39 | HTML语言：DTD到底是什么？ | 极客时间

winter 2019-04-25

你好，我是 winter。今天，我们来聊一聊 HTML 语言。

我们平时写 HTML 语言，都习惯把关注点放到各种标签上，很少去深究它的语法。我想你应该会有模糊的感觉，HTML 这样的语言，跟 JavaScript 这样的语言会有一些本质的不同。

实际上，JavaScript 语言我们把它称为“编程语言”，它最大的特点是图灵完备的，我们大致可以理解为“包含了表达一切逻辑的能力”。像 HTML 这样的语言，我们称为“标记语言（mark up language）”，它是纯文本的一种升级，“标记”一词的概念来自：编辑审稿时使用不同颜色笔所做的“标记”。

在上世纪 80 年代，“富文本”的概念在计算机领域的热门，犹如如今的“AI”和“区块链”，而 Tim Berners-Lee 当时去设计 HTML，也并非是凭空造出来，他使用了当时已有的一种语言：SGML。

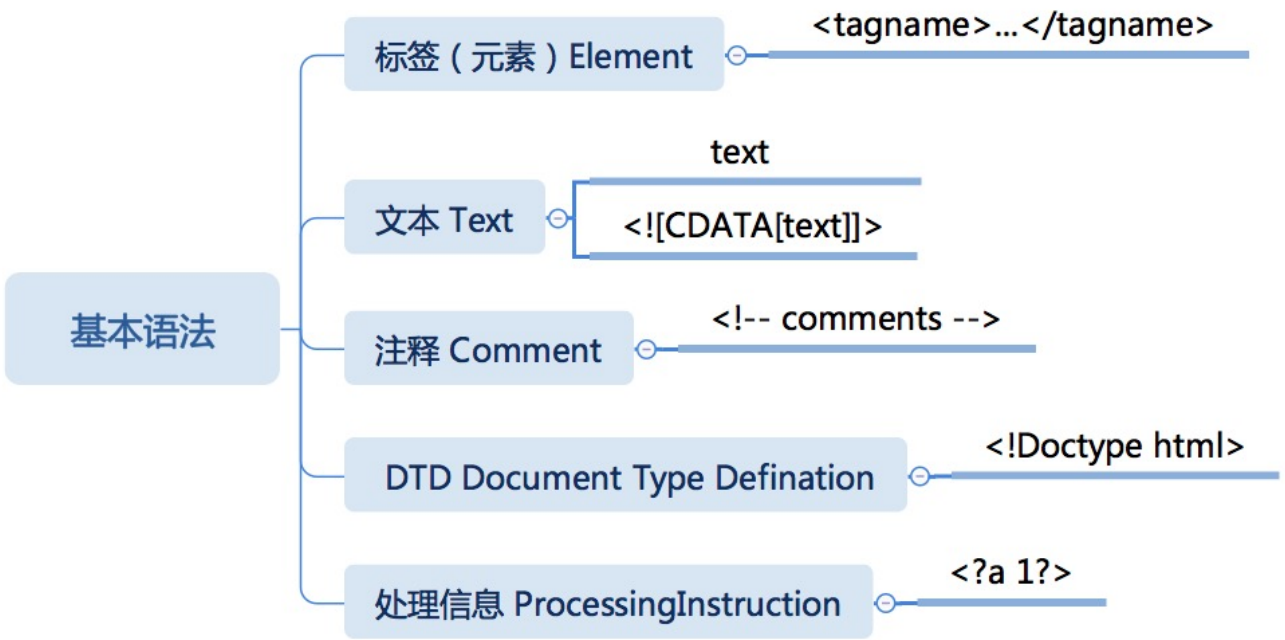
SGML 是一种古老的标记语言，可以追溯到 1969 年 IBM 公司所使用的技术，SGML 十分复杂，严格来说，HTML 是 SGML 中规定的一种格式，但是实际的浏览器没有任何一个是通过 SGML 引擎来解析 HTML 的。

今天的 HTML 仍然有 SGML 的不少影子，那么接下来我们就从 SGML 的一些特性来学习一下 HTML。这里我最想讲的是 SGML 留给 HTML 的重要的遗产：基本语法和 DTD。

基本语法

首先，HTML 作为 SGML 的子集，它遵循 SGML 的基本语法：包括标签、转义等。

SGML 还规定了一些特殊的节点类型，在我们之前的 DOM 课程中已经讲过几种节点类型，它们都有与之对应的 HTML 语法，我们这里复习一下：



这里我们从语法的角度，再逐个具体了解一下。

标签语法

标签语法产生元素，我们从语法的角度讲，就用“标签”这个术语，我们从运行时的角度讲，就用“元素”这个术语。

HTML 中，用于描述一个元素的标签分为开始标签、结束标签和自闭合标签。开始标签和自闭合标签中，又可以有属性。

- 开始标签：`<tagname>`
  - 带属性的开始标签：`<tagname attributename="attributevalue">`
- 结束标签：`</tagname>`
- 自闭合标签：`<tagname />`

HTML 中开始标签的标签名称只能使用英文字母。

这里需要重点讲一讲属性语法，属性可以使用单引号、双引号或者完全不用引号，这三种情况下，需要转义的部分都不太一样。

属性中可以使用文本实体（后文会介绍）来做转义，属性中，一定需要转义的有：

- 无引号属性：`<tab>` `<LF>` `<FF>` `<SPACE>` `&`五种字符
- 单引号属性：`&apos;` `&`两种字符
- 双引号属性：`&quot;` `&`两种字符

一般来说，灵活运用属性的形式，是不太用到文本实体转义的。

文本语法

在 HTML 中，规定了两种文本语法，一种是普通的文本节点，另一种是 CDATA 文本节点。

文本节点看似是普通的文本，但是，其中有两种字符是必须做转义的：`<` 和 `&`。

如果我们从某处拷贝了一段文本，里面包含了大量的 `<` 和 `&`，那么我们就有麻烦了，这时候，就轮到我们的 CDATA 节点出场了。

CDATA 也是一种文本，它存在的意义是语法上的意义：在 CDATA 节点内，不需要考虑多数的转义情况。

CDATA 内，只有字符组合]]>需要处理，这里不能使用转义，只能拆成两个 CDATA 节点。

注释语法

HTML 注释语法以<!--开头，以-->结尾，注释的内容非常自由，除了-->都没有问题。
如果注释的内容一定要出现 -->，我们可以拆成多个注释节点。

DTD 语法（文档类型定义）

SGML 的 DTD 语法十分复杂，但是对 HTML 来说，其实 DTD 的选项是有限的，浏览器在解析 DTD 时，把它当做几种字符串之一，关于 DTD，我在本篇文章的后面会详细讲解。

ProcessingInstruction 语法（处理信息）

ProcessingInstruction 多数情况下，是给机器看的。HTML 中规定了可以有 ProcessingInstruction，但是并没有规定它的具体内容，所以可以把它视为一种保留的扩展机制。对浏览器而言，ProcessingInstruction 的作用类似于注释。

ProcessingInstruction 包含两个部分，紧挨着第一个问号后，空格前的部分被称为“目标”，这个目标一般表示处理 ProcessingInstruction 的程序名。

剩余部分是它的文本信息，没有任何格式上的约定，完全由文档编写者和处理程序的编写者约定。

DTD

现在我们来讲一下 DTD，DTD 的全称是 Document Type Defination，也就是文档类型定义。SGML 用 DTD 来定义每一种文档类型，HTML 属于 SGML，在 HTML5 出现之前，HTML 都是使用符合 SGML 规定的 DTD。

如果你是一个上个时代走过来的前端，一定还记得 HTML4.01 有三种 DTD。分别是严格模式、过渡模式和 frameset 模式。

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
□复制代码
```

严格模式的 DTD 规定了 HTML4.01 中需要的标签。

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
□复制代码
```

过渡模式的 DTD 除了 html4.01，还包含了一些被贬斥的标签，这些标签已经不再推荐使用了，但是过渡模式中仍保留了它们。

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Frameset//EN" "http://www.w3.org/TR/html4/frameset.dtd">
□复制代码
```

frameset 结构的网页如今已经很少见到了，它使用 frameset 标签把几个网页组合到一起。

众所周知，HTML 中允许一些标签不闭合的用法，实际上这些都是符合 SGML 规定的，并且在 DTD 中规定好了的。但是，一些程序员喜欢严格遵守 XML 语法，保证标签闭合性，所以，HTML4.01 又规定了 XHTML 语法，同样有三个版本：

版本一

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
```

版本二

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "
http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
```