Unidad 1. Introducción a los lenguajes de marcas

Objetivos:
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 existentes

¿Qué son los lenguajes de marcas?

Formas de representar la información

El ordenador solo es capaz de representar la información en forma de 0's y 1's, es decir, en **formato digital.** Este formato **está muy alejado de la forma de representar la información del ser humano**: texto, números, imágenes, videos, música…etc. El ser humano diferencia claramente entre un texto o una imagen o una canción. Para el ordenador todos son 0's y 1's , es decir, todo es binario.

Es por ello, que se hace necesario la **codificación de la información** original a un formato comprensible por el ordenador. La codificación de información binaria a la representación que usa el ser humano y viceversa, es complicada, este hecho genera la necesidad del uso de **estándares de codificación**.

Tipos de datos

Datos binarios

Cualquier dato que no sea texto: imágenes, videos, sonidos, archivo Excel, programa ejecutable…

Forma de codificación

Depende del tipo de dato.

Ej. Imagen: cada pixel se representa mediante una secuencia de 0's y 1's

Características

Ventajas

- Ocupan menos espacio que los archivos de texto, ya que optimizan mejor su codificación a binario
- Son más rápidos de manipular por parte del ordenador (se parecen más al lenguaje nativo del ordenador
- Permiten el acceso directo a los datos. Los archivos de texto siempre se manejan de forma secuencial, más lenta
- En cierto modo permiten cifrar el contenido que de otra forma sería totalmente visible por cualquier aplicación capaz de entender textos (como el bloc de notas o cualquier editor de texto básico), es decir los datos no son fácilmente comprensibles.

Desventajas

- Su manipulación es compleja, necesitan de un software específico para poder ser modificados.
- Son menos transportables y comprensibles por los sistemas informáticos.

Datos de texto

Cualquier dato que represente texto. Es la forma más habitual de transmitir información.

Forma de codificación

Cada carácter se representa mediante una **secuencia de 0's y 1's .**

Se utilizan **diferentes sistemas de codificación,** desgraciadamente, hay una falta de estandarización generalizada por la existencia de numerosos estándares.

Características

Ventajas

- Son ideales para almacenar datos para exportar e importar información a o desde cualquier dispositivo electrónico ya que cualquiera es capaz de interpretar texto

- Son directamente modificables, sin tener que acudir a un software específico
- Su manipulación es más sencilla que la de los archivos binarios
- Son directamente transportables y entendibles por todo tipo de redes

Desventajas

- Sólo son capaces de almacenar texto plano sin indicar ningún formato o añadir información no textual

No almacenan por ejemplo imágenes o elementos multimedia, por ese motivo se intenta que el propio texto sirva para almacenar otros datos. Para ello dentro del archivo habrá contenido que no se interpretará como texto sino "como mostrar el texto", por ejemplo.

Codificación

En ordenadores diferentes se pueden usar **sistemas de codificación distintos,** lo cual hace imposible el intercambio y compartición de documentos entre ellos.

Por ejemplo:

"A": 10111011

"A": 11000010

Aparecen entonces los **estándares de codificación de caracteres** para intentar que todos los ordenadores codifiquen los caracteres de igual forma.

ASCII

ASCII (acrónimo inglés de American Standard Code for Information Interchange, Código Estándar Estadounidense para el Intercambio de Información).

Es un código de caracteres basado en el **alfabeto latino**, tal como se usa en inglés moderno.

Fue creado en **1963** por el Instituto Estadounidense de Estándares Nacionales o **ANSI.**

El código ASCII inicialmente **utilizaba 7 bits** para representar los caracteres.

A menudo se llama incorrectamente ASCII a otros códigos de caracteres de 8 bits, como el estándar ISO-8859-1, que es una extensión que utiliza 8 bits para proporcionar caracteres adicionales usados en idiomas distintos al inglés, como el español.

Incluye las letras del alfabeto inglés, en minúsculas y mayúsculas, caracteres de puntuación, símbolos especiales, símbolos de control…

En la actualidad se define además códigos para 32 caracteres no imprimibles, de los cuales la mayoría son caracteres de control que tienen efecto sobre cómo se procesa un texto.

Casi todos los sistemas informáticos actuales utilizan el código ASCII o una extensión compatible para representar textos y para el control de dispositivos que manejan texto con el teclado.

Este estándar sufría de **ciertas carencias**: En países de habla no inglesa no tenían suficientes bits para representar los caracteres que necesitaban.

Aparecen por ese motivo los ASCII extendidos

Cada carácter se representaba mediante **8 bits.** Eso proporcionaba **128 caracteres disponibles más**.

Los ASCII extendidos incluían:

- 128 primeros caracteres originales
- 128 siguientes: caracteres extra (dependía del estándar usado)

Aparecen las normas ISO

Tablas formadas por 256 caracteres que correspondían a los estándares ASCII extendidos.

Desafortunadamente seguían siendo insuficientes para codificar todos los alfabetos del planeta

Hay que indicar el sistema de codificación utilizado, para saber cómo interpretar los códigos del archivo

Así en 8859_1 el código 245 es el carácter õ y en 8859_2 es el carácter ő

UNICODE

Unicode es un **estándar de codificación de caracteres diseñado para facilitar el tratamiento informático, transmisión y visualización de textos de múltiples lenguajes y disciplinas técnicas, además de textos clásicos de lenguas muertas**.

El término Unicode procede de los tres objetivos perseguidos: universalidad, uniformidad y unicidad.

Unicode trata los caracteres alfabéticos, ideográficos y símbolos de forma equivalente, lo que significa que se pueden mezclar en un mismo texto sin la introducción de marcas o caracteres de control.

Este estándar es mantenido por el **Unicode Technical Committee (UTC),** integrado en el **Consorcio Unicode,** del que forman parte con distinto grado de implicación empresas influyentes de informática, como: Microsoft, Apple, Adobe, IBM, Oracle, SAP, Google o Yahoo, instituciones como la Universidad de Berkele …..

El establecimiento de Unicode ha sido un **ambicioso proyecto** para reemplazar los esquemas de codificación de caracteres existentes, **muchos de los cuales están muy limitados en tamaño y son incompatibles** con entornos plurilingües.

Unicode se ha vuelto el más extenso y completo esquema de codificación de caracteres, siendo el dominante en la internacionalización y adaptación local del software informático. El estándar ha sido implementado en un número considerable de tecnologías recientes, que incluyen XML, Java y sistemas operativos modernos.

Los 128 primeros son los originales de ASCII para **mantener la compatibilidad con ISO-8859_1**

Existen varias normas disponibles:

- **UTF_8: usa 8 bits**
- **UTF_16: usa 16 bits**
- **UTF_32: usa 32 bits**

Ha conseguido **incluir los caracteres de todas las lenguas de planeta**… más de 50.000 símbolos

Definición de lenguaje de marca

Los lenguajes de marcas, también llamados lenguajes de marcado o de etiquetas o tags , 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**.

Se consigue gracias a **etiquetas o marcas o tags** intercaladas en el contenido.

El lenguaje de marcas 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 SOLO 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:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<agenda>
  <alumno>
    <nombre>jose</nombre>
    <apellido>alonso</apellido>
    <nia>123456789</nia>
  </alumno>
  <alumno>
    <nombre>pepe</nombre>
    <apellido>fernandez</apellido>
    <nia>2525258</nia>
  </alumno>
</agenda>
```

Los lenguajes de marcas han de diferenciarse de los lenguajes de programación. El lenguaje de marcas no tiene funciones aritméticas, variables o estructuras de control.

Evolución de los lenguajes de marcas

Origen

Históricamente, el marcado **se usaba y se usa en la industria editorial y de la comunicación**, así como entre autores, editoriales e imprentas. Procede del término inglés "**marking up"** y significa **"marcar manuscritos con lápiz de color para hacer anotaciones".**

Ejemplo: hacer una anotación sobre el tipo de letra a emplear en la edición de un libro en una imprenta

Los lenguajes de marcas se llaman así por la práctica tradicional de marcar **los manuscritos con instrucciones de impresión en los márgenes**. En la época de la imprenta, esta tarea ha correspondido a los **marcadores**, que indicaban el tipo de letra, el estilo y el tamaño, así como la corrección de errores, para que otras personas compusieran la tipografía. Esto condujo a la creación de un grupo de **marcas estandarizadas**. Con la introducción de las computadoras, se trasladó un concepto similar al mundo de la informática.

Los lenguajes de marcas no son lenguajes en el sentido de los lenguajes de programación de aplicaciones. 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 sí poseen los lenguajes de programación.

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.

El concepto de lenguaje de marcas fue expuesto por vez primera por **William W. Tunnicliffe** en 1967. Sin embargo, quien es considerado el padre de los lenguajes de marcas es **Charles Goldfarb**, investigador para la compañía IBM. Goldfarb participó en la creación del **lenguaje GML,** y posteriormente dirigió el comité que elaboró el estándar **SGML**, la piedra angular de los lenguajes de marcas.

Es en esas fechas es cuando se estandariza el lenguaje SGML (Standard Generalizad 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. Del que acabó derivando HTML entre otros lenguajes..

En cualquier caso, y a pesar de las controversias sobre su origen, es comúnmente aceptado que la idea de los lenguajes de marcas **surgió de forma independiente varias veces durante los 70, y que se generalizó en los años 80.**

A finales de los 80 dentro del CERN se creó un lenguaje de marcado pensado para **compartir información usando las redes de ordenadores,** y de forma más general, a través de Internet. Este lenguaje se basaba en algunos principios del SGML y lo denominaron **HTML (HiperText 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 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 méramente comerciales.

A mediados de los 90 el **consorcio W3C (Word Wide Web Consortium)** comenzó una iniciativa para intentar dotar a la Web de un lenguaje más potente y que pudiera dar una estructura 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.

Algunos de los principales lenguajes de marcas que surgieron entonces fueron:

Goldfarb

Es considerado por muchos autores como el **primer lenguaje de marcas** propiamente dicho. Fue creado por **Charles Goldfarb** , un investigador de IBM que es considerado el padre de los lenguajes de marcas.

Los documentos contenían información que inidicaban el formato con el que debían aparecer. Sobre este lenguaje se basó el lenguaje **GML** de IBM, base del futuro **SGML** ideado por el propio Charles Goldfarb.

Tex

Es un lenguaje de marcas creado por **Donald E. Knuth en la década de los 70.**

Fue muy **popular en el entorno académico**, especialmente entre las comunidades de matemáticos, físicos e informáticos. Obtuvo **mucho éxito en la comunidad científica** gracias a sus 300 comandos y tipos de fuentes de gran calidad

Ha conseguido sustituir con creces al lenguaje "troff", otro lenguaje de tipografía habitual en Unix.

TeX se considera generalmente una de las **mejores formas de componer fórmulas matemáticas** . Permite elaborar documentos científicos de gran calidad en los resultados.

Tiene como características que usa una tipografía especial: fuentes Modern Computer

Pero tiene como inconveniente que necesita un **programa específico** capaz de convertir el archivo TeX a un formato de impresión.

Latex

Es un lenguaje **derivado de Tex** creado en **1984 por Leslie Lamport**, que permite la composición de textos con una alta calidad tipográfica.

LaTeX está formado por un **gran conjunto de macros de TeX** con la intención de **simplificar y facilitar el uso** del lenguaje de composición tipográfica, creado por Donald Knuth, Tex.

Por sus características y posibilidades, es usado de forma especialmente intensa en la **generación de artículos y libros científicos** que incluyen, entre otros elementos, expresiones matemáticas.

Es muy utilizado para la composición de artículos académicos, tesis y libros técnicos, dado que la calidad tipográfica de los documentos realizados con LaTeX es comparable a la de una editorial científica de primera línea.

RTF

Rich Text Format (formato de texto enriquecido, a menudo abreviado como RTF) es un lenguaje de marcas desarrollado por Microsoft en 1987 para el **intercambio de documentos.**

Richard Brodie, Charles Simonyi, y David Luebbert, miembros del equipo de desarrollo de Microsoft Word, crearon el RTF original en los años 1980. Su sintaxis se ve influenciada por el lenguaje de composición tipográfica TeX.

La mayoría de los procesadores de texto pueden leer y escribir documentos RTF.

Básicamente consisten e**n documentos de texto con anotaciones de formato para su presentación.**

Se utiliza en **entornos Windows** como formato de intercambio entre distintos procesadores de texto por su potencia.

El procesador de texto Word Pad lo utiliza

SGML

El lenguaje de marcado generalizado estándar ó SGML **(por sus siglas en ingles de Standard Generalized Markup Language) es un lenguaje para el tratamiento de la información.** Además de un estándar para **definir otros lenguajes** de marcado para documentos. Fue definido por **ISO** y se considera un **estándar mundial** .

Básicamente , permite el tratamiento de la información y sirve de base para **definir lenguajes de etiquetas o de marcado**, por ejemplo HTML.

Sucesor de **GML ,** el cual fue creado por IBM , padre del lenguaje **XML** y base del lenguaje **HTML.**

HTML en teoría era un ejemplo de **un lenguaje basado en SGML** .

Otros ejemplos son DocBook SGML y LinuxDoc

PostScript

1976 John Warnock fue uno de los desarrolladores de este lenguaje en la empresa **Xerox.**

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

Este lenguaje se diferenció, fundamentalmente, por utilizar un lenguaje de programación para describir una imagen de impresión.

Imagen que más tarde sería impresa en una impresora láser o algún otro dispositivo de salida de gran calidad, en lugar de una serie de secuencias de escapes de bajo nivel.

Proporciona indicaciones muy potentes para componer imágenes, que posteriormente serán impresas. Puede incluir texto y el tipo de letra del mismo, píxeles individuales y formas vectoriales (líneas, curvas). **Sus posibilidades son muy amplias.**

Se puede considerar un **lenguaje de programación tradicional** en muchos sentidos

En 1985 John Warnock acabará creando la empresa Adobe Systems.

HTML

Fue creado por **Tim Berners-Lee en 1991 en el CERN** . Está basado en el lenguaje de marcas SGML.

HTML, siglas en inglés de **HyperText Markup Language (lenguaje de marcas de hipertexto),** hace referencia al lenguaje de marcado para la **elaboración de páginas web** o documentos transportables a través de Internet.

Es un estándar que sirve de referencia para la **elaboración de páginas web** en sus diferentes versiones. Este lenguaje define una estructura básica y un código (denominado código HTML) para la definición de contenido de una página web, como texto, imágenes, videos, juegos, entre otros. **Hoy en día casi todo en Internet se ve a través de documentos HTML.**

Inicialmente estos documentos se veían con ayuda de intérpretes de texto (como por ejemplo el Lynx de Unix) que simplemente coloreaban el texto y remarcaban el hipertexto.

Aparecieron navegadores con capacidad más gráfica para mostrar formatos más avanzados y visuales: elementos media

Es un estándar a cargo del **World Wide Web Consortium (W3C) o Consorcio WWW,** organización dedicada a la estandarización de casi todas las tecnologías ligadas a laSweb, sobre todo en lo referente a su escritura e interpretación.

Se considera el **lenguaje web más importante** siendo su invención crucial en la **aparición, desarrollo y expansión de la World Wide Web (WWW) y en el éxito rotundo de Internet .** Es el estándar que se ha impuesto en la visualización de páginas web y es el que todos los navegadores actuales han adoptado.

El HTML se escribe en forma de «etiquetas», rodeadas por corchetes angulares (<,>,/). El HTML también puede describir, hasta un cierto punto, la apariencia de un documento, y puede incluir o hacer referencia a un tipo de programa llamado script, el cual puede afectar el comportamiento de navegadores web y otro programas. HTML consta de varios componentes vitales, entre ellos los elementos y sus atributos, tipos de data y la declaración de tipo de documento.

XML

XML, siglas en inglés de **eXtensible Markup Language ("lenguaje de marcas Extensible"),** es un lenguaje que permite definir lenguajes de marcas, desarrollado por el **World Wide Web Consortium (W3C)** y es **utilizado básicamente para almacenar datos en forma legible**.

Procede del lenguaje SGML y también permite definir lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML) para estructurar documentos grandes.

Intenta mejorar el propio SGML: su sintaxis es más estricta, pero más comprensible

XML es una **tecnología sencilla** que tiene a su alrededor **otras tecnologías que la complementan** y la hacen mucho más grande y con unas posibilidades mucho mayores.

Tiene un papel muy importante en la actualidad ya que **permite la compatibilidad entre sistemas para compartir la información** de una manera segura, fiable y fácil.

A diferencia de otros lenguajes, XML **da soporte a bases de datos**, siendo útil cuando varias aplicaciones deben comunicarse entre sí o integrar información.

XML no ha nacido sólo para su aplicación para Internet, sino que se propone como un estándar para el intercambio de información estructurada entre diferentes plataformas. Se puede usar en bases de datos, editores de texto, hojas de cálculo y casi cualquier cosa imaginable.

Su popularidad le ha convertido en el **lenguaje de marcado más importante** de la actualidad para el **almacenamiento e intercambio de la información.** Permite la **exportación e importación** de información entre sistemas informáticos de una forma muy exitosa.

JSON

JSON, acrónimo de **JavaScript Object Notation,** es un formato de texto ligero para el **intercambio de datos.** Fue creado en 2002 y está basado en el lenguaje JavaScript estándar . **No se considera lenguaje de marcas al 100%**

Compite claramente con XML. Una de las ventajas de JSON sobre el lenguaje XML como formato de intercambio de datos es que es mucho **más sencillo** .

JSON se emplea habitualmente en entornos donde el tamaño del flujo de datos entre cliente y servidor es de vital importancia (de aquí su uso por Yahoo, Google, etc. , que atienden a millones de usuarios) .

Clasificación de los 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**.

Por ejemplo, el HTML contiene etiquetas puramente procedimentales, como la B de bold (negrita), junto con otras puramente descriptivas. El HTML, por ejemplo, también incluye el elemento PRE, que indica que el texto debe representarse tal y como está escrito.

Lenguajes orientados a presentación

Son aquellos que **indican el formato del texto**. Se añaden palabras encerradas en símbolos especiales al texto que indican que formato se le debe dar. **Los traductores (programas) interpretan esas palabras y le dan el formato correspondiente.**

Este tipo de marcado es útil para maquetar la **presentación de un documento** para su lectura, pero resulta **insuficiente para el procesamiento automático de la información**.

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.

Este tipo de lenguajes son los usados tradicionalmente por los **procesadores de texto** como puede **ser Microsoft Word.**

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 web. Varios programas pueden deducir la estructura del texto basándose en esta clase de datos, aunque el resultado suele ser bastante imperfecto.

Generalmente **las marcas de los lenguajes orientados a presentación se ocultan al usuario** lo que permite obtener un efecto **WYSIWYG** ("What You See Is What You Get", "lo que ves es lo que obtienes")

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. Este tipo de lenguajes de marcas **no suelen ser flexibles ni reutilizables.**

Lenguajes orientados a procedimientos

En este tipo de lenguajes las etiquetas son también orientadas a presentación pero **se integran dentro de un marco procedural o funcional** que permite definir macros (es decir, secuencias de acciones) y subrutinas que van a dar el formato al texto.

Se añaden **palabras que se interpretan como órdenes**, como si fueran líneas de código o instrucciones de un lenguaje de programación tradicional .**Esas órdenes indican cómo se debe presentar la información.**

Estos lenguajes **son visibles para el usuario** que edita el texto.

El programa que muestra 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. En sistemas más avanzados se utilizan macros o pilas que facilitan el trabajo.

Algunos ejemplos de lenguajes orientados a procedimientos son **nroff, troff, TeX , LaTeX y Postscript…**

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

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

Lenguajes descriptivos

Se añaden palabras encerradas en símbolos especiales al texto (tags o etiquetas) que **indican el significado ó la semántica del texto** . No indican cómo deben presentarlo. Los **traductores interpretarán esas palabras y le darán el formato correspondiente dependiendo del traductor.**

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 se están 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.

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

Por ejemplo, el lenguaje Atom, 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 mostrar este dato de múltiples maneras.

Una de las ventajas 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, posteriormente los buscadores los emplearon para localizar nuevas páginas con información relacionada, o para evaluar la popularidad de determinado sitio web.

El marcado descriptivo también **simplifica la tarea de formatear un texto,** debido a que la información del formato, su apariencia, está separada del propio contenido.

Por ejemplo, un fragmento marcado como énfasis en HTML (<EM>texto</M>), puede mostrarse en cursiva, negrita ... depende del programa traductor o navegador que se utilice.

El marcado descriptivo está evolucionando hacia el **marcado genérico**. Los nuevos sistemas de marcado descriptivo estructuran los documentos en una estructura arborescente. Esto **permite tratarlos como bases de datos**, en las que el propio almacenamiento tiene en cuenta la estructura.

Etiquetas, elementos y atributos

Existen tres términos comúnmente usados para describir los componentes de un documento escrito con un lenguaje de marcas: etiquetas, elementos y atributos.

Etiquetas

Es un **carácter o cadena de caracteres encerrados entre símbolos especiales** que se encuentran insertados dentro del texto a mostrar al usuario.

En HTML una etiqueta, también llamada "tag", es un texto que va entre el símbolo menor que (<) y el símbolo mayor que (>).

Normalmente existe una **etiqueta de apertura y otra de etiqueta de cierre** que se diferencian por algún otro carácter especial

<nombre> **</nombre>**

Los caracteres especiales pueden variar de un lenguaje a otro, pero normalmente so :

< **>{}[]**

Las etiquetas o marcas le confieren un significado especial al texto encerrado en ellas y permite que sea tratado de forma diferente por el programa traductor a la hora de mostrarlo al usuario.

Elementos

Los elementos son los que van a **definir la estructura del texto** y van a darle un significado.

Los elementos están formados por **las etiquetas de inicio y fin y todo lo que se encuentran entre ambas.**

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 etiquetas.

Los elementos van a formar una **estructura jerarquizada** donde se van a poder distinguir:

- Elementos padre
- Elementos hijos

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

Atributos

Dentro de las etiquetas podemos encontrarnos con otros objetos que no son más que **propiedades que modifican o matizan el significado de la etiqueta**.

Un atributo es un par **"nombre=valor"** que se encuentra dentro de la etiqueta de inicio de un elemento e indica las propiedades que pueden llevar asociadas ese elemento.
El formato suele ser:

nombre del atributo = valor del atributo

Además del símbolo = podemos encontrar también otros símbolos como …. **:** **==**

Organizaciones desarrolladoras

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

Organización Internacional para la Estandarización (ISO, Internacional Organization for Standarization)

Se formó después de la Segunda Guerra Mundial y es el **organismo encargado de promover el desarrollo de normas e 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 la **estandarización de normas de productos** y seguridad para las empresas u organizaciones a **nivel internacional.**

Es una **red de Institutos de más 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 ello 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 Generalizad Markup Languajge (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 W3C. 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.

W3C está integrado por:

- **Miembros:** a Abril de 2010 contaba con 330 miembros
- **Equipo (W3C Team):** 65 investigadores y expertos de todo el mundo
- **Oficinas (W3C Offices):** centros regionales establecidos en Alemania y Austria (oficina conjunta), Australia, Benelux (oficina conjunta), China, Corea del Sur, España, Finlandia, Grecia, Hong Kong, Hungría, India, Israel, Italia, Marruecos, Suecia y Reino Unido e Irlanda (oficina conjunta)

Utilización de lenguajes de marcas en entornos Web

Los lenguajes de marcado son la herramienta fundamental en el **diseño de la WWW.**

Gracias a estos lenguajes **nació el servicio más usado en Internet: la WWW**. Este servicio de Internet, no solo permite acceder a la información (texto , imágenes, vídeos, música, etc. …) interrelacionada entre sí mediantes hiperenlaces, sino que además permite mostrar dicha información con un determinado formato.

Los hiperenlaces son cadenas de caracteres que pueden enlazar ese documento con otros documentos de similares características

Una **página Web** es un documento de texto adaptado para ser **mostrado en un navegador**. Básicamente está escrita en un lenguaje de marcas, generalmente, HTML, aunque puede tener código de otros lenguajes incrustado (scripts).

Una página Web forma parte de un **sitio WEB**, y está guardada en un servidor Web junto con otras páginas Web entre otros elementos: imágenes, videos, sonidos...

Un sitio Web está formado por cientos de páginas escritas habitualmente en HTML

La página Web está compuesta, principalmente, **por información (solo texto o módulos multimedia), así como por hiperenlaces;** además, puede contener o tener asociados **datos de estilo (CSS)** para especificar cómo debe visualizarse, y también **aplicaciones embebidas** para hacerla interactiva.

El contenido de la página puede ser predeterminado y fij : **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**, **no existe una norma fija, pero, algunos organismos, es especial el W3C, suelen establecer directivas** con la intención de normalizar el diseño y así facilitar y simplificar la visualización e interpretación del contenido de una página Web.

Gramáticas

Todo documento escrito en un lenguaje de marcas tiene en común una **gramática o una sintaxis**

La sintaxis del lenguaje o simplemente gramática es definida mediante un **DTD (Definition Type Document, Definición de Tipo de Documento) o estándar .**

La gramática describe los componentes del lenguaje: **nombre, significado, donde puede ser utilizado, que valores puede tomar, relación entre los elementos…**

DTD

La DTD (Definición de Tipo de Documento)describe la sintaxis de un lenguaje, es decir, establece las reglas de formación del lenguaje formal,: 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 componentes: el nombre, su significado, dónde pueden ser utilizados y qué pueden contener….

- **Elementos** (nombre, significado , donde pueden ser usados y que pueden contener)
- **Etiquetas (**nombre, significado, donde pueden usarse)
- **Atributos (**nombre, significado, donde pueden usarse, valores

Es posibles que existan **varios tipos de DTD's** para un mismo lenguaje de marcas…

Lenguaje HTML

Las especificaciones del W3C para el lenguaje de marcas HTML contempla tres tipos de DTD's:

DTD Estricta

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

Existen otros elementos, etiquetas y atributos mejores que los han sustituido.

Funcionan en la mayoría aplicaciones, herramientas o navegadores, pero no en todos.

DTD Transicional o flexible

Incluye todo lo que la anterior más los elementos y atributos desaprobados (deprecated)

DTD para documentos con marcos

Engloba todo lo incluido en la transicional más lo relativo a la creación de documentos con marcos (frames)

Marcos o frames

Son utilidades que ofrecen la posibilidad de **dividir la ventana del navegador en secciones y visualizar varias páginas Web a la vez en cada una de ellas**. Cada uno de los documentos HTML son independientes entre sí.

Algunos navegadores no soportan frames.

Aunque la especificación recomienda ceñirse a los recursos de la DTD estricta, utilizar el resto de los elementos y atributos no es incorrecto, aunque si poco recomendable.

Lenguaje XML

La DTD de un documento XML:

- Define todos los elementos posibles
- 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.

La DTD es el método más sencillo usado para validar un documento XML , es decir, comprobar si está correctamente escrito.

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

Nota Aclaratoria del material:

Material tomado del IES DOÑANA – Cádiz
Existen muchos enlaces y descargas en la web. No dispongo si es material de dominio público.

PRÁCTICAS DE CLASE

PRÁCTICA 1

- Investigar tipos de Metadatos
- Investiga Otros Estándares de Metadatos
  - Estándares de información geográfica en España
  - Estándares en Educación
- Estándares de Redes Sociales
  - Exportar por Whatsapp un contacto (tipo de metadato VCARD)
- Investiga cómo se almacena una representación gráfica (imagen) en un ordenador
  - Optimización de bits (repetitivo

PRÁCTICA 2

- Qué es la codificación de caracteres
- Ejemplos de Sistemas de representación o codificación
  - Código ASCII extendido
  - Unicode
- Exporta / importa datos de una hoja de cálculo (fichero CVS)
  - Generamos un fichero RTF y comprobamos al modificar etiquetas

Práctica 3

- Realiza una presentación con la historia de los lenguajes de marcas:
  - GML
  - SGML (ISO 8879)
  - HTML
  - XML
  - XHTML

- En dicha presentación resalta los aspectos más importantes de cada una de ellas
  - Ejemplo falta de estandarización en los formatos de información
  - Organismo o creador de cada lenguaje

- Tipos de Lenguaje de marcas
  - Presentación, procedimientos, descriptivos o semánticos

PRÁCTICA 4

- DESCOMPRIME UN FICHERO DOCX
  - OBSERVAS LAS DIVERSAS CARPETAS
  - ABRE EL FICHERO DE LA CARPETA WORD / DOCUMENT.XML
- Cambia la extensión a un fichero DOCX por ZIP
  - Descomprime