LENGUAJES DE MARCAS Y SISTEMAS DE GESTIÓN DE INFORMACIÓN.

UNIDAD 1. RECONOCIMENTO DE LAS CARACTERÍSTICAS DE LENGUAJES DE MARCAS.

1.1 Concepto y características generales, ventajas para el tratamiento de la información.

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 del formato del documento tales como especificación de la fuente, línea espaciamiento, y disposición de página requerida por Script. Usando GML, un documento está marcado con las etiquetas que definen cuáles son el texto, en términos de párrafos, listas, tablas, y así sucesivamente. El documento se puede entonces ajustar al formato automáticamente para varios dispositivos simplemente especificando un perfil para el dispositivo. Por ejemplo, es posible ajustar a formato un documento para una impresora laser o para una pantalla simplemente especificando un perfil para el dispositivo sin cambiar el documento.

El principal problema, antes de usar GML, era que cada aplicación utilizaba sus propias marcas para describir los diferentes elementos. Las marcas son códigos que indican a un programa cómo debe tratar su contenido y así, si se desea que un texto aparezca con un formato determinado, dicho texto debe ir delimitado por la correspondiente marca que indique como debe ser mostrado en pantalla o impreso. Y lo mismo ocurre con todas las demás características de cualquier texto.

Conociendo este sistema y conociendo a la perfección el sistema de marcas de cada aplicación sería posible pasar información de un sistema a otro sin necesidad de perder el formato indicado.

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.

Los lenguajes de marcas no son equivalentes a los lenguajes de programación aunque se definan igualmente como "lenguajes". Son sistemas complejos de descripción de información, normalmente documentos, que si se ajustan a SGML, se pueden controlar desde cualquier editor ASCII. Las marcas

más utilizadas suelen describirse por textos descriptivos encerrados entre signos de "menor" (<) y "mayor" (>), siendo lo más usual que existan una marca de principio y otra de final.

Metalenguaje: Existen varias definiciones para especificar este término, por ejemplo:

- Es el lenguaje que se utiliza para hacer referencias a otros lenguajes.
- Lenguaje utilizado para describir un sistema de lenguaje de programación.
- Lenguaje que se usa para hablar del lenguaje.

SGML (Standard Generalized Markup Language, 1986): es un metalenguaje que permite definir lenguajes de marcado.

- Especifica la sintaxis para la inclusión de marcas en los textos, así como la sintaxis del documento que especifica qué etiquetas están permitidas y dónde: el Document Type Definition (DTD).
- La definición de la estructura y el contenido de un tipo de documento se realiza por medio de su DTD (Document Type Definition).

Sin embargo, la potencia de SGML implica una dificultad en su aprendizaje y uso.

En definitiva, SGML apareció para resolver temas de compatibilidad para el intercambio de documentos estrictamente formateados entre diferentes plataformas de ordenadores. Anteriormente, existían opciones limitadas para el intercambio electrónico de documentos en un formato coherentemente útil para el receptor. Así los creadores de SGML desarrollaron un formato para documentos, independientemente de la plataforma.

Ventajas de SGML:
- Reutilización de los datos.
- Integridad y mayor control sobre los datos.
- Portable.
- Flexible.
- Perdurabilidad de la información.

Inconvenientes de SGML:
- Alta complejidad

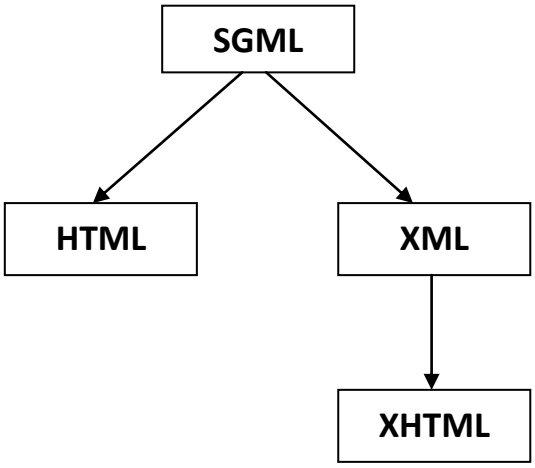
Ejemplo SGML:

```
<antologia>
    <poema>
        <titulo>Soneto Número 1</titulo>
        <estrofa>
            <verso>Un soneto me manda hacer Violante,</verso>
            <verso>que en mi vida me he visto en tanto
```

```
            aprieto;</verso>
                <verso>catorce versos dicen que es soneto,</verso>
                <verso>burla burlando van los tres delante.</verso>
        </estrofa>
            <!-- resto del poema -->
    </poema>
    <!-- otros poemas de la antologia -->
</antologia>
```

A partir de aquí podemos decir que:

– El HTML se crea a partir del SGML.
– XML surge como respuesta al desorden que supuso el rápido crecimiento del HTML, por lo que podemos decir que su precedencia es SGML.



Se puede decir que existen tres utilizaciones básicas de los lenguajes de marcas en documentos: los que sirven principalmente para describir su contenido, los que sirven más que nada para definir su formato y los que realizan las dos funciones indistintamente. Las aplicaciones de bases de datos son buenas referencias del primer sistema, los programas de tratamiento de textos son ejemplos típicos del segundo tipo, y aunque no lo parezca, el HTML es la muestra más conocida del tercer modelo.

1.2 Características de los lenguajes de marcas.

1) 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 más usados son: ASCII, ISO-8859-1 o Latín-1, Unicode, etc...

Una de las principales ventajas del lenguaje de marcas es que puede ser interpretada directamente, dado que son archivos de texto plano. 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 70, 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.

2) Compacidad (o compactación).

Las instrucciones de marcado se entremezclan con el propio contenido en un único archivo o flujo de datos. Este es un ejemplo en diferentes lenguajes de marcas:

Ejemplos	**HTML**	**LaTeX**	**Wikitexto**
Título	Título	\section{Título}	== Título ==
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 ' '

El código entre corchetes como <ul>, o con códigos \section, son instrucciones de marcado, también llamados etiquetas. Estas etiquetas en concreto son descriptivas de la estructura del documento, pudiendo ser su presentación visual de varias maneras. La etiqueta **i** (de italics, cursiva), por el contrario, específica que el texto se debe mostrar en cursiva, sin especificar el motivo de esta diferenciación: es una etiqueta de presentación. El texto entre estas instrucciones es el propio contenido del documento.

3) Facilidad de procesamiento.

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, ha 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 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 del XML.

4) Flexibilidad.

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 XHTML+SMILy de XHTML+MathML+SVG.

1.3 Clasificación e identificación de los más relevantes. Utilización en distintos ámbitos.

Normalmente los lenguajes de marcas se suelen clasificar en tres tipos, aunque en la práctica nos podemos encontrar varias clases en el mismo documento. Por ejemplo HTML tiene etiquetas del tipo puramente procedimental, como la etiqueta <b> (**bold**, para establecer una serie de caracteres en negrita), junto con las puramente descriptivas, como <href …> que sirve para crear enlaces. La clasificación es la siguiente:

. De presentación:

- Indica el formato del texto (información para el maquetado).

El ***marcado de presentación*** es aquel que indica el formato del texto. 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. 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 ejemplo: Rich Text Format (RTF), S 1000D, TeX, troff, HTML…

. De procedimientos:

- Orientado también a la presentación pero, en este caso, se indican los procedimientos que deberá realizar el SW 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. En sistemas más avanzados se utilizan macros o pilas que facilitan el trabajo.

Algunos ejemplos de marcado de procedimientos son nroff, troff, TeX. Este tipo de marcado se ha usado extensamente en aplicaciones de edición profesional, manipulados por tipógrafos cualificados, ya que puede llegar a ser extremadamente complejo.

. Descriptivo o semántico:

- 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.

Las etiquetas pueden utilizarse para añadir al contenido cualquier clase de metadatos. Por ejemplo, el estándar Atom, un lenguaje de gestión, 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, no como en los grandes objetos binarios (blobs) como en el pasado. Estos sistemas no tienen un esquema estricto como las bases relacionales, por lo que a menudo se las considera bases semiestructuradas.

Por ejemplo: ASN.1, EBML, YAML.

Podemos encontrar muchas más clasificaciones, pero la expuesta anteriormente es la más descriptiva.

Algunos ejemplos de lenguajes de marcas podrían ser:

- **Documentación electrónica.**
  - RTF
  - TeX
  - Wikitexto
  - DocBook

- **Tecnologías de internet.**
  - HTML, XHTML
  - RDF
  - RSS

- **Otros lenguajes especializados.**
  - MathML
  - VoiceXML
  - SVG
  - MusicXML

1.4 Introducción a los principales lenguajes de marcas.

Comentaremos en este apartado las principales características de los lenguajes HTML y XML.

Con respecto a HTML.

Las páginas web pueden ser vistas por el usuario mediante un tipo de aplicación llamada navegador. Podemos decir por lo tanto que el HTML es el lenguaje usado por los navegadores para mostrar las páginas webs al usuario, siendo hoy en día la interface más extendida en la red. Es un estándar compuesto por recomendaciones publicadas en un consorcio internacional: el World Wide Web Consortium (W3C).

Este lenguaje nos permite aglutinar textos, sonidos e imágenes y combinarlos a nuestro gusto. Además, y es aquí donde reside su ventaja con respecto a libros o revistas, el HTML nos permite la introducción de referencias a otras páginas por medio de los enlaces hipertexto.

El HTML se creó en un principio con objetivos divulgativos. No se pensó que la web llegara a ser un área de ocio con carácter multimedia, de modo que, el HTML se creó sin dar respuesta a todos los posibles usos que se le iba a dar y a todos los colectivos de gente que lo utilizarían en un futuro. Sin embargo, pese a esta deficiente planificación, si que se han ido incorporando modificaciones con el tiempo, estos son los estándares del HTML. Numerosos estándares se han presentado ya.

Esta evolución tan anárquica del HTML ha supuesto toda una serie de inconvenientes y deficiencias que han debido ser superados con la introducción de otras tecnologías accesorias capaces de organizar, optimizar y automatizar el funcionamiento de las webs. Ejemplos que pueden sonaros son las CSS, Javascript u otros.

Otros de los problemas que han acompañado al HTML es la diversidad de navegadores presentes en el mercado los cuales no son capaces de interpretar un mismo código de una manera unificada. Esto obliga al webmaster a, una vez creada su página, comprobar que esta puede ser leída satisfactoriamente por todos los navegadores, o al menos, los más utilizados.

Además del navegador necesario para ver los resultados de nuestro trabajo, necesitamos evidentemente otra herramienta capaz de crear la página en sí. Un archivo HTML (una página) no es más que un texto. Es por ello que para programar en HTML necesitamos un editor de textos.

Es recomendable usar el Bloc de notas que viene con Windows, u otro editor de textos sencillo. Hay que tener cuidado con algunos editores más complejos como Wordpad o Microsoft Word, pues colocan su propio código especial al guardar las páginas y HTML es **únicamente texto plano**, con lo que podremos tener problemas.

Existen otro tipo de editores específicos para la creación de páginas web los cuales ofrecen muchas facilidades que nos permiten aumentar nuestra productividad. No obstante, es aconsejable en un principio utilizar una herramienta lo más sencilla posible para poder prestar la máxima atención a nuestro código y familiarizarnos lo antes posible con él. Siempre tendremos tiempo más delante de pasarnos a editores más versátiles con la consiguiente ganancia de tiempo.

Es importante tener claro todo ello puesto que en función de vuestros objetivos puede que, más que aprender HTML, resulte más interesante aprender el uso de una aplicación para la creación de páginas.

Así pues, una página es un archivo donde está contenido el código HTML en forma de texto. Estos archivos tienen extensión .html o .htm (es indiferente cuál utilizar). De modo que cuando programemos en HTML lo haremos con un editor de textos y guardaremos nuestros trabajos con extensión .html, por ejemplo mipágina.html

Consejo: Utiliza siempre la misma extensión en tus archivos HTML. Eso evitará que te confundas al escribir los nombres de tus archivos unas veces con .htm y otras con .html. Si trabajas con un equipo en un proyecto todavía más importante que os pongáis todos de acuerdo en la extensión.

Con respecto a XML.

Es una simplificación y adaptación del SGML y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML). Por lo tanto 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 y casi cualquier cosa imaginable.

XML es una tecnología sencilla que tiene a su alrededor otras 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.

ISO

La **Organización Internacional para la Estandarización** o **ISO**, nacida tras la Segunda Guerra Mundial, 23 de febrero de 1947), 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.

La ISO es una red de los institutos de normas nacionales de 163 países, sobre la base de un miembro por país, con una Secretaría Central en Ginebra (Suiza) que coordina el sistema. La Organización Internacional de Normalización (ISO), está compuesta por delegaciones gubernamentales y no gubernamentales subdivididos en una serie de subcomités encargados de desarrollar las guías que contribuirán al mejoramiento ambiental.

Las normas desarrolladas por ISO son voluntarias, comprendiendo que ISO es un organismo no gubernamental y no depende de ningún otro organismo internacional, por lo tanto, no tiene autoridad para imponer sus normas a ningún país...

Está compuesta por representantes de los organismos de normalización (ON) nacionales, que produce normas internacionales industriales y

comerciales. Dichas normas se conocen como *normas ISO* y su finalidad es la coordinación de las normas nacionales, en consonancia con el Acta Final de la Organización Mundial del Comercio, con el propósito de facilitar el comercio, el intercambio de información y contribuir con normas comunes al desarrollo y a la transferencia de tecnologías.

W3C

El Consorcio World Wide Web (W3C) es una comunidad internacional donde las organizaciones Miembro, personal 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.

El objetivo del W3C es guiar la Web hacia su máximo potencial a través del desarrollo de protocolos y pautas que aseguren el crecimiento futuro de la Web.

Principios

Los siguientes principios guían el trabajo del W3C.

Web para todo el mundo

El valor social que aporta la Web, es que ésta hace posible la comunicación humana, el comercio y las oportunidades para compartir conocimiento. Uno de los objetivos principales del W3C es hacer que estos beneficios estén disponibles para todo el mundo, independientemente del hardware, software, infraestructura de red, idioma, cultura, localización geográfica, o habilidad física o mental.

Web desde cualquier dispositivo

La cantidad de dispositivos diferentes para acceder a la Web ha crecido exponencialmente. Actualmente, los teléfonos móviles, teléfonos inteligentes, PDAs, sistemas de televisión interactiva, sistemas de respuesta de voz, puntos de información e incluso algunos pequeños electrodomésticos pueden acceder a la Web.

Visión

La visión del W3C para la Web incluye la participación, compartir conocimiento y, de esta forma, construir confianza a gran escala.

Web de los Autores y Consumidores

La Web fue creada como una herramienta de comunicación para permitir el intercambio de información entre todo el mundo y desde cualquier lugar. Durante muchos años, para muchas personas la Web fue una herramienta de "solo lectura". Los blogs y wikis trajeron más autores a la Web y las redes sociales emergieron del próspero mercado para crear contenido y personalizar las experiencias en la Web. Los estándares del W3C han apoyado esta evolución gracias a la robusta arquitectura y a los principios de diseño.

Web de los Datos y Servicios

Algunas personas ven la Web como un repositorio gigante de datos enlazados mientras otros como un conjunto enorme de servicios que intercambian mensajes. Ambas vistas son complementarias y los requisitos de cada aplicación pueden ser los mejores determinantes para decidir que aproximación elegir para solucionar progresivamente los problemas complejos mediante tecnología Web.

Web de Confianza

La Web ha cambiado la forma en la que nos comunicamos. Al ocurrir esto, la naturaleza de nuestras relaciones sociales ha cambiado también. En la actualidad, las personas se "conocen en Internet", y llevan a cabo relaciones personales y comerciales sin haberse visto en persona anteriormente. El W3C reconoce que la confianza es un fenómeno social, pero el diseño de las tecnologías puede fomentar la confianza y la responsabilidad. A medida que cualquier actividad se hace a través de la Web, cada vez es más importante apoyar las interacciones complejas entre distintas partes alrededor del mundo.