2022

GRUPO STUDIUM ciclos

Tema 1: Temario

INTRODUCCIÓN A LOS LENGUAJES DE MARCAS

STUDIUM

www.grupostudium.com
informacion@grupostudium.com
954 539 952

1.1 Lenguajes de marcas

Los lenguajes de marcas (también llamados lenguajes de marcado) son aquellos que combinan la información, generalmente textual, con marcas o anotaciones relativas a la estructura del texto o a la forma de representarlo.

Especifica cuáles serán las etiquetas posibles, dónde deben colocarse y el significado que tendrá cada una de ellas.

El lenguaje de marcas más extendido es el HTML (HyperText Markup Language, lenguaje de marcado de hipertexto), fundamento del World Wide Web (entramado de comunicación de alcance mundial).

Los lenguajes de marcado suelen confundirse con lenguajes de programación. Sin embargo, no son lo mismo, ya que el lenguaje de marcado no tiene funciones aritméticas o variables, como poseen los lenguajes de programación. Históricamente, el marcado se usaba y se usa en la industria editorial y de la comunicación, así como entre autores, editores e impresores.

Un ejemplo de cómo funciona el lenguaje de marcado puede observarse en el dictado de viva voz de un documento a una persona que lo transcribe a máquina:

Ponga *estilo de carta*, ponga *comillas*, ponga *mayúsculas*, Estimado Juan, ponga *dos puntos, aparte*, *sangría*, ponga *primera letra mayúscula*, te escribo esta carta, ponga *negrillas*, de forma muy urgente, cierre *negrilla*, ya que no me has enviado..., etc.

Ejemplo: Una noticia representada mediante un lenguaje de marca podría ser así:

```
<noticia>
        <lugar>Madrid</lugar>
        <fecha>27/08/2021</fecha>
        <desc>Se ha inaugurado una estación de tren</desc>
</noticia>
```

Los lenguajes de marcas se utilizan para cualquier tipo de documento (textos, presentaciones, gráficos, tecnologías de Internet, matemáticas, música, multimedia, ...).

1.2 Clases de lenguajes de marcas

Se suele diferenciar entre tres clases de lenguajes de marcado, aunque en la práctica pueden combinarse varias clases en un mismo documento:

- Marcado de Presentación: Es aquel que indica el formato del texto, sólo se muestra la presentación, pero es difícil extraer información.
- Marcado de Procedimientos: Se incluyen instrucciones de cómo hay que procesar el texto.
- Marcado Descriptivo o Semántico: Utiliza etiquetas para describir los fragmentos de texto, pero sin especificar cómo deben ser representados, o en qué orden.

1.2.1 Lenguajes de marcado de Presentación

Son los usados tradicionalmente por los procesadores de texto como puede ser Microsoft Word®.

Codifican cómo ha de presentarse el documento, por ejemplo, indicando que una determinada palabra debe presentarse en fuente itálica.

Generalmente se ocultan al usuario lo que permite obtener un efecto WYSIWYG (What You See Is What You Get).

Este tipo de lenguajes de marcas no suelen ser flexibles ni reusables.

1.2.2 Marcado de Procedimientos

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.

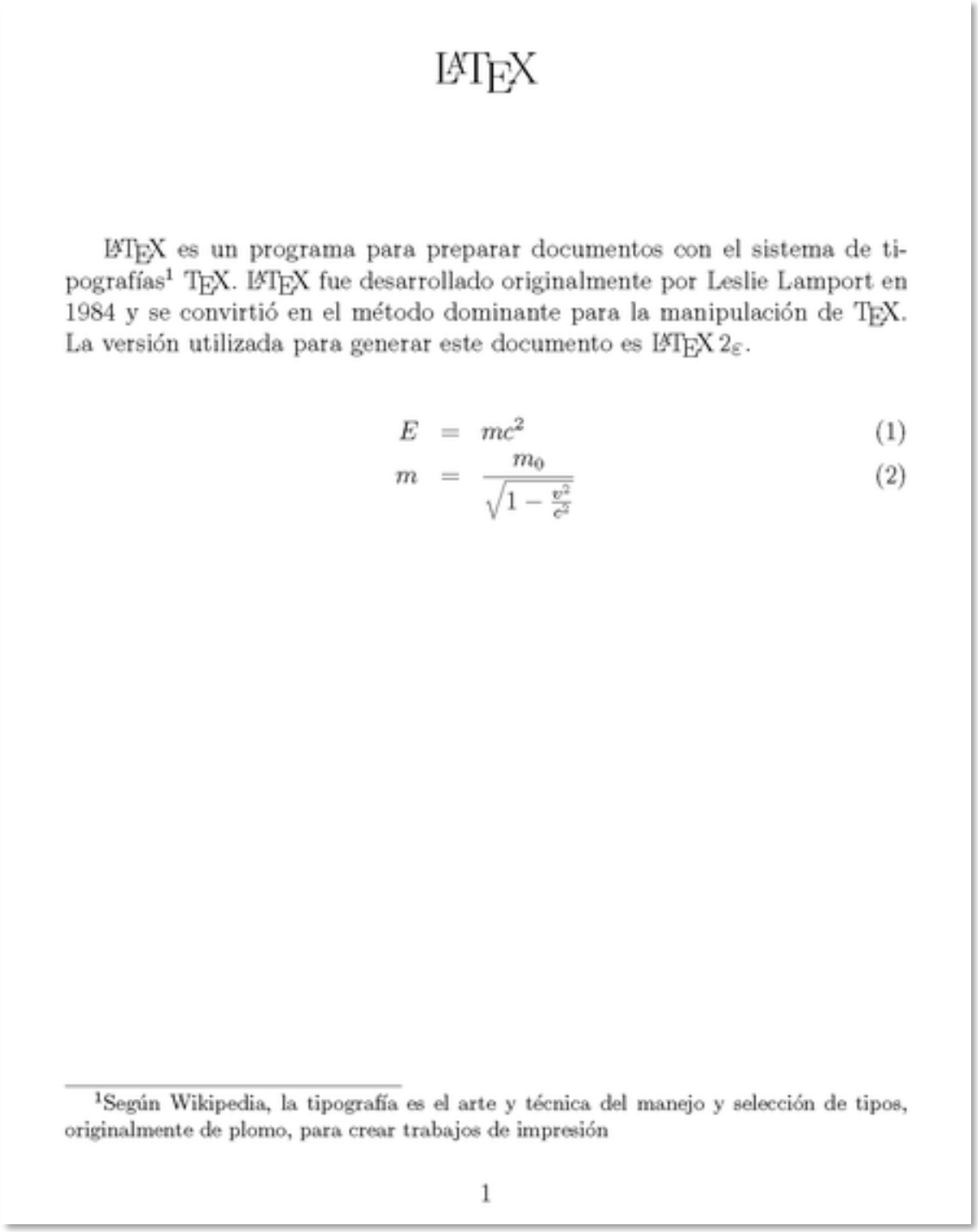
El siguiente código LaTex...

```
\documentclass[12pt]{article}
\usepackage[spanish]{babel}
\usepackage{amsmath}
\title{\LaTeX}
\date{}
% Este es un comentario, no será mostrado en el documento final.
\begin{document}
\maketitle

\LaTeX{} es un programa para preparar documentos con  el sistema de
tipograf\'ias\footnote{%nota al pie de página Seg\'un Wikipedia, la
tipograf\'ia es el arte y t\'ecnica del manejo y selecci\'on de tipos,
originalmente de plomo, para crear trabajos de impresi\'on } %fin nota al pie de página
\TeX{}. \LaTeX{} fue desarrollado originalmente por Leslie Lamport en
1984 y se convirti\'o en el m\'etodo dominante para la  manipulaci\'on
de \TeX. La versi\'on utilizada para generar  este documento es \LaTeXe.
\newline
% El siguiente código muestra la calidad de la tipografía de LaTeX
\begin{align}
E &= mc^2                                  \\
m &= \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}
\end{align}
\end{document}
```


www.grupostudium.com
informacion@grupostudium.com
954 539 952

...genera como resultado...



LaTeX

LaTeX es un programa para preparar documentos con el sistema de tipografías[1] TeX. LaTeX fue desarrollado originalmente por Leslie Lamport en 1984 y se convirtió en el método dominante para la manipulación de TeX. La versión utilizada para generar este documento es LaTeX 2_ε.

$$E = mc^2 \quad (1)$$

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (2)$$

[1]Según Wikipedia, la tipografía es el arte y técnica del manejo y selección de tipos, originalmente de plomo, para crear trabajos de impresión

1

1.2.3 Marcado Descriptivo o Semántico

Las marcas sirven para indicar qué es esa información, es decir, describen qué 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.).

1.3 Características de los lenguajes de marcas

Ya se ha comentado antes cómo diferencias un Lenguaje de Marcado de un Lenguaje de Programación. Veamos ahora las principales características de los Lenguajes de Marcado:

- Uso de texto plano
- Compacidad: las instrucciones de marcado se entremezclan con el propio contenido
- Facilidad de procesamiento
- Flexibilidad

1.4 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 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.

SGML surgió para permitir compartir información por parte de sistemas informáticos.

Tuvo una gran aceptación, pero no consiguió asentarse del todo debido principalmente a su complejidad lo que provocaba que el software que usaba SGML terminaba siendo excesivamente extenso y complejo.

A finales de los 80 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 SGML y lo denominaron HTML (Hyper-text Markup Language).

Supuso una revolución en la forma de compartir información, gracias principalmente a la sencillez de su 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.

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.

1.4.1 GML (Generalized Markup Language)

Uno de los problemas que se conocen desde hace décadas en la informática es la falta de estandarización en los formatos de información usados por los distintos programas.

Para resolver este problema, en los años 60 IBM encargó a Charles F. Goldfab la construcción de un sistema de edición, almacenamiento y búsqueda de documentos legales.

Tras analizar el funcionamiento de la empresa llegaron a la conclusión de que para realizar un buen procesado informático de los documentos había que establecer un formato estándar para todos los documentos que se manejaban en la empresa. Con ello se lograba gestionar cualquier documento en cualquier departamento y con cualquier aplicación, sin tener en cuenta dónde ni con qué se generó el documento.

El formato de documentos que se creó como resultado de este trabajo fue GML, cuyo objetivo era describir los documentos de tal modo que el resultado fuese independiente de la plataforma y la aplicación utilizada.

1.4.2 SGML (Standard Generalized Markup Language)

El formato GML evolucionó hasta que en 1986 dio lugar al estándar ISO 8879 que se denominó SGML. Éste era un lenguaje muy complejo y requería de unas herramientas de software caras. Por ello su uso ha quedado relegado a grandes aplicaciones industriales.

```
<email>
        <remitente>
                <persona>
                <nombre>Pepito</nombre>
                <apellido>Grillo</apellido>
                </persona>
        </remitente>
        <destinatario>
                <direccion>pepito_grillo@hotmail.com</direccion>
        </destinatario>
        <asunto>¿Quedamos?</asunto>
        <mensaje>Hola, he visto que ponen esta noche la película que querías ver.
¿Te apetece ir?</mensaje>
</email>
```

Un documento SGML consta de dos partes: El prólogo (contiene la declaración de que es un documento SGML y la definición del tipo de documento para indicar la sintaxis), y la instancia del documento que contiene los datos en sí.

En el ejemplo anterior hemos obviado el prólogo, pero podemos destacar el vocabulario que usamos y las reglas:

- Vocabulario: email, remitente, persona, nombre, apellido, destinatario, dirección, asunto, mensaje.
- Reglas: email contiene a remitente, destinatario, asunto y mensaje. Remitente contiene al elemento persona que a su vez contiene a nombre y apellido. Destinatario solo contiene a dirección.

1.4.3 HTML (HyperText Markup Language)

En 1989, Tim Berners-Lee creó el World Wide Web y se encontró con la necesidad de organizar, enlazar y compatibilizar gran cantidad de información procedente de diversos sistemas.

Para resolverlo creó un lenguaje de descripción de documentos llamado HTML, que, en realidad, era una combinación de dos estándares ya existentes:

- ASCII: Es el formato que cualquier procesador de textos sencillo puede reconocer y almacenar. Por tanto, es un formato que permite la trasferencia de datos entre diferentes ordenadores.
- SGML: Lenguaje que permite dar estructura al texto, resaltando los títulos o aplicando diversos formatos al texto.

HTML es una versión simplificada de SGML, ya que sólo se utilizaban las instrucciones absolutamente imprescindibles.

Era tan fácil de comprender que rápidamente tuvo gran aceptación logrando lo que no pudo SGML, HTML se convirtió en un estándar general para la creación de páginas web.

Además, tanto las herramientas de software como los navegadores que permiten visualizar páginas HTML son cada vez mejores.

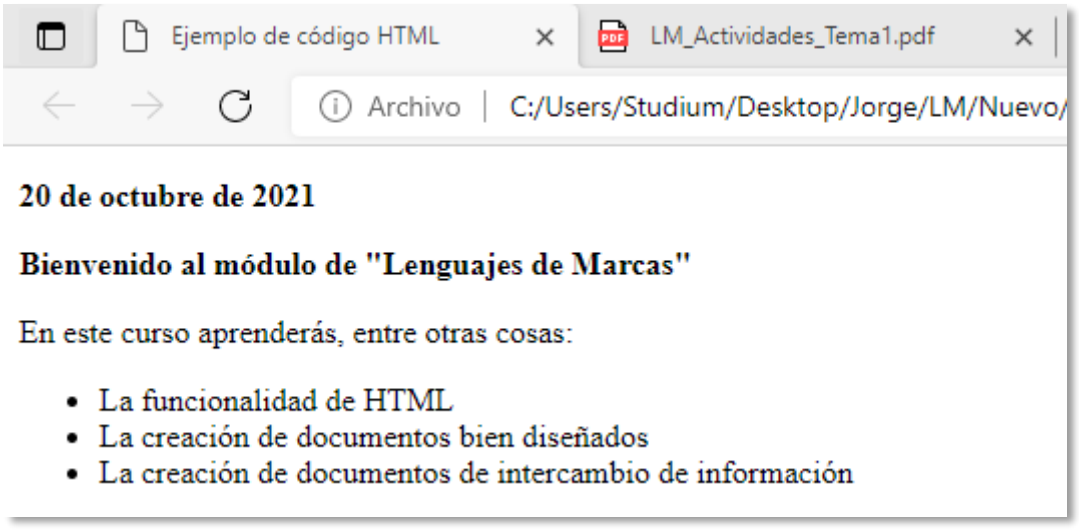
A pesar de todas estas ventajas HTML no es un lenguaje perfecto, sus principales desventajas son:

- No soporta tareas de impresión y diseño
- El lenguaje no es flexible, ya que las etiquetas son limitadas
- No permite mostrar contenido dinámico
- La estructura y el diseño están mezclados en el documento

El siguiente código...

```
<!DOCTYPE HTML>
<html>
	<head>
		<title> Ejemplo de código HTML</title>
	</head>
	<body bgcolor="#ffffff">
		<p></p>
		<p>
			<b>20 de octubre de 2021</b>
		</p>
		<p>
			<b> Bienvenido al módulo de "Lenguajes de Marcas"</b>
		</p>
		<p> En este curso aprenderás, entre otras cosas:<br/>
			<ul>
				<li>La funcionalidad de HTML </li>
				<li>La creación de documentos bien diseñados </li>
				<li>La creación de documentos de intercambio de información</li>
			</ul>
		</p>
	</body>
</html>
```

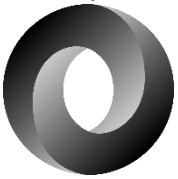
...genera, en el Navegador “Microsoft Edge” ...



www.grupostudium.com
informacion@grupostudium.com
954 539 952

1.4.4 XML (eXtensible Markup Language)

Para resolver los problemas de HTML,antes descritos, el W3C[1] establece, en 1998, el estándar internacional XML, un lenguaje de marcas puramente estructural que no incluye ninguna información relativa al diseño. Se convirtió con rapidez en el estándar para el intercambio de datos en la Web. Hoy en día, solamente amenaza su dominio un tal JSON, que también estudiaremos más adelante.



A diferencia de HTML, las etiquetas indican el significado de los datos en lugar del formato con el que se van a visualizar los datos.

XML es un metalenguaje caracterizado por:

- Permitir definir etiquetas propias.
- Permitir asignar atributos a las etiquetas.
- Utilizar un esquema para definir de forma exacta las etiquetas y los atributos.
- La estructura y el diseño son independientes.

En realidad, XML es un conjunto de estándares relacionados entre sí y que son:

- XSL, eXtensible Style Language. Permite definir hojas de estilo para los documentos XML e incluye capacidad para la transformación de documentos.
- XML Linking Language, incluye Xpath, Xlink y Xpointer. Determinan aspectos sobre los enlaces entre documentos XML.
- XML Namespaces. Proveen un contexto al que se aplican las marcas de un documento de XML y que sirve para diferenciarlas de otras con idéntico nombre válidas en otros contextos.
- XML Schemas. Permiten definir restricciones que se aplicarán a un documento XML. Actualmente los más usados son las DTD y XSD.

[1] El **Consorcio World Wide Web** (W3C) es una comunidad internacional donde las organizaciones Miembro, personal a tiempo completo y el público en general trabajan conjuntamente para desarrollar estándares Web. Liderado por el inventor de la Web Tim Berners-Lee y el Director Ejecutivo (CEO) Jeffrey Jaffe, la misión del W3C es guiar la Web hacia su máximo potencial.

1.4.5 Comparación de XML con HTML

Si comparamos XML con HTML podemos destacar las siguientes cuestiones:

- XML:
  - Especifica cómo deben definirse conjuntos de etiquetas aplicables a un tipo de documento.
  - Modelo de hiperenlaces complejo. (XML Linking Language).
  - El navegador es una plataforma para el desarrollo de aplicaciones no para visualizarlas.
  - Fin de la guerra de los navegadores y etiquetas propietarias.
- HTML:
  - Aplica un conjunto limitado de etiquetas sobre un único tipo de documento.
  - Modelo de hiperenlaces simple.
  - El navegador es un visor de páginas.
  - El problema de la "no compatibilidad" y las diferencias entre navegadores ha alcanzado un punto en el que la solución es difícil.

1.4.6 Comparación de XML con SGML

Igualmente, XML versus SGML:

- XML:
  - Su uso es sencillo.
  - Trabaja con documentos bien formados, no exige que estén validados.
  - Facilita el desarrollo de aplicaciones de bajo coste.
  - Es muy utilizado en informática y en más áreas de aplicación.
  - Compatibilidad e integración con HTML.
  - Formateo y estilos fáciles de aplicar.
  - No usa etiquetas opcionales.
- SGML:
  - Su uso es muy complejo.
  - Sólo trabaja con documentos válidos.
  - Su complejidad hace que las aplicaciones informáticas para procesar SGML sean muy costosas.
  - Sólo se utiliza en sectores muy específicos.
  - No hay una compatibilidad con HTML definida.
  - Formateo y estilos relativamente complejos.

1.5 Lenguaje de marca. Etiquetas, elementos y atributos

Un "lenguaje de marcas" es un modo de codificar un documento donde, junto con el texto, se incorporan etiquetas, marcas o anotaciones con información adicional relativa a la estructura del texto o su formato de presentación.

Permiten hacer explícita la estructura de un documento, su contenido semántico o cualquier otra información lingüística o extralingüística que se quiera hacer patente.

Todo lenguaje de marcas está definido en un documento denominado DTD (Document Type Definition).

En él se establecen las marcas, los elementos utilizados por dicho lenguaje y sus correspondientes etiquetas y atributos, su sintaxis y normas de uso.

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

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. Ejemplo: <title>Animales Amorosos</title>

Algunos elementos no tienen contenido. Se les denomina elementos vacíos y no deben llevar etiqueta de fin. Ejemplo:
<footer/> o <footer></footer>

A su vez los elementos pueden estar formados por otros elementos (elementos anidados) y/o por atributos.

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

Ejemplo: <u>Esto está subrayado</u>

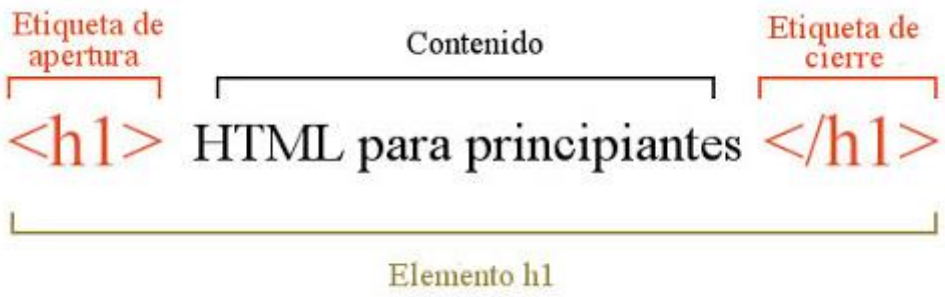
Las últimas especificaciones emitidas por el W3C indican la necesidad de que vayan escritas siempre en minúsculas para considerar que el documento está correctamente creado.

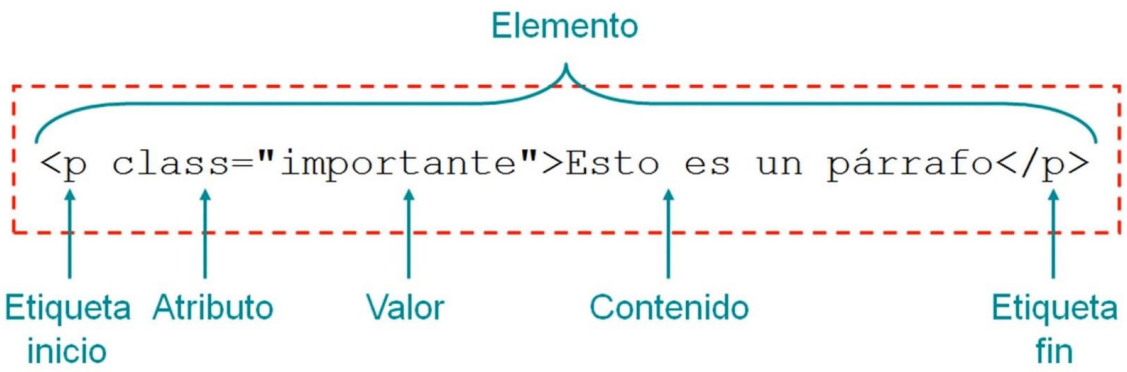
Por último, los atributos permiten añadir propiedades a los elementos de un documento.

No pueden organizarse en ninguna jerarquía, no pueden contener ningún otro elemento o atributo y no reflejan ninguna estructura lógica.

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.

Veamos algunos ejemplos de HTML:





Como se observa en el ejemplo último, los atributos se definen y dan valor dentro de una etiqueta de inicio o de elemento vacío, a continuación del nombre del elemento o de la definición de otro atributo siempre separado de ellos por un espacio.

Los valores del atributo van precedidos de un igual que sigue al nombre de este y tienen que definirse entre comillas simples o dobles.

Los nombres de los atributos han de cumplir las mismas reglas que los de los elementos, y no pueden contener el carácter menor que, <.

Y ahora, otro ejemplo, pero de XML:

Ejemplo:

```
<direccion>
        <nombre>
                <titulo>Sra.</titulo>
                <nombre>María</nombre>
                <apellidos>Merino</apellidos>
        </nombre>
        <calle>C/ Álvarez Quintero, 23</calle>
        <ciudad provincia="Cádiz">Conil</ciudad>
        <codigo-postal>34829</codigo-postal>
</direccion>
```

El elemento <nombre> contiene tres elementos hijos: <titulo>, <nombre> y <apellidos> y provincia es un atributo del elemento <ciudad> y tiene el valor "Cádiz".

1.6 Organizaciones Desarrolladoras

Hablemos un poco de las organizaciones que se han encargado de desarrollar los lenguajes de marcas, entre otras tareas.

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.

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.

Esta organización después del éxito que tuvo GML y, después de un largo proceso, publicó en 1986 el Standard Generalized Markup Languaje (SGML) 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 Keiō (Shonan Fujisawa Campus) en Japón en septiembre de 1996 como huésped asiático.

Su función principal es tutelar el crecimiento y organización de la web.

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.

1.7 Gramáticas

Todo documento de un lenguaje de marcas tiene en común una gramática que define el marcado permitido, el marcado requerido y cómo debe ser utilizado dicho marcado en la instancia del documento. Podemos distinguir dos tipos de gramáticas:

DTD (Definición de Tipo de Documento)

Establece las reglas de formación del lenguaje formal, es decir, qué combinaciones de símbolos elementales son sintácticamente correctas.

Contiene las reglas de dichos elementos: el nombre, su significado, dónde pueden ser utilizados y qué pueden contener. Define todos los elementos y define las relaciones entre los distintos elementos.

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.

Utiliza una sintaxis no-XML para definir la estructura.

Para lenguajes como XML, la DTD la debemos definir nosotros, ya que somos nosotros lo que decidimos qué conjunto de etiquetas vamos a usar, y cómo van a usarse esas etiquetas, además definimos qué etiquetas se pueden anidar, y cuáles llevan atributos y cuáles no.

Para lenguajes como HTML, la DTD ya están definidas por la W3C.

Esquema Xml

Es la evolución de la DTD también descrita por el W3C, también denominado XSD (XML Schema Definition).

Es un lenguaje de esquema más complejo y 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, supone un gran consumo en recursos y tiempo debido a su gran especificación y complejidad en la sintaxis (son más difíciles de leer y de escribir).

12/07/2022