

UD1 Introducción a los lenguajes de marcas

Introducción

Cubren la necesidad de disponer de lenguajes sencillos para el intercambio de información entre distintos sistemas. Con la llegada de Internet, algunos se convirtieron en piedras angulares sobre las que construir World Wide Web-WWW.

Los lenguajes de marcas comparten algunos aspectos comunes: fundamentos, orígenes, características básicas y elementos relacionados con su estructura.

1.1. Introducción a los lenguajes de marcas

Es aquel que permite representar información que contiene, además de los datos propiamente dichos, marcas o etiquetas que indican cómo se estructuran estos datos, su significado y cómo debe representarse desde un punto de vista gráfico o visual.

No son lenguajes de programación, no contienen instrucciones ni implican acciones directas. Son reglas que organizan la información con el objetivo de darle una estructura uniforme y facilitar su procesamiento automático.

Uno de los lenguajes de marcas más representativo es HTML, el etiquetado HTML indica al navegador web cómo debe presentar la información, es decir, proporciona junto a los datos a mostrar información a cómo han de mostrarse. A partir de HTML 5 incluye etiquetas con

información semántica, indicando el significado que los datos tienen desde un punto de vista estructural.

Como por ejemplo:

`<h1>`El origen de la vida`</h1>`

La etiqueta `<h1>` indica al navegador web que el texto debe representarse de una forma determinada y a nivel semántico el texto es el título de un bloque de información del más alto nivel posible, como por ejemplo el título de un capítulo de un libro.

Se mostraría así:

El origen de la vida

1.2. Contexto de los lenguajes de marcas

El origen de los lenguajes de marcas se remonta al origen de la computación. En el momento en que se comenzó a querer almacenar información para procesarla por un programa informático, surgió la necesidad de dar formato a dicha información.

El problema surge cuando cada programa utiliza su propio lenguaje, que obliga a crear un

conversor entre cada par de programas que comparten información.

A finales de los años sesenta y principios de los setenta del siglo XX, el equipo liderado por Charles Goldfarb, de IBM, desarrolló el precursor de los lenguajes de marcas modernos: GML. GML fue el primer paso para conseguir unificar los formatos de representación de información, centrándose en los procesadores de texto.

SGML es el estándar creado en 1986 por la ISO, para desvincular GML de un ente privado -IBM- evitando las implicaciones que podrían impedir su popularización. HTML (hasta la versión 4) y XML son lenguajes de marcas creados a partir de SGML.

XML se ha convertido en un lenguaje de intercambio de datos entre sistemas por su simplicidad, robustez y capacidad. Es la evolución natural de SGML, aunque HTML es el lenguaje de marcas más conocido.

Tim Berners-Lee desarrolló la WWW entre 1989 y 1990 trabajando para el CERN. En 1993 la web adquirió su carácter público y gratuito. Tiene como objetivo principal simplificar el acceso a la información, tanto desde un punto de vista formal (estructural) como funcional (acceso).

Estas dos perspectivas consiguen que la información sea accesible desde cualquier sistema operativo, ya que utiliza un lenguaje de marcado estándar (por ejemplo, HTML) y la obtención de información relacionada es sencilla (hipertextos).

1.3. Lenguajes de marcas: características, estructura y formatos

Como ya hemos comentado, cuando se habla de lenguaje de marcas, se está haciendo referencia a la información añadida mediante marcas a los datos propiamente dichos del documento.

Cada lenguaje de marcas tiene su propia estructura y formato. Con el objetivo de poner un ejemplo, se comenta parte de la estructura de uno de los más importantes, XML.

- Cada elemento de un documento XML está precedido por una etiqueta de apertura compuesta por un nombre y rodeada por los símbolos <y>.
- Cada elemento de un documento XML está seguido por una etiqueta de cierre, compuesta por el mismo nombre de la etiqueta de apertura y rodeada por los símbolos </y>.
- El contenido de un elemento de un documento XML puede ser un texto o uno o más elementos.
- Solo existe un elemento raíz.
- Un atributo está compuesto de un nombre del atributo, el símbolo = y el valor del atributo entrecomillado.
- Un elemento puede tener entre 0 y n atributos.

El formato en el que está escrito un documento XML es texto plano.

Un ejemplo podría ser el fichero *misjuegos.xml* con el siguiente contenido:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<videojuegos>
  <videojuego multijugador="s1">
    <titulo>Call Of Duty</titulo>
    <plataforma>Play Station 5</plataforma>
  </videojuego>
  <videojuego multijugador="s1">
    <titulo>Fortnite</titulo>
    <plataforma>XBox One</plataforma>
  </videojuego>
  <videojuego multijugador="no">
    <titulo>Manic Miner</titulos>
    <plataforma>Amstrad CPC</plataforma>
  </videojuego>
</videojuegos>
```

Es el autor del documento quien decide cómo se han de llamar las etiquetas, cuál es su orden y cómo se estructuran entre sí. Esto es debido a que XML es un metalenguaje.

Por su parte, HTML no es un metalenguaje y aunque presenta también una estructura bien

concreta, tiene definidos con precisión las etiquetas y atributos.

```
<table class="fondoGris">
  <tr>
    <td>Elemento 1</td>
    <td>Elemento 2</td>
  </tr>
  <tr>
    <td>Elemento 3</td>
    <td>Elemento 4</td>
  </tr>
</table>
```

Las etiquetas <table>, <tr> y <td> están definidas en HTML, así como el atributo *class*.

En general, las características fundamentales de los lenguajes de marcas son:

- *Independencia*: su uso no debe depender de ninguna plataforma hardware ni sistema software concreto.
- *Almacenados en texto plano*: el documento está compuesto únicamente por caracteres de texto.
- *Flexibilidad*: deben ser lo suficientemente flexibles para poder utilizarse en diferentes

contextos.

- *Compactibilidad*: las marcas y el contenido se encuentran en el mismo documento.
- *Facilidad de procesamiento*: el formato debe ser diseñado de tal manera que el procesamiento automático sea sencillo.

1.4. Lenguajes de marcas y dialectos más comunes

- HTML: el más conocido y utilizado. Actualmente se encuentra en la versión 5, en su revisión 5.3. Su principal objetivo es diseñar páginas web que van a ser representadas en un navegador. La extensión de sus ficheros suele ser *.html* o *.htm*.
- XML: es el lenguaje de intercambio de información entre sistemas por excelencia. Se puede afirmar que XML es para los sistemas el equivalente a HTML para los humanos. La última revisión es la 1.1. La extensión de los ficheros XML suele ser *.xml*.
- XHTML: versión en formato XML de HTML. La diferencia fundamental es que XHTML tiene una sintaxis más estricta y es más formal que HTML. 1.1 es su última revisión. La extensión de los ficheros suele ser *.xhtml*.
- JSON: es el formato de texto en el que JavaScript representa sus objetos. Se ha universalizado como alternativa a XML, pues es más ligero y cuenta con una sintaxis más sencilla que XML. Como desventaja no dispone de las herramientas y riqueza de XML.

JSON y XML coexisten sin problemas, pues dependiendo del contexto, interesa utilizar JSON o XML. La versión más actual es la 2.6.2. Sus ficheros tienen como extensión *.json*.

- SVG, MathML, NewsML, FXML, XAML, RSS, Atom y otros muchos: dialectos de XML con diferentes usos.