UT1. INTRODUCCIÓN A LOS LENGUAJES DE MARCAS

1. QUÉ SON LOS LENGUAJES DE MARCAS.

1.1 ORIGEN DE LOS LENGUAJES DE MARCAS

En los inicios de la informática los archivos digitales solo contenían texto sin formato. Los caracteres eran almacenados utilizando una codificación numérica (un carácter se representa por un número). ASCII fue el código más extendido.

El siguiente paso fue almacenar el formato en que se quiere visualizar el texto, en el mismo archivo digital (cuales son los párrafos, que debe ir en negrita o cursiva).

Para indicar las distintas partes del texto se utilizan marcas (*markup* en la jerga tipográfica inglesa).

Para realizar marcas se utilizan etiquetas (tag) con un significado preestablecido.

Ejemplos de documentos generado con un lenguaje de marcas:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html>
 <head>
  <title>Combinación de Estilos</title>
 </head>
 <body>
  <p><u><b>texto en negrita y subrayado</b></u><br>
```

```
  <s><i>italica y tachado</i></s><br>
  <b><tt>negrita teletipo</tt></b><br></p>
  X <u><sup>subrayado superindice</sup></u><br>
  X <i><sub>italica subindice</sub></i>
 </body>
</html>
```

```
<ficha>
        <nombre>Pepe </nombre>
        <apellido>Pérez</apellido>
        <apellido>Rodriguez</apellido>
        <datos_profesionales>
                <cuerpo>grupo A</cuerpo>
                <especialidad>informática</especialidad>
                <destino>Ayuntamiento de Burgos</destino>
        </datos_profesionales>
</ficha>
```

1.2. DEFINICIÓN

Los lenguajes de marcas combinan la información generalmente textual con marcas relativas a:

- la estructura del texto: secciones, párrafos, listas etc…
- la forma en que se ha de presentar el texto: tipo de letra, colores etc

El lenguaje de marcas es el que especifica cuales serán las etiquetas posibles, dónde se deben colocar, y el significado que tendrá cada una de ellas.

Las etiquetas no se suelen presentar al usuario final, que está interesado en el contenido del documento.

Los lenguajes de marcas son lenguajes descriptivos no son lenguajes de programación (no tienen variables, sentencias de control, funciones, etc).

1.3. CARACTERÍSTICAS DE LOS LENGUAJES DE MARCAS

- Las etiquetas suelen ser palabras entre los signos menor ‘<’ y mayor ‘>’
- Lo habitual es usar pares de marcas: una de principio y otra de fin.
- El documento generado con un lenguaje de marcas se almacena en ficheros de texto plano, es decir que contiene solo texto sin formato (como vi de unix o notepad de windows).

  Esto aporta las siguientes ventajas:

- Portabilidad ya que puede ser implementado en cualquier arquitectura hardware o software que posea un simple editor de textos.
- Reutilización del código, ya que es abierto puede ser analizado por cualquiera.
- Facilidad de mantenimiento, ya que es entendible por los humanos y puede ser tratado por cualquier editor de textos.

1.4. TIPOS DE LENGUAJES DE MARCAS

- **Lenguajes de Presentación**: aportan información sobre el formato del texto para su maquetado.

  Ejemplo: RTF (Rich Text Format), o Word de Microsoft, procesadores de texto en general.

- **Lenguajes de Procedimiento**: Esta orientado a la presentación pero indicando los procedimientos que deberá realizar el software de presentación.

  Postcript (impresoras, talleres de impresión profesional), TeX, LaTeX (utilizado para fórmulas matemáticas).

- **Lenguajes Descriptivos o Semánticos:** Describen lo que significa el contenido y las partes de que las que se compone el documento, pero no indican que se ha de hacer para representarlo. (XML).

2. PRECEDENTES

2.1. BREVE HISTORIA DE INTERNET

El origen de Internet se remonta a 1969, cuando la Agencia de Proyectos para la Investigación Avanzada de Estados Unidos, ARPA, **conectó cuatro sistemas distantes** en una red que se denominó ARPANET, cuya misión era mantener las comunicaciones en caso de guerra.

Esta agencia, dependiente del Departamento de Defensa, nació en 1958 con el objetivo de desarrollar proyectos de tecnología militar en plena Guerra Fría. EE.UU. quería contrarrestar los avances de la antigua URSS.

Hasta ese momento, Estados Unidos contaba con una red centralizada que se consideraba muy insegura en caso de guerra, ya que un solo fallo podría bloquear el sistema.

Con ARPANET llegaba una revolución en el campo de las comunicaciones porque era una red que impedía que se perdiesen las comunicaciones aún en el caso de fallo de alguno de los nodos de la red.

De las agencias militares a las universidades

El salto cualitativo se produjo cuando ARPANET se extendió por el mundo académico. Los científicos la utilizaron y la desarrollaron para compartir opiniones y colaborar en sus trabajos.

La red conectó todas las agencias y los proyectos de defensa de Estados Unidos, y en 1972 ya integraba a 50 universidades y centros de investigación diseminados por todo el país.

El número de ordenadores conectados creció, y a partir de los 80 aparecieron otras redes, lo que provocó el caos por la variedad de protocolos y estándares de redes a los que los ordenadores estaban de ordenadores conectados.

Con la unificación de esas redes nace internet que utiliza la familia de protocolos de Internet *TCP/IP como protocolo de interconexión entre nodos.*

Alrededor de 1993, las empresas e individuos que deseaban la capacidad de las comunicaciones globales digitales que proporcionaba esta red, presionaron para obtener una red global interconectada con un tráfico sin restricciones. La mayor parte de la información en ese momento era simple texto ASCII sobre temas académicos. No existía un orden concreto, ni una organización temática, y era muy laborioso encontrar un documento de interés.

La red Internet que conocemos a día de hoy cuenta con millones de equipos conectados, donde los usuarios no tienen porque ser especialistas en tecnología, tal y como sucedía en los inicios de esta breve pero intensa historia.

El TCP/IP es un conjunto de protocolos. Los dos protocolos más importantes son Protocolo de Control de Transmisión (TCP) y Protocolo de Internet (IP) que son la base de Internet y sirve para enlazar redes y computadoras que utilizan diferentes sistemas operativos. El conjunto lo forman más de 100 protocolos, los mas populares son: HTTP (HyperText Transfer Protocol) que se utiliza para acceder a las páginas web, ARP (Address Resolution Protocol) para la resolución de direcciones, FTP (File Transfer Protocol) para transferencia de archivos, y SMTP (Simple Mail Transfer Protocol) y POP (Post Office Protocol) para correo electrónico, TELNET para acceder a equipos remotos, etc…

2.2. HISTORIA DE LOS LENGUAJES DE MARCAS

- En 1969 un equipo de tres empleados de IBM, Charles Goldfarb, Edward Mosher y Raymond Lorie, trabajaron conjuntamente en el desarrollo de GML (Generalized Markup Language), un lenguaje cuyo primer objetivo era facilitar la integración de los sistemas de información en un proyecto desarrollado para despachos de abogados.

- En 1978 ANSI creó un comité de Lenguajes de Ordenador para Procesador de Texto (Computer Languages for the Processing of Text).

- En 1980 se genera la primera versión de SGML (Standard Generalized Markup Language), a partir de GML. Es un lenguaje para permitir compartir información entre diferentes sistemas informáticos. Su principal inconveniente es su complejidad.

- En 1986 se convierte en la norma ISO 8879

- En 1992 el físico Tim Berners-Lee trabajador de la CERN (Conseil Européen pour la Recherche Nucléaire) padre de la web junto a Robert Cailliau, definen los pilares de la WWW:

  - El protocolo HTTP (HyperText Transfer Protocol)
  - El sistema de localización de objetos en la web URL(Uniform Resource Locator)
  - El lenguaje de Marcas HTML (HyperText Markup Language)

- En 1995 el IETF (http://www.ietf.org/) (*Internet Engineering Task Force*) publica HTML 2.0 que es el primer estándar oficial de HTML. (Permite formularios).

- En 1994 Tim Berners-Lee fue contratado por el Laboratorio de Ciencias de la Computación e Inteligencia Artificial del Massachusetts Institute of Technology (MIT) y fundó el W3C (World Wide Web Consortium), que dirige actualmente.

- En 1997 W3C publica HTML 3.2. Incorpora applets de java y texto que fluye alrededor de las imágenes.

- En 1998 W3C publica HTML 4.0. Las novedades más destacadas que incorpora son las hojas de estilos CSS, la posibilidad de incluir pequeños programas o scripts en las páginas web, mejora de la

accesibilidad de las páginas diseñadas, tablas complejas y mejoras en los formularios (y la desafortuna inclusión de los marcos).

- En 1999 W3C publica HTML 4.1 Se trata de una revisión y actualización de la versión HTML 4.0, por lo que no incluye novedades significativas.

- En esta década de los 90 W3C comenzó una iniciativa para dotar a la web de un lenguaje más potente que incorporara estructura semántica. Crearon un lenguaje de marcas basado en la potencia de SGML y sencillo como HTML. En 1998 se publicó el nuevo estándar que se denominó XML (eXtended Markup Language). XML es un metalenguaje, es decir un conjunto de reglas para crear lenguajes de marcas.

- Desde la publicación de HTML 4.01, la actividad de estandarización de HTML se detuvo y el W3C se centró en el desarrollo del estándar XHTML. Por este motivo, en el año 2004, las empresas Apple, Mozilla y Opera mostraron su preocupación por la falta de interés del W3C en HTML y decidieron organizarse en una nueva asociación llamada WHATWG (http://www.whatwg.org/) (*Web Hypertext Application Technology Working Group*).

- La actividad del WHATWG se centró en el futuro estándar HTML 5, cuyo primer borrador oficial (http://www.w3.org/TR/html5/) se publicó el 22 de enero de 2008.

- Debido a la fuerza de las empresas que forman el grupo WHATWG y a la publicación de los borradores de HTML 5.0, en marzo de 2007 el W3C decidió retomar la actividad estandarizadora de HTML (http://www.w3.org/2007/03/html-pressrelease) .

- De forma paralela a su actividad con HTML 5.0, W3C ha continuado con la estandarización de XHTML, una versión *avanzada* de HTML y basada en XML. La primera versión de XHTML se denomina XHTML 1.0 y se publicó en Enero de 2000. Posteriormente XHTML 1.1 se revisó en Agosto de 2002.

- En Octubre de 2014 se publica como estándar HTML5

2.3. ORGANIZACIONES DESARROLLADORAS

El **W3C (World Wide Web Consortium)** es un organismo internacional de estandarización de tecnologías Web dirigido conjuntamente por el **MIT** (Massachusetts Institute of Technology) donde está su sede central, el **ERCIM** francés (European Research Consortium for Informatics and Mathematics) y la **Universidad de Keiō** en Japón. Este organismo decidió que todos sus estándares fuesen libres, es decir, que los pudiese utilizar todo el mundo libremente sin coste alguno, lo que sin lugar a dudas fue una de las grandes razones para que la Web haya llegado a tener la importancia que tiene hoy en día.

ISO (International Standards Organization) Su función principal es la de buscar la estandarización de normas de productos y seguridad para las empresas u organizaciones a nivel internacional.

ANSI (American National Standards Institute) es una organización sin ánimo de lucro que supervisa el desarrollo de estándares para productos, servicios, procesos y sistemas en los Estados Unidos.

3. WWW (WORLDWIDE WEB)

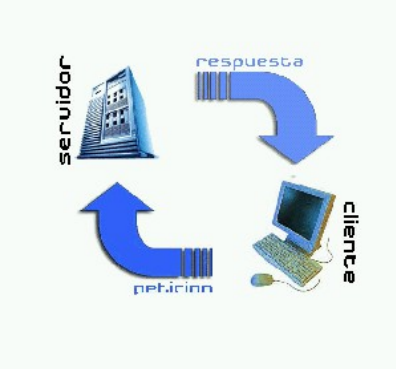
3.1. Definición de WWW

La World Wide Web, la Web, o WWW (en castellano Tela de Araña Mundial) es un sistema de documentos de hipertexto enlazados y accesibles a través de la red Internet.

El usuario utiliza un Navegador (browser) para acceder a los elementos de información que se llaman Páginas Web (web pages) y mostrarlos en pantalla. Estas páginas web están almacenadas en ordenadores llamados Servidores o también Sitios Web (websites).

Las páginas web contienen texto, imágenes, videos y otros contenidos multimedia y se navega a través de ellas usando hiperenlaces (hipertextos y hipermedios).

Los Navegadores Clientes Web (IExplorer, Chrome, Mozilla, Opera, Safari, etc.) son aplicaciones que solicitan las páginas a los Ordenadores Servidores donde estas se alojan (website), las reciben y las interpretan.



El usuario puede acceder a nuevas páginas web haciendo click en los enlaces localizados en la página que está visualizando en su navegador. Asimismo, el usuario puede enviar información al servidor web a través de estas páginas con el objeto de interactuar con el servidor.

Para que la navegación sea posible son necesarias tres cosas: Que cada documento tenga un localizador o identificador (url). Que haya un lenguaje en que las computadoras se comuniquen para pedirse y entregarse los documentos unas a otras (http). Que haya una forma de codificar los documentos para que una vez obtenido por el ordenador se represente en la pantalla o en otro medio (html).

La funcionalidad de la Web se basa en tres estándares:

- **El Identificador Uniforme de Recursos (URI)**: Especifica cómo a cada página de información se le asocia un "nombre" único.
- **El Protocolo de Transferencia de Hipertexto (HTTP)**: Especifica cómo el navegador y el servidor web intercambian información en forma de peticiones y respuestas.
- **El Lenguaje de Marcación de Hipertexto (HTML)**: Define un método para codificar la información de los documentos y sus enlaces en forma de hipertexto.

3.2. HTTP

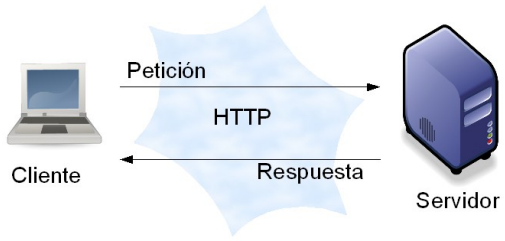
HTTP es el protocolo usado en las transferencias de páginas en la Web.

Es la abreviatura de Hypertext Transfer Protocol (Protocolo de Transferencia de Hipertexto).

HTTP es el sistema mediante el cual se envían las peticiones para acceder a una página web, y se reciben las respuestas del servidor web (las páginas web). HTTP también permite enviar datos al servidor para interactuar con él, como por ejemplo formularios de datos.

HTTP dispone de una variante segura llamada HTTPS, que permite cifrar el contenido de lo que se envía y recibe mediante el protocolo SSL.

HTTP está basado en el principio cliente/servidor. HTTP permite que la "computadora A" (el cliente) establezca una conexión con la "computadora B" (el servidor) y hacer una petición. El servidor acepta la conexión iniciada por el cliente y le envía una respuesta.



(Cuando un usuario selecciona un enlace hipertexto, el programa cliente (navegador) usa HTTP para contactar al servidor, identificando el recurso solicitado.

Una petición HTTP identifica el recurso que le interesa al cliente y le dice al servidor qué "acción" realizar en el recurso.

El servidor acepta el pedido, y entonces usa HTTP para responder o realizar la acción requerida.)

HTTP es un protocolo sin estado, es decir, que no guarda ninguna información sobre las peticiones de páginas web realizadas anteriormente. Al finalizar la transacción todos los datos se pierden.

Debido a esta limitación aparecieron las cookies, que son pequeños ficheros guardados en el ordenador cliente, y que se pueden leer desde un servidor web al establecer conexión con él. De esta forma se puede reconocer a un cliente que anteriormente estuvo accediendo a información del servidor. Gracias a esta identificación, el servidor web puede almacenar información sobre el cliente con el objeto de ofrecerle un servicio de navegación a la medida del cliente.

3.3. URL

URI (Uniform Resource Identifier) Identificador Uniforme de Recurso.

Todos los recursos disponibles en la Web (documentos HTML, imágenes, videoclips, programas, grupos de noticias, blogs, una dirección de correo electrónico, etc.) necesita un identificador único que lo diferencie de todos los demás y se codifica mediante un URI.

Un URI es una cadena de texto que nombra de forma unívoca cualquier recurso accesible en una red, pero también puede ser algo que no esté en la red (un libro, una organización, una persona, etc…). La especificación detallada se encuentra en el documento llamado "RFC-2396 - Uniform Resource Identifiers (URI): Generic Syntax".

Existen dos maneras distintas de identificar un recurso, según la finalidad que se persiga: podemos identificar un recurso por su nombre, o podemos identificarlo por su localización. Por ello existen dos tipos de URI:

- Los URN (Uniform Resource Name).
- Los URL (Uniform Resource Locator).

Un ejemplo de URN es ISBN:0-1234-98765-1 Este URI identifica un libro en base a su código ISBN, pero NO da indicación alguna de como obtener una copia de él.

Los URLs tienen una sintaxis que depende del tipo de recurso. Consta de dos partes separadas por el carácter ":"

- El tipo de esquema que utiliza la URL ("http", "ftp", "mailto", etc)
- La parte que identifica el recurso dentro del esquema. Esta parte sigue unas reglas generales de formación, pero depende del tipo de esquema que se esté utilizando.

mailto:secretaria@iesellago.net

http://www.iesellago.net

El formato de una URL genérica es:

protocolo://máquina/directorio/fichero

3.4. SGML, XML, HTML, XHTML: ¿UN LIO?

SGML:

SGML, Standard Generalized Markup Language, o "Lenguaje de Marcado Generalizado". Consiste en un sistema para la organización y etiquetado de documentos.

La Organización Internacional de Estándares (ISO) normalizó este lenguaje en 1986.

El SGML sirve para especificar las reglas de etiquetado de documentos y no impone en sí ningún conjunto de etiquetas en especial. Es un metalenguaje.

HTML está definido a partir de las normas del SGML.

XML es un estándar de creación posterior, que incorpora un subconjunto de la funcionalidad del SGML (suficiente para las necesidades comunes).

XML:

XML, eXtensible Markup Language, (lenguaje de marcas extensible), es un metalenguaje extensible de etiquetas desarrollado por el World Wide Web Consortium (W3C).

Se trata de una adaptación del SGML y permite definir la gramática de lenguajes específicos adaptados a una necesidad concreta.

XML no es realmente un lenguaje en particular, sino una manera de definir lenguajes para diferentes necesidades.

Algunos de estos lenguajes que usan XML para su definición son XHTML, SVG, MathML.

XML no ha nacido sólo para su aplicación en 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, imágenes y casi cualquier cosa imaginable. Complementado con otra serie de tecnologías, constituye el futuro del tratamiento de la información.

HTML:

El W3C define el lenguaje HTML como *"un lenguaje reconocido universalmente y que permite publicar información de forma global".*

Desde su creación, el lenguaje HTML ha pasado de ser un lenguaje utilizado exclusivamente para crear documentos electrónicos (Viene de un ambiente científico) a ser un lenguaje que se utiliza en Internet para muchas aplicaciones electrónicas como buscadores, tiendas online y banca electrónica.

El lenguaje HTML es un estándar reconocido en todo el mundo y cuyas normas define W3C.

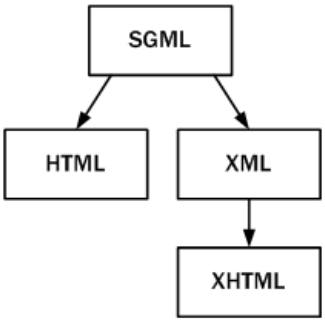
Como se trata de un estándar reconocido por todas las empresas relacionadas con el mundo de Internet, una misma página HTML se visualiza de forma muy similar en cualquier navegador de cualquier sistema operativo (En teoria).

El lenguaje HTML está definido a partir de las normas del SGML.

XHTML:

El lenguaje XHTML es muy similar al lenguaje HTML. De hecho, XHTML no es más que una adaptación de HTML al lenguaje XML.

Técnicamente, HTML es descendiente directo del lenguaje SGML, mientras que XHTML lo es del XML (que a su vez, también es descendiente de SGML).



Las páginas y documentos creados con XHTML son muy similares a las páginas y documentos HTML.