TEMA 1. Introducción: Los Lenguajes de Marcas

1.- Lenguajes de marcas. Historia.

Los lenguajes de programación son aquellos lenguajes artificiales que permiten escribir código para resolver problemas de una manera intuitiva. Ejemplos de lenguajes de programación de alto nivel son: Java, C, C++, Pascal, Delphi, ASP, Visual C, etc.

Estos lenguajes de programación tienen estructuras de código más o menos complejas como son los bucles (do-while, for, repeat) o los saltos condicionales (if, switch).

Frente a los lenguajes de programación surgieron otros lenguajes sencillos cuyo objetivo inicial era simplemente darle formato a un texto, mediante un conjunto de marcas de formato. Éstos últimos se llaman **lenguajes de marcas**, y realmente no podemos considerarlos lenguajes de programación puesto que no tienen la complejidad de éstos, sin embargo han ido evolucionando para ofrecer más prestaciones y se han convertido en la base para la creación de las páginas WEB.

Recuerda:

- Un 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.

En los años 60, IBM intentó resolver sus problemas asociados al tratamiento de documentos en diferentes plataformas a través de un lenguaje de marcas denominado **GML** (*Generalized markup Language* o Lenguaje de marcas generalizado).

GML libera al creador del documento de preocupaciones específicas de formato como pueden ser el tipo de letra, la alineación de los párrafos, el espaciado entre líneas, el uso de tablas, etc. GML fue una manera de estandarizar estos formatos de forma que un mismo documento se pudiera enviar a una impresora, mostrar en diferentes pantallas, etc.

Más tarde GML pasó a manos de ISO y se convirtió en **SGML** (ISO 8879), *Standart Generalized Markup Language*. Esta norma es la que se aplica desde entonces a todos los lenguajes de marcas, cuyos ejemplos más conocidos son el HTML y el RTF.

Conviene repetir que los lenguajes de marcas no son equivalentes a los lenguajes de programación aunque se definan igualmente como "lenguajes". Son sistemas de descripción de información, normalmente documentos, que si se ajustan a SGML, se

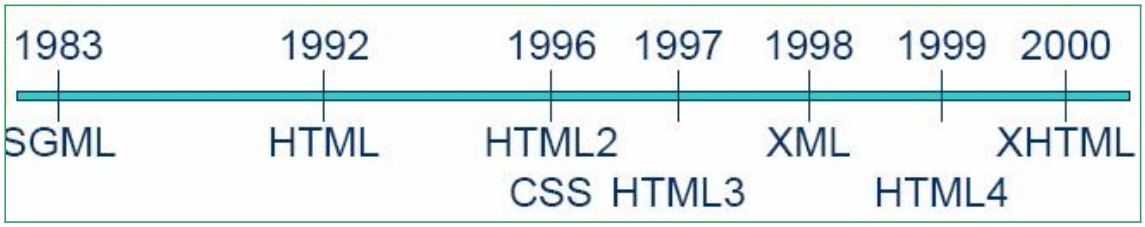
pueden controlar desde cualquier editor de texto ASCII.

Las **marcas o etiquetas** más utilizadas suelen representarse por textos descriptivos encerrados entre signos de "menor" (<) y "mayor" (>), siendo lo más usual que exista una marca de principio y otra de final.

Ejemplo de la estructura básica de un documento HTML (página web):

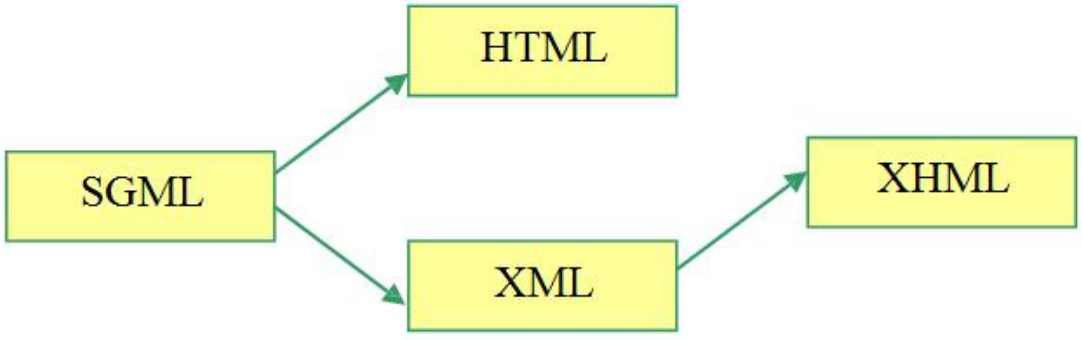
```
<HTML>
  <HEAD>
     <TITLE>Mi primera página</TITLE>
  </HEAD>
  <BODY>
     Esta es la primera página web
  </BODY>
</HTML>
```

La evolución histórica de los lenguajes de marcas más utilizados es la siguiente:



En este módulo iremos conociendo estos lenguajes en el siguiente orden: primero analizaremos el HTML, lenguaje que todavía en la actualidad es la base de escritura de casi todas las páginas WEB. Luego veremos el XHTML que es una evolución del HTML que incorpora las ventajas de XML. Y por último nos centraremos en el XML y todo el abanico de posibilidades asociadas a él, que están revolucionando el intercambio de datos entre aplicaciones, y que es el futuro inmediato de Internet.

No obstante no podemos perder de vista que históricamente la evolución de estos lenguajes sigue el siguiente diagrama:



2.- Otros ejemplos de lenguajes de marcas.

SGML (Standard Generalized Markup Language)

Es un ejemplo de lenguaje genérico que apareció con el identificador 8879 como norma ISO (International Organization for Standarization). Lo creó la comunidad de editores e imprentas, ya que consideraban que era de vital importancia contar con una manera normalizada de transmitir los documentos en un formato adecuado para los procesos de edición e impresión.

SGML es apropiado para describir texto altamente estructurado, aunque también se pueden incluir en los documentos otros elementos, como por ejemplo diagramas y gráficos, independientemente de su formato de codificación.

SGML contiene las reglas para crear una infinita variedad de lenguajes de marcado, pero no describe el formato de los documentos marcados. Esto significa que SGML sólo aporta reglas para definir nuestros propios lenguajes de marcas, es decir, es un **metalenguaje**. Esto hace posible que, mediante la utilización de una definición de tipo de documento (denominada DTD Document Type Definition), se pueda especificar la estructura lógica de una clase de escrito.

Resumiendo SGML no es un lenguaje en sí, sino una manera de crear otros lenguajes. La dificultad de su uso hizo que dejara de utilizarse, pero es el padre de otros lenguajes que utilizaremos nosotros como HTML y XML.

WML (Wireless Markup Language)

Con el auge de las páginas web, pronto se comprendió que era necesario contar con un mecanismo que las hiciera visibles en los teléfonos móviles. Para ello fue necesario trabajar en un protocolo de comunicaciones adaptado a estos dispositivos. Fue así como surgió el protocolo WAP (Wireless Application Protocol) que permite el desarrollo de aplicaciones sobre dispositivos móviles a través de redes inalámbricas

La tecnología WAP permite que los usuarios de estos dispositivos puedan acceder a servicios disponibles en Internet. Sin embargo, existen algunas consideraciones a tener en cuenta al diseñar estos servicios para usuarios móviles, fundamentalmente debidas a las características de los terminales: pantalla significativamente más pequeña que la de un ordenador personal, teclados más limitados que los de un ordenador, limitaciones en la memoria disponible, tanto memoria RAM como memoria para almacenamiento persistente, y limitaciones en la capacidad del procesador, en comparación con la memoria y procesador de un ordenador personal típico. Las redes de telefonía móvil ofrecen también unas prestaciones por lo general menores que los accesos a Internet, si bien con las redes de tercera generación como UMTS las prestaciones mejoran de manera importante..

Una vez desarrollado este protocolo fue necesario crear un lenguaje de marcas adaptado. WML y WMLScript son los equivalentes dentro del mundo “inalámbrico” al HTML y al JavaScript dentro de las redes que usan el protocolo TCP/IP como hace Internet.

Los documentos WML pueden mostrarse en teléfonos móviles, pero también en cualquier otro dispositivo que contenga un micronavegador

3.- Características de los lenguajes de marcas.

1.- Uso del texto plano.

Los archivos de texto plano son aquellos que están compuestos únicamente por texto sin formato, sólo caracteres. Estos caracteres se pueden codificar de distintos modos dependiendo de la lengua usada. Algunos de los sistemas de codificación de caracteres más usados son: ASCII, ISO-8859-1, Latín-1, Unicode, etc...

Los lenguajes de marcas se escriben en archivos de texto plano, y como principal ventaja es que estos archivos pueden ser interpretados directamente. Esto es una ventaja evidente respecto a los sistemas de archivos binarios, que requieren siempre de un programa intermediario para trabajar con ellos que lo interprete.

Un documento escrito con lenguajes de marcado puede ser editado por un usuario con un sencillo editor de textos, sin perjuicio de que se puedan utilizar programas más sofisticados que faciliten el trabajo.

Al tratarse solamente de texto, los documentos son independientes de la plataforma, sistema operativo o programa con el que fueron creados. Esta fue una de las premisas de los creadores de GML en los años 60, para no añadir restricciones innecesarias al intercambio de información. Es una de las razones fundamentales de la gran aceptación que han tenido en el pasado y del excelente futuro que se les augura.

Nota: Microsoft Word es un editor de textos, pero no de texto plano, en principio guarda sus archivos con un formato propio. Para generar archivos de texto plano la opción más sencilla es usar el Notepad o Bloc de notas en entornos Windows (está en el menú accesorios).

2.- Compactos.

Las etiquetas o marcas se entremezclan con el propio contenido en un único archivo. Veamos un ejemplo en diferentes lenguajes de marcas:

Ejemplos	HTML	LaTeX	Wikitexto
Titulo	Titulo	\section{Titulo}	== Titulo ==
Lista	 - Punto 1 - Punto 2 - Punto 3 	\begin{itemize} \item Punto 1 \item Punto 2 \item Punto 3 \end{itemize}	* Punto 1 * Punto 2 * Punto 3
texto en **negrita**	texto	\bf{texto}	''' texto '''
texto en *cursiva*	texto	\it{texto}	'' texto ''

En estos ejemplos vemos cómo diferentes lenguajes de marcas usan sus propios conjuntos de marcas. Por ejemplo si queremos que una parte de un texto aparezca en **negrita** y estamos en HTML tendremos que marcarlo así:

Texto en negrita

Es decir, en HTML la marca es para la negrita.

Si por el contrario estamos usando LaTeX tendremos que poner la marca \bf, y en Wikitexto encerramos el texto entre tres comillas simples.

Éste es el concepto de **marca o etiqueta**. Afectan a la presentación visual (negrita, cursiva, etc.), o a la estructuración del texto (listas, tablas, etc.).

3.- Fáciles de procesar:

Las organizaciones de estándares han venido desarrollando lenguajes especializados para los tipos de documentos de comunidades o industrias concretas. Uno de los primeros fue el CALS, utilizado por las fuerzas armadas de EE.UU. para sus manuales técnicos. Otras industrias con necesidad de gran cantidad de documentación, como las de aeronáutica, telecomunicaciones, automoción o hardware, han elaborado lenguajes adaptados a sus necesidades.

Esto ha conducido a que sus manuales se editen únicamente en versión electrónica, y después se obtenga a partir de ésta las diferentes versiones impresas, en línea o en CD. Un ejemplo notable fue el caso de Sun Microsystems, empresa que optó por escribir la documentación de sus productos en SGML, ahorrando costes considerables. El responsable de aquella decisión fue Jon Bosak, que más tarde fundaría el comité para el desarrollo del XML.

4.- Flexibles.

Aunque originalmente los lenguajes de marcas se idearon para documentos de texto, se han empezado a utilizar en áreas como gráficos vectoriales, servicios web, sindicación web o interfaces de usuario. 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, como en el caso de XHTM+SVG.

4.- Clasificación de los lenguajes de marcas.

4.1.- Según la naturaleza de las marcas:

1.- De presentación.
2.- De procedimientos.
3.- Semánticos.

1.- Lenguajes de presentación:

Son aquellos que están especialmente diseñados para indicar formatos del texto, es decir, la forma o apariencia que adquirirá el texto (la presentación). Por lo tanto sirven para ver los textos adecuadamente pero no resultan de utilidad para procesar de forma automática la información que contienen.

El marcado de presentación resulta más fácil de elaborar, sobre todo para cantidades pequeñas de información. Sin embargo resulta complicado de mantener o modificar, por lo que su uso se ha ido reduciendo en proyectos grandes en favor de otros tipos de marcado más estructurados.

Se puede tratar de averiguar la estructura de un documento de esta clase buscando pistas en el texto. Por ejemplo, el título puede ir precedido de varios saltos de línea, y estar ubicado centrado en la página. Varios programas pueden deducir la estructura del texto basándose en esta clase de datos, aunque el resultado suele ser bastante imperfecto, por lo que no se utilizan para analizar información, sólo para presentarla.

Ejemplos: Rich Text Format (RTF), S 1000D, TeX, troff, HTML…

2.- Lenguajes de procedimientos:

Aunque también suelen incorporar marcas para la presentación, su objetivo general es indicar los procedimientos que deberá realizar el software de representación. El marcado de procedimientos está enfocado hacia la presentación del texto, sin embargo, también es visible para el usuario que edita el texto. El programa que representa el documento debe interpretar el código en el mismo orden en que aparece.

Por ejemplo, para formatear un título, debe haber una serie de directivas inmediatamente antes del texto en cuestión, indicándole al software instrucciones tales como centrar, aumentar el tamaño de la fuente, o cambiar a negrita. Inmediatamente después del título deberá haber etiquetas inversas que reviertan estos efectos.

Ejemplos: nroff, troff, TeX.

Este tipo de marcado se ha usado extensamente en aplicaciones de edición profesional, manipulados por tipógrafos calificados, ya que puede llegar a ser extremadamente complejo.

3.- Lenguajes semánticos:

No se encargan de la presentación, sino de la estructura de la información almacenada. Es decir, describen las diferentes partes en las que se estructura el documento pero sin especificar cómo deben representarse.

El marcado descriptivo o semántico utiliza etiquetas para describir los fragmentos de texto, pero sin especificar cómo deben ser representados, o en qué orden.

Los lenguajes expresamente diseñados para generar marcado descriptivo son el SGML y el XML, así como el nuevo HTML5.

Las etiquetas pueden utilizarse para añadir al contenido cualquier clase de metadatos. Por ejemplo, el estándar Atom, un lenguaje para la sindicación de contenidos (RSS), proporciona un método para marcar la hora "actualizada", que es el dato facilitado por el editor de cuándo ha sido modificada por última vez cierta información. El estándar no especifica cómo se debe representar, o siquiera si se debe representar. El software puede emplear este dato de múltiples maneras, incluyendo algunas no previstas por los diseñadores del estándar.

Una de las virtudes del marcado descriptivo es su flexibilidad: los fragmentos de texto se etiquetan tal como son, y no tal como deben aparecer. Estos fragmentos pueden utilizarse para más usos de los previstos inicialmente. Por ejemplo, los hiperenlaces fueron diseñados en un principio para que un usuario que lee el texto los pulse. Sin embargo, los buscadores los emplean para localizar nuevas páginas con información relacionada, o para evaluar la popularidad de determinado sitio web.

El marcado descriptivo también simplifica la tarea de reformatear un texto, debido a que la información del formato está separada del propio contenido. Por ejemplo, un fragmento indicado como cursiva (texto), puede emplearse para marcar énfasis o bien para señalar palabras en otro idioma. Esta ambigüedad, presente en el marcado de presentación y en el procedimental, no puede soslayarse más que con una tediosa revisión a mano. Sin embargo, si ambos casos se hubieran diferenciado descriptivamente con etiquetas distintas, podrían representarse de manera diferente sin esfuerzo.

El marcado descriptivo está evolucionando hacia el marcado genérico. Los nuevos sistemas de marcado descriptivo estructuran los documentos en árbol, con la posibilidad de añadir referencias cruzadas. Esto permite tratarlos como bases de datos, en las que el propio almacenamiento tiene en cuenta la estructura

Ejemplos: XML, SGML, ASN.1, EBML, YAML.

Ejemplo de filosofía de XML:

```
<persona>
    <nombre>Javier</nombre>
    <apellido>González</apellido>
    <apellido>Márquez</apellido>
    <edad>34</edad>
</persona>
```

Podemos ver cómo se estructura una persona, pero no tenemos información de cómo mostrar sus datos en la pantalla.

4.2.- Según el uso que se da al lenguaje:

1.- Elaboración de documentos electrónicos:

- RTF
- TeX
- Wikitexto
- DocBook

2.- Uso en Internet (páginas web, canales de noticias, etc.):

- HTML, XHTML
- RDF
- RSS

3.- Especializados en ámbitos:

- MathML
- VoiceXML
- SVG
- MusicXML