, el contenido informativo y documental, se ve complementado por un **conjunto de etiquetas o marcas**. La lógica subyacente es que cualquier contenido, sea texto, datos, enlaces, etc, se puede marcar, y esa esa marca significará algo para cualquier aplicación que pueda leer ese documento. Si una aplicación, un programa, recibe un documento con etiquetas o marcas, y sabe qué significan esas marcas, entonces puede ejecutar sobre el contenido marcado todas las instrucciones que se le den.

Los lenguajes de marcado se usan en documentos que tienen un contenido informativo y documental. A diferencia de los lenguajes de programación, que sirven para escribir algoritmos (órdenes, secuencias y acciones que una máquina debe ejecutar), **los lenguajes de marcado no se ejecutan: simplemente, etiquetan o marcan elementos**. El marcado o etiquetado de los elementos no es arbitrario: todo lenguaje tiene una estructura lógica. **Cada etiqueta o marca lo que hace es indicar un atributo de aquello que está marcando**.

Los lenguajes de marcado se utilizan para todo en internet. No sólo para las páginas web: gran parte de **la información que circula entre máquinas a través de la red adopta la forma de documentos etiquetados con lenguajes de marcas**. El **intercambio de datos entre aplicaciones se hace usando lenguajes de marcado**. En los ordenadores de escritorio se pueden encontrar muchas aplicaciones que guardan información en ficheros etiquetados, aunque el usuario o usuaria sea ignorante de ello. Los documentos con información etiquetada mediante lenguajes de marcado son omnipresentes.

Los lenguajes de marcado no son para humanos: **son para máquinas**. Aunque es común marcar elementos cuando se está creando un documento (por ejemplo, al crear una página web usando directamente etiquetas o marcas de HTML), son las máquinas y su software el destinatario de estos marcados. En el mismo caso del HTML, se marcan elementos para que el navegador sepa cómo presentarlos al usuario. Al usuario se le aísla de las etiquetas o marcas: cuando se crea un contenido en un blog, por ejemplo, el editor visual lo que está haciendo es trasladar a lenguaje de etiquetado, de manera transparente, lo que el usuario está introduciendo y maquetando.

Material complementario

- Wikipedia: lenguaje de marcado.

1.2. Nociones básicas sobre lenguajes de marcado

La definición más común establece que un lenguaje de marcado o lenguaje de marcas es una forma de codificar un documento que, junto con el texto, incorpora etiquetas o marcas que contienen información acerca de la estructura del texto o de su formato de presentación. Estos lenguajes pueden hacen explícita la estructura del documento que se trate, pueden indicar el contenido semántico, o pueden señalar e indicar cualquier otro tipo de información que pueda ser relevante para un uso dado.

Los lenguajes de marcas se dividen en tres grandes grupos:

- **Lenguajes de presentación**: son aquellos orientados a definir el formato o la capa de presentación del texto. Suelen ocultar las etiquetas y mostrar al usuario solamente el texto con su formato. El conocido RTF para ficheros de texto es un marcado de este tipo.
- **Lenguajes de procedimientos**: orientados también a la presentación, pero además incorporan elementos que la aplicación o programa que representa el documento debe interpretar para ejecutar acciones en función de éstos. El HTML de las páginas web es un ejemplo.
- **Lenguajes descriptivos o semánticos**: son lenguajes diseñados para representar las diferentes partes en las que se estructura un documento, y para definir su contenido. Sin embargo, y a diferencia de los anteriores, no especifican cómo deben representarse los documentos en su capa visual. Son los utilizados para facilitar el intercambio de información y datos entre aplicaciones. XML es el estándar actual para ello.

El funcionamiento de los lenguajes de marcado es simple: un elemento se destaca del resto de información mediante una marca o etiqueta:

El documento marcado se somete a un procesador, que interpreta las marcas y genera un documento final, o bien ejecuta una serie de acciones sobre el contenido etiquetado (presentación en pantalla, incorporación a una base de datos, relación con otros elementos etiquetados...). Es importante señalar que **un mismo documento puede contener al mismo tiempo diferentes lenguajes de marcado**, y que será el procesador de documentos o la aplicación que lo trate, en cada caso, la que decida que debe hacer con el contenido marcado.

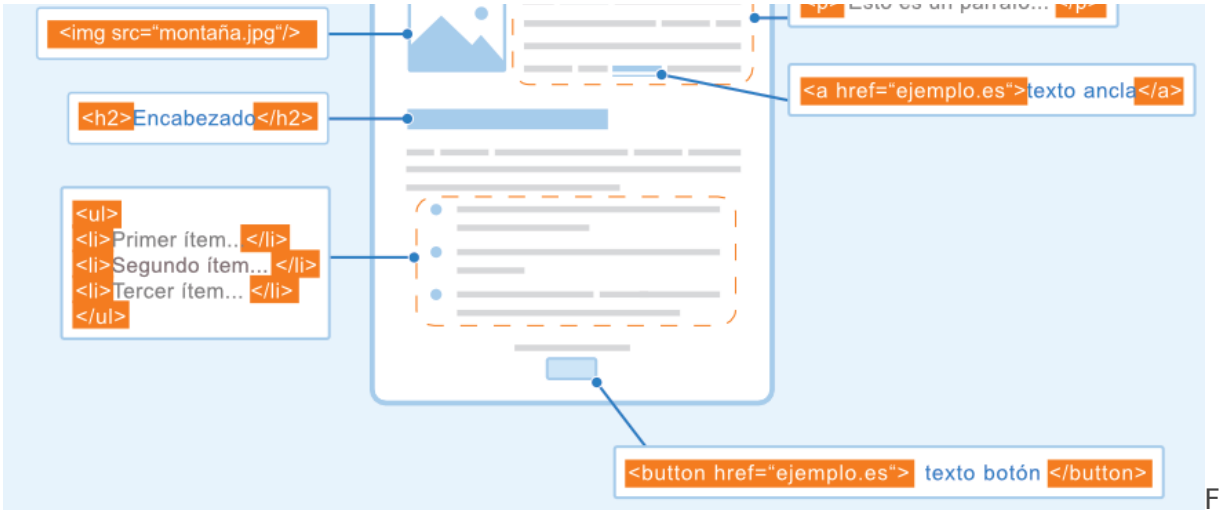

Fig. 1.

Ejemplo de marcado en HTML y su representación visual (fuente original).

Los lenguajes de marcas se usan para etiquetar, para marcar, elementos dentro de un documento. En el párrafo anterior se ha indicado que este documento se somete a un procesador. Ahora bien ¿cómo sabe el procesador cómo procesar, valga la redundancia, las marcas? Esto es posible porque todos y cada uno de **los lenguajes de marcado tienen un documento de referencia en el que se explicitan las reglas sobre cómo se estructuran los documentos, que marcas y etiquetas se usan, lo que significan, y cómo se aplican y relacionan**. Estos documentos de referencia se identifican con las siglas *DTD* (*Document Type Definition*), o *XML Schema*. Por ejemplo, hay un DTD/XML Schema para HTML, otro para XML, etc. Estos documentos de referencia se publican en internet de manera abierta, de forma que cada procesador o aplicación pueda acudir a la url correspondiente, y cargarlo para saber cómo actuar ante cada documento.

El **flujo de trabajo** que se establece es el siguiente:

1. Una persona o una máquina crea un **documento de texto, sobre cuyo contenido aplica un lenguaje de marcado.**
2. Una aplicación o un **procesador accede al contenido** de ese documento.
3. **Identifica**, generalmente en la cabecera o primeras líneas de texto, los **DTD o XML Schema que debe usar para procesar el contenido**.
4. **Utiliza los url de los DTD/XML Schema para** ir a la localización original, y **cargar su contenido** como parámetros de trabajo.
5. Una vez cargados, **procesa el documento y su contenido marcado, de acuerdo con las reglas obtenidas** del DTD/XML Schema.

El resultado de este procesamiento puede ser la visualización de una página web, la creación de un nuevo documento, la incoporación de datos a una base de datos, u otros que puedan haber sido programados en la aplicación correspondiente. Por ejemplo, la posibilidad de crear un documento HTML, o en formato EPUB, desde un documento generado por un procesador de textos, es un ejemplo del uso de lenguajes de marcado y del procesamiento que crea diferentes tipos de resultados en virtud de diferentes reglas de procesamiento, todo ello de forma transparente para el

usuario final.