

Para comenzar nuestra aventura **HTML**, vamos a introducir el concepto de documento **HTML**, un archivo de texto en formato **HTML**.

Iniciamos la lección introduciendo el concepto de documento **HTML** y de sus componentes básicos: los elementos, los atributos, el contenido y los comentarios. A continuación, presentamos los elementos principales de todo documento **HTML**: `<html>`, `<head>` y `<body>`. Finalmente, examinamos la validación de documentos.

Al finalizar la lección, el estudiante sabrá:

- Qué es un documento **HTML**.
- Qué es un elemento.
- Qué es un atributo de elemento.
- Qué es el contenido de un elemento.
- Cuál es la estructura de un documento **HTML**.
- Qué es la validación.

## Introducción

Un **documento HTML** (*HTML document*) es un archivo de texto escrito bajo el lenguaje de marcas **HTML**. Por convenio, este tipo de archivos tiene la extensión **.html**. Están formados principalmente por elementos.

## Elementos

El lenguaje **HTML** define un conjunto de elementos, donde un **elemento** (*element*) representa un componente individual del documento que describe algo como, por ejemplo, un párrafo, un texto en cursiva, una imagen o un enlace. Estos elementos se delimitan por **etiquetas** (*tags*), marcas o señales que se colocan para determinar el límite del elemento y su contenido.

He aquí un ejemplo:

```
<p>Esto es un ejemplo de párrafo.</p>
```

En **HTML5**, los elementos suelen presentar una etiqueta de inicio y otra de fin, aunque no es necesario, siendo obligatoria únicamente la de inicio. En el ejemplo, la etiqueta de inicio es `<p>` y la de fin `</p>`. Así pues, la **etiqueta de inicio** (*start tag*) es el punto en el que comienza el elemento y contiene su nombre seguido de cero, uno o más atributos. Por otra parte, la **etiqueta de fin** (*end tag*), formada por una barra seguida del nombre del elemento, marca el final del elemento.

## Atributos de elemento

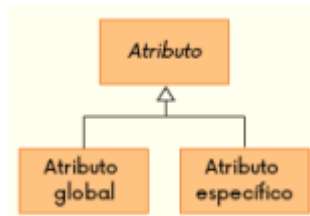
Un **atributo** (*attribute*) no es más que una propiedad a través de la cual proporcionar información adicional sobre un elemento como, por ejemplo, en qué fuente debe visualizarse el contenido del elemento, el tamaño que debe destinarse, etc.

Se ubican en la etiqueta de inicio, entre el nombre del elemento y el símbolo `>` que cierra esta etiqueta. Se especifican mediante pares **nombre=valor**, donde el valor debe de estar delimitado por comillas simples (`'`) o dobles (`"`). También es posible indicar sólo el **nombre**, en cuyo caso su valor será **true**.

A continuación, se muestra un elemento `<meta>` con dos atributos, **name** y **content**:

```
<meta name="author" content="Nodemy Media Labs">
```

Atendiendo a qué elementos se puede adjuntar, los clasificamos en globales y específicos.



Un **atributo global** (*global attribute*) es puede usar en cualquier elemento. Los define el estándar para todos los elementos **HTML**. En cambio, un **atributo específico** (*specific attribute*) es aquel que es particular de un determinado elemento **HTML**. También los define el estándar, pero a nivel de elemento particular.

A lo largo del curso, iremos presentando los atributos más utilizados e importantes.

## Contenido

Cuando un elemento tiene una etiqueta de inicio y de fin, el texto que aparece entre ambas etiquetas se conoce formalmente como **contenido** (*content*).

## Comentarios

Un **comentario** (*comment*) es un texto que no tiene ningún significado gramatical para el analizador de **HTML**. Se delimitan mediante `<!--` y `-->`. Se pueden expandir en varias líneas.

He aquí un ejemplo ilustrativo:

```
<!-- esto es
      un comentario que se
      expande en varias líneas -->
```

## Estructura de un documento HTML

La estructura define el formato de un documento **HTML**. Todo documento debe tener una directiva de tipo de documento, seguida de un elemento `<html>`, el cual delimita el contenido del documento:

```
<!doctype html>

<html>
  ...
</html>
```

## Tipo de documento

El **tipo de documento** (*document type*) es una directiva que indica ante qué tipo de documento estamos y bajo qué especificación se redactó. En el caso de **HTML5**, su sintaxis es la siguiente:

```
<!doctype html>
```

Se especifica al comienzo del documento.

## Elemento raíz

El **elemento raíz** (*root element*) es con el que delimita el contenido del documento. Se trata del elemento `<html>`. Este elemento tiene atributos y un modelo de contenido.

## Modelo de contenido

El **modelo de contenido** (*content model*) define qué otros elementos puede contener un elemento. En el caso del elemento raíz, `<html>`, el modelo de contenido es el siguiente:

```
<html>
<head>
  <!-- ... -->
</head>

<body>
```

```
<!-- ... -->
</body>
</html>
```

El elemento raíz debe estar formado por un elemento de cabecera, `<head>`, que contiene metadatos del documento como, por ejemplo, su autor, las palabras claves, etc., seguido de un cuerpo, `<body>`.

## Atributos del elemento html

Los atributos más frecuentes del elemento `<html>` son los siguientes:

Atributo	Tipo de datos	Descripción
----------	---------------	-------------

<code>lang</code>	Texto	Idioma principal del contenido del documento como, por ejemplo, Español, Inglés, Italiano, etc.
-------------------	-------	---

### Atributo lang

El atributo `lang` no es específico del atributo `<html>`, puede utilizarse en otros muchos elementos. Especifica el idioma del contenido del elemento. En el caso del elemento `<html>`, indica el idioma principal del contenido del elemento; el cual podrá cambiarse en otros elementos que definan su correspondiente atributo `lang`.

Se recomienda encarecidamente indicar el idioma del documento. Así, los motores de búsqueda no tendrán problemas a la hora de determinarlo y, por ende, indexarlos correctamente.

El valor del atributo tiene un formato, estandarizado por el [RFC 5646](#). El formato principal es el siguiente:

idioma  
idioma-país

El **idioma** indica el idioma y el **país** la variante hablada del idioma en un determinado país. Se habla de **idioma neutral** (*neutral language*) como aquel que hace referencia a un determinado idioma, sin indicar su variante local. Mientras que si se indica también el país, se habla de **idioma específico** (*specific language*).

A continuación, se muestra una tabla de valores del atributo `lang`:

Valor	Tipo	Descripción
es	Neutral	Español
es-cl	Específico	Español-Chile
es-cr	Específico	Español-Costa Rica
es-es	Específico	Español-España
fr	Neutral	Francés
fr-be	Específico	Francés-Bélgica
fr-ch	Específico	Francés-Suiza
fr-fr	Específico	Francés-Francia
en	Neutral	Inglés
en-gb	Específico	Inglés-Reino Unido
en-ie	Específico	Inglés-Irlanda
en-us	Específico	Inglés-Estados Unidos

He aquí un ejemplo de uso del Español en su variante neutral:

```
<!doctype html>

<html lang="es">
  <head>...</head>
  <body>...</body>
</html>
```

## Cabecera del documento

El **elemento de cabecera** (*head element*) es aquel que se sitúa al comienzo del elemento raíz y contiene metadatos del documento como, por ejemplo, su título, su autor, sus descripción, las palabras reservadas, etc. Se representa mediante el elemento `<head>`.

## Modelo de contenido

El modelo básico de contenido del elemento `<head>` es el siguiente:

```
<head>
  <title>
  <base>
  <meta>
  <link>
  <style>
  <script>
</head>
```

Mediante elementos `<style>` y `<script>` se utilizan para añadir hojas de estilo **CSS** y código **JavaScript** a la página.

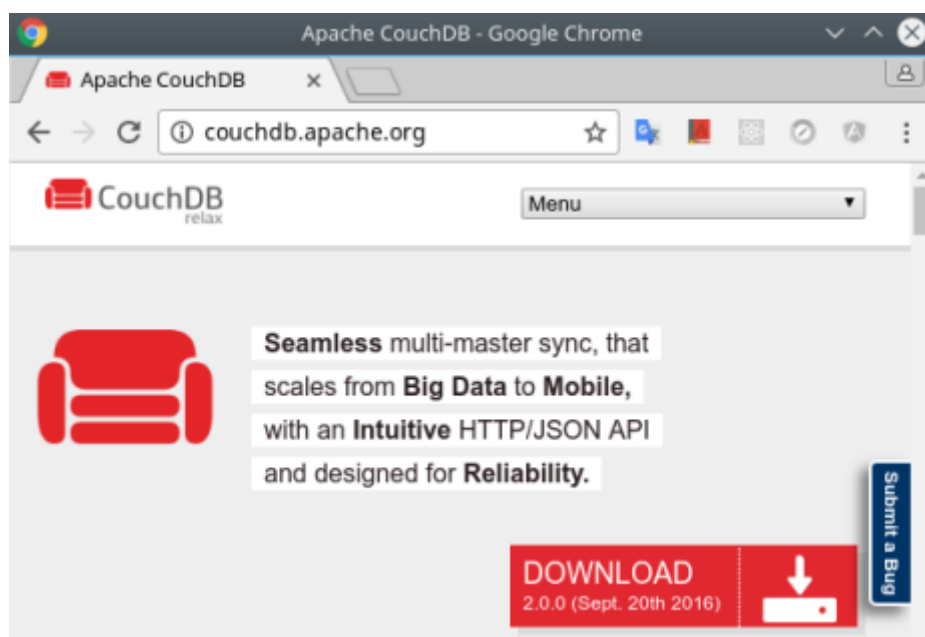
### Título del documento

El **título del documento** (*document title*) es el elemento de la cabecera que en forma de palabra o frase *describe* el contenido del documento. Los navegadores suelen visualizarlo en la pestaña asociada a la página web o en la barra de título. Se indica mediante el elemento `<title>`:

```
<title>el título</title>
```

Se recomienda que todo documento disponga de un título descriptivo, generalmente de menos de 15 palabras de como máximo 50-60 caracteres en total. Los motores de búsqueda lo utilizan para la indexación en sus bases de datos de la página y aparece en sus resultados, por lo que todos los usuarios lo verán si la página aparece en el resultado.

He aquí un ejemplo ilustrativo de visualización del título en la pestaña del navegador:



Generalmente, el título contiene el nombre descriptivo de la página, pero también el de su sitio web. Algunos suelen poner el nombre del portal primero y, después, el título de la página. Otros suelen hacerlo al revés.

Veamos otro ejemplo ilustrativo dentro del sitio web anterior:



### URL base

La **URL base** (*base URL*) representa la URL base del sitio web que debe utilizarse con los URLs relativos que aparecen en el documento como, por ejemplo, en las imágenes, en los enlaces, etc. Se indica mediante el elemento `<base>`:

```
<base href="URL">
```

Se recomienda encarecidamente el uso de la URL base. Generalmente, cuando un portal web tiene varias aplicaciones, cada una de ellas asignada a un directorio particular de URLs. Así, por ejemplo, si tenemos un portal el cual tiene una aplicación web que se encarga de las noticias a la que se le ha asignado el directorio `/noticias`, tendríamos una URL base en todas las páginas de la aplicación de noticias como la siguiente:

```
<base href="http://mi.portal.com/noticias">
```

### Enlaces

Un **enlace** (*link*) es una referencia a un recurso utilizado por el documento como, por ejemplo, una hoja de estilo **CSS** o la ubicación de una versión en otro idioma o formato del documento. Cada enlace se representa mediante un elemento `<link>`:

```
<!-- Hoja de estilo CSS -->
```

```
<link rel="stylesheet" href="url" title="Título descriptivo">
```

```
<!-- Documento alternativo -->
```

```
<link rel="alternate" href="url" hreflang="idioma" type="tipo" title="Título">
```

```
<link rel="alternate" href="url" media="dispositivo" title="Título">
```

### Metadatos

Un **metadato** (*metadatum*) es un dato que contiene información adicional del documento como, por ejemplo, sus palabras claves, su autor, los derechos de autor, etc. Se representa mediante un elemento `<meta>`:

```
<meta name="keywords" content="palabras claves">
```

```
<meta name="author" content="info del autor">
```

```
<meta name="description" content="descripción de la página">
```

```
<meta charset="codificación">
```

Las **palabras claves** (*keywords*) son una lista separada por comas de palabras o frases que indican los términos claves o más importantes que describen el contenido del documento. Ya no son tan importantes como antes. Antiguamente, los motores de búsqueda los utilizaban para indexar la página en sus bases de datos. Cada palabra o frase representa un término con los que el autor piensa que los usuarios podrían buscar su página en los buscadores.

Las palabras claves se indican mediante el elemento `<meta name="keywords">`.

Otros dos elementos de metadatos muy utilizados son `<meta name="author">` que contiene la información sobre el autor o *webmaster* de la página. Y `<meta name="description">` para describir brevemente el contenido de la página, utilizado también por los motores de búsqueda.

Otro metadato muy importante es el que indica el formato de codificación en que se encuentra almacenado el documento. Por lo general, es **utf-8**. Un **formato de codificación de caracteres** (*character encoding format*) es una manera de almacenar el documento en el disco. Digamos que cada

carácter tiene una secuencia de ceros y unos. Mediante el formato, indicamos cuál estamos usando para que así el navegador pueda interpretar los dígitos a los caracteres correctos. Se indica mediante `<meta charset>`. Ejemplo:

```
<meta charset="utf-8">
```

Por convenio y buenas prácticas, el elemento `<meta charset>` es el primero de la cabecera. Debe de encontrarse en los primeros 1024 bytes del documento.

Hay muchos más, pero éstos son los más importantes.

## Atributos

Los atributos más frecuentes del elemento `<head>` son:

Atributo	Tipo de datos	Descripción
id	Texto	Identificador del elemento.

## Cuerpo del documento

El **elemento de cuerpo** (*body element*) aloja el contenido del documento a mostrar a los usuarios. Se delimita mediante un elemento `<body>`.

## Modelo de contenido

El modelo básico de contenido del elemento `<body>` es el siguiente:

```
<body>
```

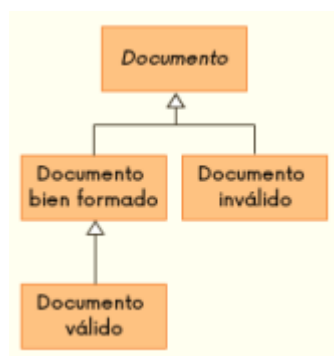
```
...  
</body>
```

El cuerpo del documento puede tener muchos elementos, muchos de los cuales iremos presentando a lo largo del curso como, por ejemplo, `<a>`, `<abbr>`, `<article>`, `<aside>`, `<audio>`, `<b>`, `<blockquote>`, `<br>`, `<button>`, `<canvas>`, `<cite>`, `<code>`, `<div>`, `<em>`, `<fieldset>`, `<figure>`, `<footer>`, `<form>`, `<h1>`, `<h2>`, `<h3>`, `<h4>`, `<h5>`, `<h6>`, `<header>`, `<hr>`, `<i>`, `<img>`, `<input>`, `<label>`, `<nav>`, `<object>`, `<p>`, `<pre>`, `<progress>`, `<table>`, `<textarea>`, `<time>`, `<u>`, etc.

## Validación de documentos

La **validación** (*validation*) es la operación mediante la cual se comprueba que un documento **HTML** sigue la especificación **HTML**, validando para ello que el documento está correctamente estructurado y que los elementos usados siguen la sintaxis esperada. La herramienta que realiza esta operación se conoce formalmente como **validador** (*validator*). Se puede utilizar validadores de línea de comandos, módulos incorporables a nuestras propias aplicaciones o el validador alojado en el **W3C**, <http://validator.w3.org>.

Atendiendo a cómo de conforme es el documento con respecto al estándar, se distingue entre documentos inválidos, bien formado y válidos.



Un **documento bien formado** (*well formed document*) es aquel que cumple las reglas básicas de la especificación **HTML**, principalmente, construir con una sintaxis correcta el documento. Mientras que un **documento válido** (*valid document*) es un documento bien formado que utiliza únicamente los elementos definidos en el estándar **HTML**. Finalmente, un **documento inválido** (*invalid document*) es aquel que no

está bien formado y, por ende, no cumple el estándar **HTML** mínimamente.