TEMA 1

INTRODUCCIÓN A LOS LENGUAJES DE MARCAS

Módulo: **Lenguages de Marcas y Sistemas de Gestión de la Información**

Ciclo: **DAW**

Introducción

Un **lenguaje de marcas, o lenguaje de marcado,** es una forma de codificar un documento donde, junto al propio texto, se incorporan una serie de etiquetas, marcas o anotaciones con el objetivo de proporcionar información adicional sobre la estructura del texto o su formato de presentación. Por tanto, los lenguajes de marcado permiten hacer explícita la estructura de un documento, su contenido semántico o cualquier otra información lingüística o extralingüística que se quiera resaltar.

Cada lenguaje de marcas se define en base a un documento llamado DTD[1] en el que se establecen los diferentes elementos utilizados por dicho lenguaje: las etiquetas reconocidas, la forma de construir nuevas etiquetas, los atributos y sus valores válidos, la sintaxis y las reglas de uso, en general

A continuación se muestra un ejemplo de un documento creado utilizando el lenguaje de marcado XML:

```
<carta>
      <fecha>22/11/2024</fecha>
      <presentación>Estimado cliente:</presentación>
      <contenido>Aquí está la copia de la factura que solicitó
      </contenido>
      <firma>María Segunda Castro</firma>
</carta>
```

Una forma habitual de clasificar los lenguajes de marcado es según su objeto funcional (aunque en la práctica, en un mismo documento se pueden combinar varios lenguajes de marcado diferentes), distinguiendo:

- **Presentacional :** este tipo de lenguaje de marcado define el formato del texto.
- **Procedural :** también orientado a la presentación, pero siendo un programa que interpreta el código en orden secuencial.
- **Descriptivos o semánticos:** describen las diferentes partes en las que se estructura el documento pero sin especificar cómo deben representarse.

[1] Document Type Definition (Definición de Tipo de Documento) .

Algunos ejemplos de lenguajes marcarios (ya sea por su uso habitual o por la particularidad de su objeto) a lo largo de la historia son:

- **RTF[2] :** Diseñado para la producción de documentos de texto enriquecido, fue introducido por Microsoft en 1987. Es compatible con los procesadores de texto más utilizados, lo que lo ha convertido en un estándar *de facto* para el intercambio de documentos.
- **TeX :** es un sistema tipográfico desarrollado por Donald. E. Knuth en 1978 y que permite representar ecuaciones matemáticas complejas así como crear documentos de texto enriquecido.
- **Wikitext :** así se llama a cualquiera de los muchos lenguajes de marcado para la creación de wikis, entre los que destaca la variante que utiliza el *software* MediaWiki que soporta proyectos de Wikimedia (Wikipedia, Wikiquote , Wikcionario , etc.). El primer lenguaje *wiki* fue desarrollado por Ward Cunningham en 1995 para albergar un repositorio de patrones .
- **DocBook :** es una variante del estándar SGML/XML utilizado especialmente para la elaboración de documentación técnica y cuya característica más destacable es que separa la estructura lógica del documento de su formato, facilitando la maquetación sin necesidad de alterar el contenido original.
- **HTML[3]/XHTML[4]:** nacidos en 1991 y 2000 respectivamente, son los dos lenguajes fundamentales para la creación de páginas web .
- **XML[5]:** Fue introducido por el W3C[6] en 1998 y en este tiempo se ha convertido en el formato de intercambio de documentos entre sistemas más utilizado.
- **RSS[7]:** es una variante de XML presentada en 1999 y tiene como objetivo facilitar la distribución de contenidos a través de Internet, siendo quizás su uso más conocido el de alimentar a los agregadores de noticias.
- **MathML :** también basado en XML, este formato pretende facilitar la comunicación de expresiones en notación matemática para que diferentes máquinas puedan entenderla, así como combinarse con XHTML para mostrar contenidos matemáticos vía web .

---

[2] Rich Text Format (Formato de Texto Enriquecido).

[3] *HyperText Markup Language.*

[4] *HTML extensible* .

[5] *eXtensible Markup Language.*

[6] World Wide Web Consortium.

[7] Really Simple Syndication.

- **VoiceXML :** tiene como finalidad el intercambio de información de audio entre usuarios y aplicaciones de síntesis y reconocimiento de voz a través de servidores web (siendo también integrable en sistemas telefónicos).
- **MusicXML :** Introducido en 2011, este dialecto de XML permite representar partituras en notación occidental utilizando archivos XML que luego facilitan su intercambio entre máquinas así como su interpretación en *sintetizadores de software.*

Evolución de los lenguajes de marcas

En la década de 1970, con la expansión de la informática comercial y el crecimiento asociado a ella en el intercambio electrónico de documentos, **surgieron los primeros lenguajes informáticos destinados no a programar, sino a facilitar el flujo de información que requiere el mercado**. Con el desarrollo de los editores y procesadores de texto aparecen los primeros lenguajes especializados en tareas de descripción y estructuración de información: **los lenguajes de marcado** . Paralelamente se introdujeron otros lenguajes informáticos encaminados a la representación, almacenamiento y consulta eficiente de grandes volúmenes de información: **los lenguajes y sistemas de bases de datos.**

Los primeros lenguajes de marcado aparecieron como resultado del conjunto de códigos de formato que los procesadores de texto insertaban en los documentos para dirigir el proceso de presentación a través de una impresora. Al igual que con los lenguajes de programación originales, y como sigue sucediendo con los lenguajes de programación de bajo nivel, esos códigos de formato dependían de la arquitectura de la máquina subyacente y del programa que los generaba, lo que no permitía a los programadores y diseñadores abstraerse de las características del sistema ni expresar independientemente de él la estructura y lógica interna del documento.

Un ejemplo (de sintaxis ficticia, pero que bien podría haber sido auténtico) sería algo como el siguiente:

```
<times 14><color blue><center>Este texto sirve como ejemplo del
uso primitivo de marcas en el procesamiento y formato de
documentos.</center></color></times>
<color rojo><times 10><italic>En este ejemplo utilizamos marcas de
nuestra invención.</italic> Las partes importantes se pueden
enfatizar usando <black>negro</black> o
<underline>subrayado</underline>.</times></color>
```

Imprimiendo lo anterior obtendríamos algo similar a esto:

Este texto sirve como ejemplo del uso primitivo de marcas en el procesamiento y formato de documentos.

En este ejemplo utilizamos marcas de nuestra invención. Las partes importantes se pueden enfatizar usando **negrita** o subrayado .

Posteriormente se integraron los lenguajes de marca como medio de presentación en pantalla. Actualmente, los códigos están ocultos de forma predeterminada (al menos en la mayoría de los procesadores de texto) y el formateo se realiza seleccionando acciones, ya sea haciendo clic con el mouse o presionando una combinación de teclas predeterminada; pero en realidad los códigos siguen ahí, y se insertan donde corresponde en el documento para que el formato final coincida con el marcado por el usuario.

La necesidad de estándares que favorecieran el intercambio en la industria, dio lugar al desarrollo de los primeros lenguajes de marcas con una sintaxis preestablecida, que luego podía traducirse al lenguaje de la máquina en uso.

GML

Desarrollado entre 1969 y 1970 en los laboratorios de IBM por Charles Goldfarb , Edward Mosher y Raymond Lorie , GML[8] fue concebido como parte de la construcción de un sistema de edición, almacenamiento y búsqueda de documentos legales. Tras estudiar el escenario de la empresa, Goldfarb , Mosher y Lorie concluyeron que para realizar un buen procesamiento informatizado de los documentos era necesario establecer un formato estándar para todos los documentos que en ella se manejaban. Lo anterior lograría el objetivo de gestionar cualquier documento de cualquier departamento y con cualquier aplicación, independientemente de dónde o cómo se generó dicho documento. El formato tenía que ser válido para los diferentes tipos de documentos legales utilizados por IBM, por lo que tenía que ser flexible y adaptable a diferentes situaciones.

SGML

El uso del lenguaje GML se ha ido extendiendo en el mundo empresarial a lo largo de sus primeros quince años de existencia hasta convertirse en un estándar *de facto* para el intercambio de información estructurada. Tanto es así que en 1986 la ISO[9] , organismo encargado de establecer estándares comunes para la industria a nivel global, creó SGML[10] que se definió como un estándar *de iure* con la norma **ISO 8879** .

SGML es un lenguaje complejo y generalmente se utilizaba a partir de herramientas *de software propietario* con licencias relativamente caras, por lo que su uso se limitaba a grandes empresas industriales.

---

[8] *Generalized Markup Language*

[9] *International Organization for Standardization .*

[10] *Standard GML*.

Un documento SGML simple, por ejemplo, se vería así:

```
<correo electrónico>
  <remitente>
    <persona>
       <nombre>Juan</nombre>
       <apellido>Perillan</apellido>
    </persona>
  </remitente>
  <destinatario>
     <dirección>perillan@correo.gal</dirección>
  </destinatario>
  <subject>¿Tomamos algo?</subject>
  <mensaje> Hola, sé que los bares están abiertos nuevamente hoy.
Habrá que ir, ¿no?
  </mensaje>
</correo electrónico>
```

HTML

Entre 1989 y 1991 Tim Berners-Lee desarrolla (inicialmente solo y luego con la colaboración de Robert Caillau), mientras trabaja en el CERN, el Mundo Ancho Web , que supone la creación del primer servidor web , el primer navegador web y el primer lenguaje de programación web , HTML[11]. La necesidad de satisfacer la idea original de la Web de vincular y compatibilizar una gran cantidad de información procedente de diversos sistemas lleva a la creación de un lenguaje de descripción de documentos (HTML) que, en la práctica, es una combinación y reestructuración de dos pre-normas existentes:

- **ASCII[12]:** código de caracteres basado en la variante inglesa moderna del alfabeto latino y que es compatible con todos los sistemas informáticos de uso general.
- **SGML:** proporciona estructura y permite formatear el texto.

La naturaleza gratuita de la mayoría de las herramientas web (navegadores, servidores, editores, etc.) más el hecho de que HTML, como dialecto simplificado de SGML (la versión original sólo utilizaba las etiquetas absolutamente esenciales) era fácil de entender, dio lugar a una rápida aceptación de este lenguaje. De esta manera, HTML logró algo que SGML no pudo: ser un estándar de uso global.

[11] *HyperText Markup Language*.

[12] *American Standard Code for Information Interchange.*

A pesar de esas ventajas, HTML no es un lenguaje ideal y también tiene sus desventajas (que se solucionaron posteriormente con la adición de lenguajes complementarios para la Web en el estándar HTML5):

- No soporta tareas de impresión y diseño .
- Las etiquetas son limitadas, lo que reduce la flexibilidad del idioma.
- No permite mostrar contenido dinámico[13].
- Mezcla estructura y diseño en un mismo documento.

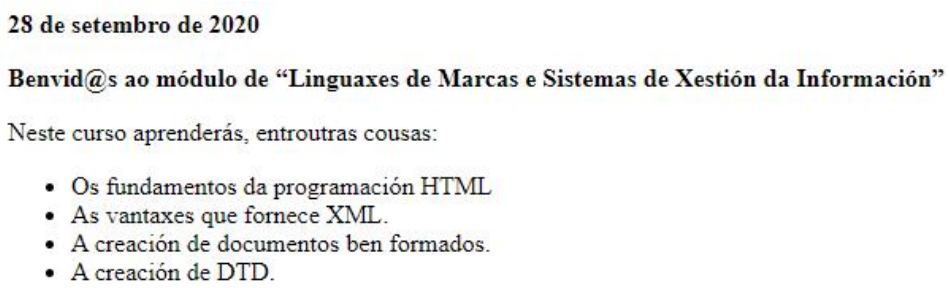
Un ejemplo de un documento HTML típico (el código de una página web) sería el siguiente:

```
<html>
  <head>
    <title>Exemplo de código HTML</title>
  </head>
  <body bgcolor="#ffffff">
    <p></p>
    <p>
      <b>28 de setembro de 2020</b>
    </p>
    <p><b>Benvid@s ao módulo de "Linguaxes de Marcas e Sistemas de
Xestión da Información"</b></p>
    <p>Neste curso aprenderás, entroutras cousas:<br/>
      <ul>
        <li>Os fundamentos da programación HTML</li>
        <li>As vantaxes que fornece XML.</li>
        <li>A creación de documentos ben formados.</li>
        <li>A creación de DTD.</li>
      </ul>
    </p>
  </body>
</html>
```

Ejemplo 1.1: código HTML de muestra.

[13] En Web nos referimos a contenidos dinámicos que son resultado de la interacción con el usuario y que, por tanto, cambian en función de ella.

Si tomamos el código anterior, lo escribimos en un documento de texto sin formato, cambiamos la extensión del archivo a .htm o .html y lo abrimos con un navegador , deberíamos ver algo similar a esto:



28 de setembro de 2020

Benvid@s ao módulo de "Linguaxes de Marcas e Sistemas de Xestión da Información"

Neste curso aprenderás, entroutras cousas:

- Os fundamentos da programación HTML
- As vantaxes que fornece XML.
- A creación de documentos ben formados.
- A creación de DTD.

XML

Con el objetivo de solucionar algunos de los problemas propios del HTML, el W3C presentó, en 1998, el estándar XML, un lenguaje de marcado puramente estructural que no incluye ningún tipo de información de diseño. En sus más de veinte años de vida, XML se ha convertido en el estándar preferido para el intercambio de datos vía Web , así como la principal forma de exportar cierto tipo de información (archivos de configuración, *logs* , etc.).

A diferencia de HTML, las etiquetas XML siempre indican el significado de los datos que contienen.

Entre las principales características del metalenguaje XML cabe destacar:

- Le permite definir sus propias etiquetas.
- Le permite asignar atributos a las etiquetas.
- Utiliza un esquema conocido para definir con precisión etiquetas y atributos.
- La estructura y el diseño son independientes.

Realmente, bajo las siglas XML encontramos una serie de estándares relacionados, como por ejemplo:

- **XSL[14]:** le permite definir hojas de estilo para documentos XML, incluida la capacidad de transformar documentos.
- ***Enlace* XML *Idioma* :** que incluye tecnologías como XPath , XLink o XPointer. Se utilizan para determinar aspectos de los vínculos entre documentos XML.
- **XML *Espacios de nombres :*** Proporcionan un contexto al que se aplican las marcas de un documento XML, lo que permite distinguirlos de otros con el mismo nombre en otros contextos.

[14] *eXtensible Style Language.*

- ***Esquemas* XML :** le permiten definir restricciones que se aplicarán a un documento XML.

A continuación vemos un ejemplo de un archivo XML que representa los datos relacionados con las copias de una biblioteca:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE biblioteca">
<biblioteca>
  <exemplar tipo_ex="libro" titulo="XML practico"
editorial="Ediciones Eni">
    <tipo>
      <libro isbn="978-2-7460-4958-1" edicion="1"
paginas="347"></libro>
    </tipo>
    <autor nome="Sebastien Lecomte"></autor>
    <autor nome="Thierry Boulanger"></autor>
    <autor nome="Ángel Belinchon Calleja"
funcion="traductor"></autor>
    <emprestado lector="Xan Perillán">
      <data_pres dia="13" mes="mar" ano="2019"></data_pres>
      <data_devol dia="21" mes="xun" ano="2019"></data_devol>
    </emprestado>
  </exemplar>
  <exemplar tipo_ex="revista" titulo="Todo Linux 101.
Virtualización en GNU/Linux" editorial="Studio Press">
    <tipo>
      <revista>
        <data_publicacion mes="abr" ano="2009"></data_publicacion>
      </revista>
    </tipo>
    <autor nome="Varios"></autor>
    <emprestado lector="Uxía Moure">
    <data_pres dia="12" mes="ene" ano="2010"></data_pres>
    </emprestado>
  </exemplar>
</biblioteca>
```

Comparativas

HTML frente a XML

XML:

- Es un dialecto de SGML.
- Especifica cómo se deben unir los conjuntos de etiquetas aplicables a un tipo de documento.
- Soporta un modelo de hipervínculo complejo (Redireccionamiento a una URL habitual y, posteriormente, a una sección de su página).
- Utiliza el navegador como plataforma para desarrollar aplicaciones.
- Significó el fin de la "guerra de los navegadores" y de las etiquetas propietarias (es decir, aquellos que solo se podían utilizar para una tecnología o empresa específica).

HTML:

- También es un dialecto de SGML.
- Aplica un conjunto limitado de etiquetas predeterminadas a un solo tipo de documento.
- Es un modelo de hipervínculo simple (Redireccionamiento a una URL habitual).
- Utilice el navegador como visor de páginas web.
- Ganó con su estabilidad la "guerra de los navegadores " y dio comienzo el nacimiento de XHTML.

XML frente a SGML

XML:

- Su uso es mucho más sencillo.
- Trabajar con documentos bien formados, sin necesidad de validarlos.
- Facilita el desarrollo de aplicaciones de bajo coste.
- Tiene muchas áreas de aplicación más allá de la pura informática (Finanzas, desarrollo web, comunicación de datos entre aplicaciones…).
- Presenta alta compatibilidad e integración con HTML.

SGML:

- Su uso es muy complejo.
- Trabaja únicamente con documentos validados.
- Su complejidad encarece mucho las aplicaciones informáticas que lo procesan.

- Solo se utiliza en sectores muy concretos.
- No hay soporte HTML definido. **SGML es un lenguaje de descripción de lenguajes de marcado**, y **HTML es uno de esos lenguajes** que puede derivarse de SGML. Pero por sí solo, SGML no "soporta" HTML directamente; se necesita un conjunto de reglas o especificaciones para definir cómo debería funcionar HTML en el marco de SGML.

En conclusión:

- HTML y XML tienen su origen en SGML.
- **SGML No es un lenguaje de marcado por sí mismo**, sino un conjunto de reglas que permite crear lenguajes de marcado como HTML o XML.
- HTML es un lenguaje de marcado específico basado en SGML, pero enfocado en la visualización de contenido web.
- XML es una simplificación de SGML, diseñada para ser más fácil de usar y enfocada en la estructuración de datos.

Etiquetas o marcas

Las etiquetas o **marcas** son el elemento definitorio de los lenguajes de marca, y constan de caracteres o conjuntos de estos reservados que permiten estructurar los documentos de texto, facilitando la interpretación de la información contenida en ellos **.**

Aunque hubo y hay muchas formas diferentes de representar etiquetas, la forma más común y extendida es la que utiliza los signos **<** y **>** para limitar las marcas, y que fue definida por SGML y heredada por lenguajes como HTML y XML. Esta es la sintaxis de la que hablaremos aquí.

Se suele utilizar una etiqueta de inicio y una etiqueta de fin para indicar que el elemento que queríamos presentar ha finalizado. En el caso de la sintaxis que nos ocupa, la etiqueta de cierre difiere porque lleva la *etiqueta de avance barra diagonal* seguida del carácter de apertura. Por ejemplo, si nuestra etiqueta de apertura es:

```
<persona>
```

Entonces la etiqueta de cierre correspondiente será:

```
</persona>
```

Las últimas especificaciones emitidas por el W3C indican la necesidad de que las etiquetas se escriban siempre en minúsculas para considerar que el documento está creado correctamente.

Ventajas de los lenguajes de marca

A modo de resumen, podemos concluir este tema citando algunas de las ventajas de trabajar con lenguajes de marca, como por ejemplo:

- **Texto plano:** Los documentos escritos utilizando lenguajes de marcado se pueden editar utilizando editores de texto sencillos, ya que las marcas se intercalan con el contenido, aunque se podrían utilizar editores más potentes para facilitar el trabajo.

  Sumado a lo anterior, el hecho de tratarse de texto plano hace que los documentos sean independientes de la plataforma, sistema operativo o programa con el que fueron creados.
- **Compacto:** las instrucciones de marcado se intercalan con el propio contenido, en un único archivo.
- **Facilidad de procesamiento:** las etiquetas del lenguaje de marcado son fácilmente interpretables y procesables por computadora.
- **Flexibilidad:** Aunque los lenguajes de marcado fueron diseñados originalmente para representar documentos de texto, con el tiempo se han utilizado con éxito en otras áreas como gráficos vectoriales, servicios web o interfaces de usuario. Estas nuevas aplicaciones aprovechan la simplicidad y el poder de los lenguajes de marca.