Unidad 1. Introducción a los lenguajes de marcas

1. ¿Qué es un lenguaje de marcas?

Según Wikipedia: 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 adicional acerca de la estructura del texto o su presentación.

Un lenguaje de marcado utiliza una notación especial para marcar diferentes secciones en un documento. Los documentos creados usando un lenguaje de marcas constan de los c**aracteres de marcado, o etiquetas**, más **texto o datos carácter o caracteres de contenido**. Los caracteres de marcado varían según el lenguaje de marcado, por ejemplo: los símbolos < y >, la coma (,), el guión (-), o incluso código binario se pueden usar para marcar texto.

Para cada lenguaje de marcado, los programadores pueden desarrollar aplicaciones que lean e interpreten documentos escritos con ese lenguaje. Las etiquetas definen cómo debe ser interpretado el texto del documento por las aplicaciones que lo lean.

> ***Ejemplo 1***. La siguiente línea en HTML
>
> *<h1> Introducción</h1>*
>
> Contiene las etiquetas *<h1>* y *</h1>* y el texto *Introducción*. Una aplicación que lea HTML, por ejemplo, un navegador web, al leer esta línea sabrá que el texto *Introducción* debe mostrarse usando el tipo de fuente h1 (heading 1).

2. Un poco de historia

El nombre de lenguaje de marcas viene de la práctica tradicional de marcar los documentos que iban a ser impresos con anotaciones en los márgenes con instrucciones de impresión. Estas instrucciones indicaban el formato de impresión del texto: tipo de letra, tamaño, estilo, etc. A este proceso se le denomina “marcar la página”, y dio lugar a un grupo de marcas estandarizadas. En la época de la imprenta esta tares la realizaban personas conocidas como marcadores. Con el uso de los ordenadores, este concepto de marcado de documentos se ha trasladado de forma similar al mundo de la informática, para referirse a las instrucciones sobre estructura y formato incluidas en los documentos digitales.

El concepto de lenguaje de marcas nace en la industria editorial a finales de los años sesenta. La principal novedad consistía en la separación entre la estructura del texto y la presentación del mismo. Así surgió, entre otros, el estándar denominado *GenCode*.

Durante los años 70 surge de forma independiente el concepto de los lenguajes de marcas tal y como se entienden actualmente, y durante los años 80 se generaliza su uso.

Dentro de la industria editorial cabe mencionar lenguajes como *Scribe* y *TeX*. *Scribe* puede considerarse el primer lenguaje que diferenció claramente la estructura de la presentación, e influyó en el desarrollo de lenguajes posteriores. *TeX* se centra en la estructura detallada del texto y la tipografía. Requiere amplios conocimientos para ser utilizado, por lo que su uso se encuentra sobre todo en entornos académicos, para la publicación de artículos en varias disciplinas científicas. El software más extendido para el uso de TeX es LaTeX.

Sin embardo, la base de los lenguajes de marcas actuales es el lenguaje *GML* (Generalized Markup Language), diseñado a finales de los años 60 por IBM (Charles F. **G**oldarf y colaboradores), como solución para mantener grandes cantidades de documentos. Este lenguaje heredó de GenCode la idea de separación entre presentación y contenido. El marcado se centra en definir la estructura del texto, y no su presentación visual.

GML fue un gran éxito, y pronto se extendió en otros ámbitos, siendo adoptado por el gobierno de Estados Unidos. Así, surgió la necesidad de estandarizarlo. En 1980, a partir de GML, surge la primera versión de SGML Standard Generalized Markup Language). En 1986 la ISO (International Standars Organization) publica el estándar *SGML*, con el código ISO 8879.

SGML es un lenguaje de marcas independiente de cualquier aplicación. Esto presenta la gran ventaja de que los documentos escritos con SGML pueden ser compartidos por un gran y variado número de usuarios, por ejemplo, clientes, proveedores, y diferentes departamentos dentro de una organización.

SGML no define etiquetas específicas, sino que permite que el autor del documento emplee las marcas que desee, pudiendo así elegir nombres de etiquetas que sean significativos para cada documento en cuestión. SGML describe la sintaxis que deben seguir las marcas que se incluyen en los textos, así como la estructura permitida para los documentos. Esto último se realiza mediante **Document Type Definitions (DTDs)** o **esquemas.** Por tanto, SGML es en realidad un **metalenguaje**, del que derivan otros lenguajes de marcas (como es el caso de XML y HTML). Estos lenguajes se pueden considerar aplicaciones de SGML.

SGML es un lenguaje muy versátil y potente, pero tiene la desventaja de ser complejo y difícil de aprender y usar. La especificación de SGML tiene más de 500 páginas, con más de 100 anexos, y está orientada a sistemas grandes y complejos. Por estos motivos, aunque SGML tuvo una gran aceptación, hoy día se emplea sólo en campos que requieran documentación a gran escala. Sin embargo, SGML ha sido un lenguaje clave en el desarrollo de los lenguajes de marcas, ya que la gran mayoría derivan de él.

Con el desarrollo de Internet surgieron nuevos retos. Así Tim Berners-Lee, investigador en el CERN (*Conseil Européen pour la Recherche Nucléaire*), se encontró con un problema: grandes cantidades de información, que había que organizar, enlazar y hacer accesible y compatible entre unos sistemas y otros. Como respuesta a estos problemas nació, en 1989, el ***HTML* (Hypertext Markup Language)**, también como una aplicación de SGML. HTML se caracteriza por su sencillez, de tal manera que cualquier usuario sin grandes conocimientos de informática puede crear documentos en este formato. Esta característica sin duda contribuyó a su espectacular éxito, de manera que HTML es tal vez el formato de documento más empleado en el mundo, y que sin duda ha contribuido a promover el amplio uso de Internet. Sin embargo, este crecimiento exponencial durante los años 90 dio lugar a la producción de gran cantidad de documentos HTML, pero muchos de ellos mal estructurados. A esto se unen otros problemas del HTML como son la competencia entre distintas empresas y la falta de adherencia a un estándar, que resultó en incompatibilidades entre navegadores web. HMTL ha sido un buen lenguaje para el desarrollo inicial de Internet, pero con el avance y desarrollo del mismo surgió la necesidad de un lenguaje que pueda usarse para propósitos más complejos y de mayor escala.

En 1996 el World Wide Web Consortium (W3C) comenzó a desarrollar un nuevo lenguaje de marcas estándar que fuera más sencillo de utilizar que SGML, pero con una estructura más rígida que HTML, comenzando así el proceso de desarrollo de ***XML (eXtensible Markup Language)***. XML es también un metalenguaje que sigue los principios de SGML, es decir, todo documento XML es a su vez SGML. Al igual que SGML, XML permite crear etiquetas adaptadas a las necesidades de los documentos (de ahí el calificativo de eXtensible). El estándar XML define la sintaxis de las etiquetas y qué se puede hacer con ellas, y es especialmente estricto en cuanto a los requisitos que deben cumplir los documentos escritos en XML. XML logra un equilibrio entre simplicidad y flexibilidad, solventando a su vez los problemas de HTML como los de la internacionalización o falta de estándares. Existe una gran variedad de lenguajes basados en XML para aplicaciones en muy diversos campos y disciplinas, como pueden ser la transacción de datos entre servidores, el intercambio de información financiera, y, como no, la presentación de documentos en entornos web. El ejemplo más claro es el lenguaje **XHTML**, que es una redefinición de HTML que cumple las reglas sintácticas de XML.

3. Clasificación de los lenguajes de marcado

Un lenguaje de marcado cumple con dos objetivos esenciales a la hora de diseñar y procesar un documento digital:

1. Especifica las operaciones tipográficas y las funciones que debe ejecutar el programa navegador/visualizador sobre dichos elementos. Las operaciones tipográficas son instrucciones de formato que se aplican a cada uno de los elementos de un documento digital como, por ejemplo, imprimir un título en negrita y a un determinado tamaño.

2. Separa un texto en los elementos de los que se compone, como por ejemplo un párrafo, un capítulo, un encabezamiento, etc.

Así, pues, se pueden diferenciar 2 tipos de marcación:

1) De presentación: describe cómo ha de formatearse el documento: la fuente, el tamaño, el color, etc.

2) Estructural: describe la estructura del documento: el titular, los párrafos, etc.

Atendiendo a este criterio, se pueden distinguir 3 tipos de lenguajes de marcado:

a) **Procedimentales**: las marcas del documento describen operaciones tipográficas, es decir, la forma y el significado de las operaciones tipográficas que van a ser aplicadas a cada uno de los elementos del documento. Por ejemplo, una regla de un lenguaje de procedimiento indicaría que el título de la sección de un texto debe ser impreso en una sola línea con una fuente de seis puntos más grande que el resto del texto. Se refiere, pues, a la apariencia física o formato (fuente, estilo de letra, tamaño, etc.) con que debe presentarse el documento. HTML es un ejemplo de un lenguaje procedimental.

b) **Estructurales o descriptivos**: En los lenguajes estructurales las marcas describen la estructura lógica del documento y/o la descripción del contenido, pero no su tipografía, es decir, cómo deben ser representados y en qué orden. A esta categoría pertenecen SGML y XML.

c) **Híbridos**: forman una combinación de los dos anteriores. Ejemplo: VRML (Virtual Reality Modelling Language)

4. Características comunes

Aunque existen una gran variedad de lenguajes de marcado para muy diversos usos, se pueden identificar una serie de características comunes:

- **Texto plano:** los documentos pueden ser interpretados directamente, sin necesidad de un programa intermediario (como ocurre con los archivos binarios)
- **Independientes de la plataforma, sistema operativo o programa con el que fueron creados:** esta es una gran ventaja y una de las principales razones de su éxito
- **Compactibilidad:** en el mismo archivo se incluyen los caracteres de marcado junto con los datos carácter
- **Sencillez y facilidad de uso**
- **Flexibilidad:** Aunque originalmente los lenguajes de marcas se idearon para documentos de texto, se han empezado a utilizar en multitud de áreas como servicios web, sindicación web o interfaces de usuario, entre muchas otras. Estas nuevas aplicaciones aprovechan la sencillez y potencia del lenguaje XML. Esto ha permitido que se pueda combinar varios lenguajes de marcas diferentes en un único archivo.

- **Fácil procesamiento por parte de las aplicaciones que los manejan**.

5. Ejemplos de lenguajes de marcado y ámbitos de aplicación

a) Documentación electrónica: RTF, TeX, Wikitexto, DocBook

b) Tecnologías de Internet:

- *World Wide Web:* HTML, XHTML
- *Sindicación de contenidos*: RSS
- *Servicios Web:* WSDL, XML-RPC

c) Lenguajes especializados:

- *Gráficos:* SVG, VML
- *Matemáticas:* MathML
- *Música:* musicXML
- *Voz:* voiceXML
- *etc.*