

21世纪网络工程丛书 开发设计系列

XML 技术精粹

胡海静 王育平 等编著

 机械工业出版社
China Machine Press

21 世纪网络工程丛书 开发设计系列

XML 技术精粹

胡海静 王育平 等编著



机械工业出版社

XML 全称 eXtensible Markup Language, 即可扩展标识语言。本书对 XML 技术、扩展样式语言 XSL、XML 链接语言 Xlink、XML 编辑工具 XMLwriter、同步多媒体合成语言 SMIL 等作了系统的介绍。本书从基本概念出发, 力求讲解简单明了、清晰顺畅, 并尽量使用典型的实例配合讲解, 深入浅出地讲解 XML 及其相关技术的基本原理、语法要点和使用方法。本书是学习 XML 及其相关技术的一本非常好的参考书。

图书在版编目 (CIP) 数据

XML 技术精粹 / 胡海静等编著. —北京: 机械工业出版社, 2001.11

(21 世纪网络工程丛书·开发设计系列)

ISBN 7-111-09513-8

I. X… II. 胡… III. 可扩充语言, XML IV. TP312

中国版本图书馆 CIP 数据核字 (2001) 第 075865 号

机械工业出版社 (北京市百万庄大街 22 号 邮政编码 100037)

责任编辑: 张秀恩

封面设计: 姚 毅 责任印制: 路 琳

中国建筑工业出版社密云印刷厂印刷·新华书店北京发行所发行

2002 年 1 月第 1 版第 1 次印刷

787mm×1092mm 1/16 16.5 印张·404 千字

0001—4000 册

定价: 26.00 元

ISBN 7-111-09513-8/TP·2193

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

本书购书热线电话 (010) 68993821、68326677-2527

[Http://www.machineinfo.gov.cn/book/](http://www.machineinfo.gov.cn/book/)

前 言

XML 的出现，在网络界又掀起了一场革命。XML 的一个主要目标市场是电子商务，有人称 XML 是“电子商务应用的基石”，还有人认为 XML 是“下一代 Web 语言”，甚至有人预言，XML 将全面取代 HTML 而成为网络中的天皇巨星。

XML 是 SGML（标准通用置标语言）的一个子集，全称是 eXtensible Markup Language，即可扩展标识语言。XML 不像 HTML 那样只有固定的形式，能够被用来使 SGML 在万维网上应用自如。XML 是一种类似于 HTML 的语言；XML 是被设计用来描述数据的语言；在 XML 中标志是没有预先定义的，必须要自定义需要的标志；XML 是能够自解释的语言；XML 使用 DTD（文档类型定义）来显示这些数据。XML 是用于标记电子文件使其具有结构性的标记语言。

XML 有很多功能，比如能够被用来分隔用户的数据和 HTML；能够被用来在 HTML 文档内部保存数据；XML 能够被用来作为一种交换信息的格式；XML 能够被用来将数据保存到文件或数据库中；等等。在实际应用中，XML 支持 EDI（电子数据交换），而且能够帮助 Java 技术。XML 的信息独立于平台之间。

本书对 XML 技术、扩展样式语言 XSL、XML 链接语言 Xlink、XML 编辑工具 XMLwriter、同步多媒体合成语言 SMIL 等作了系统的介绍。

本书从基本概念出发，力求讲解简单明了、清晰顺畅，并尽量使用典型的实例配合讲解，深入浅出地讲解 XML 及其相关技术的基本原理、语法要点和使用方法。本书是学习 XML 及其相关技术的一本非常好的参考书。

本书由胡海静、王育平组织编写，参加编写工作的还有张桂玲、李德奎、安琪、李正红、李正春、李春杰、郑春玥、汤艺佳、张青松、杨荣国等，参加材料收集和整理工作的有方帅、刘涛、宁翔、吴磊、李政斌、程岩、肖彤、葛正春、汪乃群、孙晓芳、潘庆锋、张学静、唐学军、高洪亮等。整个工作离不开大家的辛勤工作，是大家集体智慧的结晶。

作 者

2001 年 7 月

目 录

前言

第 1 章 XML 简介.....	1
1.1 置标语言.....	1
1.2 SGML	2
1.3 HTML	2
1.4 什么是 XML?.....	3
1.4.1 XML 简介	3
1.4.2 XML 与 HTML 的区别.....	3
1.4.3 来源.....	5
1.4.4 XML 三要素	5
1.4.5 XML 的设计目标	7
1.4.6 XML、SGML、HTML 之间的关系.....	7
1.5 XML 标准.....	7
1.6 XML 术语.....	8
1.7 XML 文件.....	16
1.7.1 规范的 XML 文件	16
1.7.2 字符.....	16
1.7.3 通用语法成分	16
1.7.4 字符数据和标记	17
1.8 XML 的应用.....	17
第 2 章 从 HTML 到 XML	29
2.1 书写格式良好的 HTML	29
2.1.1 标记必须是封闭的	30
2.1.2 禁止交叉标记	30
2.1.3 字符匹配	30
2.1.4 引用属性	30
2.1.5 使用单根	30
2.1.6 避免脚本块	31
2.2 页面标记.....	31
2.2.1 HTML 文件结构.....	31
2.2.2 标题.....	35
2.2.3 网页颜色.....	36
2.2.4 背景图片.....	36
2.2.5 页面空白.....	37
2.2.6 标题.....	37
2.2.7 段落.....	38

2.2.8 换行	38
2.2.9 水平线	39
2.2.10 边框和阴影	40
2.3 字体	41
2.3.1 字体大小	41
2.3.2 字体类型	42
2.3.3 字体效果	43
2.3.4 逻辑字体	44
2.3.5 预格式化文本	45
2.3.6 文字对齐方式	46
2.4 表格	46
2.4.1 基本语法	47
2.4.2 跨多行、多列的表格	48
2.4.3 表格尺寸设置	48
2.4.4 表格内容的对齐方式	50
2.4.5 表格在页面中的对齐方式	52
2.4.6 表格的标题	53
2.4.7 表格的颜色和图片设置	53
2.5 图片标记	54
2.5.1 基本语法	54
2.5.2 图片与文本对齐方式	54
2.5.3 图片在页面中的对齐方式	56
2.5.4 图片的间距	56
2.5.5 边框	56
2.6 项目符号和编号	56
2.6.1 无格式项目列表	56
2.6.2 编号列表	58
2.6.3 目录列表	59
2.6.4 定义列表	59
2.6.5 菜单列表	60
2.6.6 紧凑布局	60
2.7 超链接	61
2.7.1 外部链接	62
2.7.2 内部链接	62
2.7.3 框架的链接	62
2.7.4 图像映射	63
2.8 框架	63
2.8.1 基本指令	64
2.8.2 指定框架内容	64
2.9 表单	65
2.9.1 单行文本框	65

2.9.2 滚动文本框	65
2.9.3 复选框	65
2.9.4 单选按钮	66
2.9.5 下拉菜单	66
2.9.6 按钮	66
2.9.7 综合举例	66
2.10 多媒体页面	69
2.10.1 基本语法	69
2.10.2 背景音乐	69
2.10.3 视频剪辑	69
2.11 字幕	70
2.11.1 基本构成	70
2.11.2 文字移动的方向	70
2.11.3 文字的表现方式	70
2.11.4 重复设置	70
2.11.5 速度	71
2.11.6 对齐方式	71
2.11.7 设置背景颜色	71
2.11.8 字幕大小	71
2.12 XML 数据岛	71
2.12.1 XML 数据岛简介	71
2.12.2 类似于文档对象访问一个 XML 数据岛	72
2.12.3 使用 XSL 转换数据岛内容	73
2.13 用 HTML 显示 XML 数据	73
2.14 XHTML	74
第 3 章 XML 技术	76
3.1 从与 HTML 的比较中看 XML 的结构	76
3.2 用 XML 声明作为开头	79
3.2.1 XML 声明的形式	79
3.2.2 version 属性	80
3.2.3 standalone 属性	80
3.2.4 encoding 属性	81
3.3 区分大小写	82
3.4 逻辑结构	82
3.4.1 逻辑结构概述	82
3.4.2 起始标签	82
3.4.3 结束标签	83
3.4.4 空元素标签	83
3.5 文档类型定义	84
3.5.1 内部 DTD	86
3.5.2 外部 DTD	86

3.5.3 内部与外部文档类型定义的复合	89
3.6 元素类型声明	90
3.6.1 元素	90
3.6.2 字符数据	90
3.6.3 标记	91
3.6.4 名字和名字单词符号	93
3.6.5 元素类型声明简介	93
3.6.6 元素型内容	94
3.6.7 子元素声明	94
3.6.8 声明多个子元素（按顺序）	95
3.6.9 声明可选的子元素	95
3.6.10 声明零个或多个子元素	96
3.6.11 声明一个或多个子元素	96
3.6.12 子元素的选择	97
3.6.13 嵌套元素	97
3.6.14 混合内容	99
3.6.15 实例解析 1	100
3.6.16 实例解析 2	104
3.7 属性列表声明	106
3.7.1 属性-值对的规范化	106
3.7.2 属性列表声明	107
3.7.3 String 属性类型	108
3.7.4 Tokenized 属性类型	108
3.7.5 enumerated 属性类型	110
3.7.6 属性默认值	111
3.7.7 实例解析	112
3.8 预定义属性	120
3.8.1 语言标识	120
3.8.2 空白	121
3.8.3 综合举例	122
3.9 实体类型声明	122
3.9.1 实体	122
3.9.2 字符和实体引用	123
3.9.3 普通实体声明	124
3.9.4 在实体内应用实体	126
3.9.5 预定义实体	127
3.9.6 参数实体声明	127
3.10 符号声明	128
3.11 CDATA	129
3.12 处理指令	130
3.13 条件部分	131

3.14	注释.....	132
3.15	文字与图形.....	133
3.16	Schema 简介.....	136
3.17	实例.....	137
第 4 章	扩展样式语言 XSL.....	143
4.1	XSL 概述.....	143
4.2	层叠样式表 CSS.....	144
4.2.1	CSS 的主要用途.....	144
4.2.2	CSS 的特点.....	145
4.2.3	简单的 CSS 层叠样式表的例子.....	146
4.2.4	选择符.....	147
4.2.5	将样式表植入 HTML 文件中.....	148
4.2.6	将外部样式表链接到 HTML 文件上.....	149
4.2.7	将外部样式表输入到 HTML 文件中.....	150
4.2.8	将样式表加入到 HTML 文件行中.....	151
4.3	XSL 的组成和设计原则.....	152
4.3.1	XSL 的组成.....	152
4.3.2	XSL 的设计原则.....	152
4.4	XSL 样式表概述.....	153
4.4.1	XSL 样式表.....	153
4.4.2	CSS 与 XSL 比较.....	154
4.5	样式表结构.....	155
4.5.1	样式表元素.....	155
4.5.2	作为样式表的文字结果元素.....	156
4.5.3	向前兼容性处理.....	157
4.5.4	组合样式表.....	159
4.5.5	嵌套样式表.....	159
4.6	条件模板.....	160
4.6.1	XSL 条件性选择.....	160
4.6.2	XSL 条件性假设.....	165
4.7	XSL 过滤问题.....	168
4.8	客户端 XSL.....	169
4.9	服务器端 XSL.....	171
4.10	XSL 排序.....	173
4.11	XSL 转换.....	177
第 5 章	XML 链接语言 XLink.....	179
5.1	XLink 标记词汇表.....	179
5.2	简单链接.....	180
5.3	链接角色.....	181
5.4	链接行为.....	182
5.4.1	show 属性.....	183

5.4.2	actuate (横向激发) 属性	184
5.4.3	behavior 属性	184
5.5	扩展链接	185
5.5.1	locator 元素	185
5.5.2	链接组	188
第 6 章	XML 工具软件 XMLwriter	190
6.1	XMLwriter 简介	190
6.2	XMLwriter 的特色	190
6.3	XMLwriter 的安装	193
6.4	XMLwriter 窗口	197
6.5	File (文件) 菜单	198
6.6	Edit (编辑) 菜单	205
6.7	View (查看) 菜单	209
6.8	Project (项目) 菜单	211
6.9	Tools (工具) 菜单	212
6.10	Window (窗口) 菜单	220
6.11	Help (帮助) 菜单	222
6.12	主工具栏	226
6.13	用户工具栏	228
6.14	键盘快捷键	229
6.15	参数设置	230
6.15.1	General 选项卡	230
6.15.2	Editor 选项卡	231
6.15.3	Directories 选项卡	232
6.15.4	.File Types 选项卡	233
6.15.5	Tools 选项卡	234
6.15.6	Format 选项卡	236
6.15.7	参数设置实例	236
第 7 章	同步多媒体合成语言 SMIL	238
7.1	smil 元素	239
7.1.1	body 元素	239
7.1.2	head 元素	244
7.2	layout 元素	244
7.3	meta 元素	245
7.4	region 元素	245
7.5	SMIL DTD	247

第1章 XML 简介

1996年11月,从波士顿“SGML”世界年会上传出一条新闻,与会代表一致同意将终结目前的 Internet HomePage 撰写标准“HTML”。取而代之的是采用最新电子文件格式化通用标准的可延伸的标识语言“XML”。

XML 是 eXtensible Markup Language 的缩写,译成中文是“扩展标识语言”。XML 的出现虽然只有短短的两年时间,但是非常令人吃惊,众多的 Web 开发人员却对其青睐有加,认为 XML 将成为 HTML 的后继者,在不远的将来会成为整个 Web 的统治力量。

XML 可以改善传统网页设计语言 HTML 的描述方式,未来将取代 HTML 成为新一代的网页标识语言,这一转变将革命性地改变以 Web 为基础的电子商业应用形态。利用 XML 所具有的可延伸性以及自我描述性,Web 文件可以在企业间的应用程序中自动传输、处理及储存,不同厂商的电子商品可以在一个使用者的界面上同时展现,资讯的搜寻变得更为精确快速,不同系统间畅通无阻,而且中小企业也可以轻易享受电子商务的好处。

全球最大的软件公司——微软公司已经加入 XML 支持者的行列,该公司代号为 Shiloh 的下一代数据库软件将支持 XML;已经将 XML 集成到 Microsoft Office 2000 中,这种软件以 XML 格式存储 Word、Excel、PowerPoint 以及其他类型的文件;最新发布的 Internet Explorer 5 也广泛使用了 XML。其公司的 XML 产品主管 Dave Wascha 说,“XML 确实影响了所有的软件产品系列,我们仍然在观察它所能带来的好处。”另外, Netscape 新版本也肯定会支持 XML。其他公司,包括 Adobe、ArborText、Sun 和 Xerox 也宣布支持 XML。

XML 将给 Web 注入一股新的生命。它带来的应该是一个全新的视野,一种对于整个电子商务架构在观念上的颠覆。到底 XML 有哪些特别之处,人们缘何对它情有独钟,我们接下来就对 XML 做一个简单的介绍。

1.1 置标语言

XML 是置标语言 (Markup Language) 的成员,所以我们必须首先知道什么是置标语言,而这又要求我们知道什么是置标。所谓置标,顾名思义,是根据数据自身信息进行编码的方法。而如果在语言中为了标识某些信息而加入一些标记,这种语言就被称为置标语言。具体地说,这里所指的置标语言,实际上是一种为了计算机处理而设计的置标语言,其中所用到的标记,往往使用代表一定含义的文字或数字表示。通常根据需要定义一套标记,然后将这套标记添加到语言中去,就变成了置标语言。

置标语言不同于一般的控制程序设计语言,基本上可以被视为一种数据流的文档结构描述语言。在计算机处理过程中,置标语言的标记既可以作为数据,也可以作为控制语句来使用。在计算机程序语言中,一般把一种语言的定义整体称为某某程序设计语言,而把用这种语言编写的文字段落,称为某某语言程序。置标语言的定义整体称为某某置标语言,而把用这种置标语言编写的文字段落,称为某某语言文档。

当需要通过标记将有用的信息告知一组用户时,必须有一个标准,用来描述什么是有

效的标记。而在 HTML 中，标记就是所谓的标签。另外，还要有一个标准描述每个标记的具体含义。而在 HTML 中，每一个标签都表明了一种显示的格式。

同样，“置标”的应用在计算机世界中也甚为广泛。文字编辑器借助置标来定义格式与外观，通信程序依靠置标来理解线路上所传输信息的语意，数据库通过置标来将数据字段与一定的含义相连，并表明字段之间的关系，多媒体应用中置标则被用来标识什么是图像和声音的源数据。当这些数据被传送给计算机或应用程序时，它必须自身携带一些必要的信息，以表明这些数据的含义，以及接收者应该如何处理这些数据。

在简单介绍了置标语言之后，我们开始介绍 SGML。

1.2 SGML

SGML 全称 Standard Generalized Markup Language，即标准通用置标语言。在 Web 还没有出现之前，SGML 就已经存在了，正如它的名称指出的那样，SGML 是用标记来描述文档的通用语言，它包含一系列的文档类型定义（Document Type Definition，简称 DTD），DTD 中定义了标记的含义，因而 SGML 的语法是可以扩展的。

SGML 是一种置标语言，实际上它是一种通用的文档结构描述置标语言，主要用来定义文献模型的逻辑和物理类结构。SGML 是 ISO 组织于 1986 年发布的 ISO 8879 国际标准。

一个 SGML 语言程序，要由三部分组成，即语法定义、文件类型定义（DTD）和文件实例。语法定义，定义了文件类型定义和文件实例的语法结构；文件类型定义，定义了文件实例的结构和组成结构的元素类型。文件实例是 SGML 语言程序的主体部分。

SGML 的实际使用中，每一个特定的 DTD 都定义了一类文件。例如，所有的数学文档都可以使用同一个 DTD。因此，习惯上把具有某一特定 DTD 的 SGML 语言，称为某某置标语言。例如用于国际互联网的 HTML 语言。这样 SGML 就成为那些派生语言的元语言。

1.3 HTML

HTML 的全称是“HyperText Markup Language”（超文本标识语言）。它是 SGML 在网络上的一个特殊应用。HTML 是一种用于全球广域网上文档中的标准标记语言。

HTML 的优点是易学和通用，但是它的链接容易断开，格式也很有限。经过一段时间的发展，人们已经对它进行了许多修改和扩充，增加了表格、框架、脚本语言等内容。然而，HTML 依旧是 HTML，缺点依然明显。对于 HTML 的缺点的具体内容，我们在下文还要讨论，这里不再赘述。

现在，几种最新的技术将有望解决 HTML 固有的问题，比如 XML（eXtensible Markup Language，可扩展的标识语言）、CSS（Cascading Style Sheets，层叠样式表）和动态 HTML（DHTML）等。它们的优点表现在方方面面。比如 XML 更加侧重于组织和查找数据；CSS 侧重于 Web 页的继承性和表现；DHTML 则侧重于 Web 内容的动态表现。

在这几种新技术中，XML 是 World Wide Web 联合组织（W3C）认可的标准，它代表着人们编写标准 HTML 习惯的最大转变。XML 定义文档的结构而不是定义浏览器应该如何显示文档，这将给 Web 开发商提供更多的灵活性。XML 改变了浏览器的显示、组织和搜索信息的方法。它甚至可以使链接不会再断开。当然，HTML 与 XML 之间并不是竞争关系，

而是相互补充的关系。各种浏览器都将能够处理这两种语言,而且将来的 HTML 标准可能允许在同一文档中混合使用 HTML 和 XML。

DHTML 着眼于以更少而又更快的页面下载更丰富的图形和数据。CSS 能够让用户创建继承其他页面属性的页面。目前, CSS 是用 HTML 实现的。而 XML 也会有使用可扩展样式语言的样式表,即 XSL (eXtensible Style Language)。

W3C 的文档对象模式(DOM),即现在作为 DHTML 规范一部分的建议草案,允许 HTML 和 XML 脚本和其他程序在程序控制下访问结构化数据。DOM 还为页面设计和布局增加了对象方向。例如, HTML 的各个元素表现为属性和方法的对象或集合。开发商可以用 DOM 和脚本语言,如 JavaScript 或 VB Script 来控制 DOM 而实现动态的样式、内容和定位。脚本可以控制定位属性,从而在 HTML 页面上生成动画。

1.4 什么是 XML?

1.4.1 XML 简介

XML 全称是 eXtensible Markup Language (可扩展标识语言)。之所以称之为可扩展,是因为它不像 HTML 那样只有固定的形式。它是被用来使 SGML 在万维网上能应用自如。XML 是用于标记电子文件使其具有结构性的标记语言。

XML 是一种类似于 HTML 的语言; XML 是被设计用来描述数据的语言;在 XML 中标志是没有预先定义的,必须要自定义需要的标志; XML 是能够自解释 (self describing) 的语言; XML 使用 DTD (Document Type Definition, 文档类型定义) 来显示这些数据。XML 是用于标记电子文件使其具有结构性的标记语言。

总结一下,可以这样认为:

可扩展标识语言,缩写为 XML,描述了一类称为 XML 文件的数据对象,同时也部分地描述了处理这些数据对象的计算机程序的动作。XML 是 SGML (标准通用标识语言, ISO 8879) 针对应用的一个子集,或者说是 SGML 的一种受限形式。根据定义, XML 文件是合乎规范的 SGML 文件。

XML 文件由称为实体的存储单元组成,实体可以包含已析数据或未析数据。已析数据由字符组成,其中一些字符组成字符数据,另一些字符组成标记。标记中包含了对文件存储格式 (storage layout) 和逻辑结构的描述。XML 提供了一种机制用于约束存储格式和逻辑结构。

称为 XML 处理器的软件模块用于读取 XML 文件,存取其中的内容和结构。XML 处理器被设想为是为另一个称为应用的模块做处理。本规范从 XML 处理器应如何读取 XML 数据以及应向应用提供哪些信息的这两个方面,描述了要求 XML 处理器做出的动作。

图 1-1 是 XML 格式的网页实例。

1.4.2 XML 与 HTML 的区别

严格说来, XML 本身并不是一个独立的、预定义、单一的标识语言,它是一种元语言 (meta-language)。它是用来描述其他语言的语言,可以被用来定义任何一种新的标识语言。像 HTML 之类的传统标识语言,是用来定义某一类文件的格式,以便于展示或列印;而 XML

则是可以用来创造类别文件的格式定义,也就是在 XML 之中可以创造出很多不同的标识语言,用来定义不同的文件类别。XML 允许用户设计自己的标识,而 HTML 只是一种预定义的标识语言。HTML 只是在一类特定的文件中定义一种描述信息的方法。而 XML 允许用户在不同的文件中定义自己设计的标识语言。

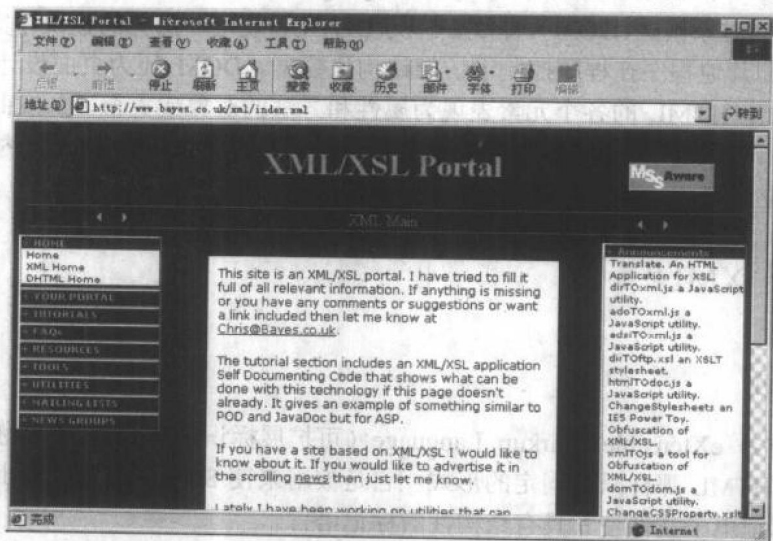


图 1-1 使用 XML 格式的网页实例

XML 继承了 SGML 的可扩展性、结构性及可校验性,与 HTML 语言相比,区别主要在三方面:

- (1) 可扩展性方面: HTML 不允许用户自行定义自己的标识或属性,而在 XML 中,用户能够根据需要,自行定义新的标识及属性名,以便更好地从语义上修饰数据。
- (2) 结构性方面: HTML 不支持深层的结构描述,XML 的文件结构嵌套可以复杂到任意程度,能表示面向对象的等级层次。
- (3) 可校验性方面: HTML 没有提供规范文件以支持应用软件对 HTML 文件进行结构校验;而 XML 文件可以包括一个语法描述,使应用程序可以对此文件进行结构确认。

需要指出的一点是,HTML 与 XML 两者相互之间并不是竞争关系,而是互相补充的关系。各种浏览器都将能够处理这两种语言,而且将来的 HTML 标准可能允许在同一文档中混合使用 HTML 和 XML。

确切地说,使用 XML,就可以以容易而一致的方式格式化和传送数据。实际上,将 XML 设计成 SGML 的一个子集,它是简化的,而且目标是面向 Web。HTML 是 SGML 的另外一个子集。事实上,如果很熟悉 HTML,那么再来学习 XML 就会很省劲。XML 是根据 SGML 所制定出来的,所以 XML 可看成是 SGML 的一个子集,且其设计目的是能广泛灵活地在 World Wide Web 上被应用。

HTML 和 XML 语言中的标记都是用尖括号“<”和“>”括起来的文本字符串,而且对读入 HTML 或 XML 文档的应用程序来说,它们是伪指令。

XML 描述的是数据本身,而 HTML 描述了如何显示 Web 页中的数据。换句话说,作为一种数据描述语言的 XML,使用频率会更高。它允许将数据组织成数据结构,甚至是很

复杂的数据结构；用户可以按照自己的需要剪裁数据；用户可以创建自己的标记，这就允许用户在 XML 文档中以自己的方式组织数据。

可以这样说，XML 不是 HTML 的替代品，XML 和 HTML 的用途是不同的。具体体现在如下几个方面：

(1) XML 是被设计用来描述数据的；而 HTML 是被设计用来显示数据的。

(2) XML 是关于数据的语言，是关于如何描述信息的；而 HTML 是关于如何显示信息的。

(3) XML 是可扩展的，允许用户自定义自己的标记和自己的文档结构；在 HTML 中所有的标记和文档结构都是预先定义的，用户只有权利使用标准的 HTML 标记。

(4) XML 是 HTML 的一种补充，XML 不是 HTML 的一种替代品。可以使用 XML 来描述网站所需数据和网站的结构，而使用 HTML 来格式化和显示这些数据。

XML 有很多功能，比如被用来分隔用户的数据和 HTML；被用来在 HTML 文档内部保存数据；被用来作为一种交换信息的格式；被用来将数据保存到文件或数据库中；等等。

XML 能够被用来分隔用户的数据和 HTML。HTML 页面是用来显示数据的，所以数据一般都保存在 HTML 页面内部。而使用 XML，这些数据能够被保存到一个单独的 XML 文件中。这种方法就能够让用户集中来使用 HTML 格式化和显示，并确保对数据的任何改动都不必重新修改 HTML 代码。

XML 也能够被用来在 HTML 文档内部保存数据。XML 的数据也可以被当成数据岛(Data Islands)保存在 HTML 页面中，这样用户仍然可以使用 HTML 来格式化和显示数据。有关数据岛的内容，请用户参考本书有关章节。

XML 能够被用来作为一种交换信息的格式。在现实世界里，计算机系统和数据库中保存的数据格式是不兼容的，所以在很多时候，开发者面临的问题是如何通过 Internet 来交换这些数据，如果把这些数据转换成 XML 能够大大降低工作的复杂性。

XML 能够被用来将数据保存到文件或数据库中，运用程序可以用来读写和显示这些数据信息。

1.4.3 来源

XML 是 SGML (Standard Generalized Markup Language, 通用标记语言) 的一个精简子集，主要用于 Internet 的信息传递描述的标记语言。SGML 并没有像当初的设计者所预想的那样为网络所广泛采纳，一个原因是它非常复杂，XML 保留了 SGML 中的优秀的特性：可扩展、文本描述、面向 Internet 的设计，并且具有简单性。XML 是由 W3C XML 工作组维护和升级，它是一个开放的标准。

1.4.4 XML 三要素

虽然 XML 有许多部分，但是只需要了解其中的三个就可以了解它是怎样工作的。它们是文档类型定义 (Document Type Definition, 简称为 DTD)，也就是 XML 的布局语言；可扩展的样式语言 (eXtensible Style Language, 简称为 XSL)，也就是 XML 的样式表语言；以及可扩展链接语言 (eXtensible Link Language, 简称为 XLL)。

首先，我们看一下文档类型定义 (DTD)。DTD 规定文档的逻辑结构。它可以定义文档的语法，而文档的语法反过来能够让 XML 语法分析程序确认某张页面标记使用的合法性。

DTD 定义页面的元素、元素的属性以及元素和属性之间的关系。例如, DTD 能够规定某个表项只能在某个列表中使用。

理想的定义应面向描述与应用程序有关的数据结构而不是如何显示数据。换句话说, 应把一个元素定义为一个标题行, 然后让样式表和脚本定义如何显示标题行。

DTD 不是强制性的。对于简单应用程序来说, 开发商不需要建立自己的 DTD (因为没有意义)。可以使用预先定义的公共 DTD, 或者根本就不使用。即使某个文档已经有了 DTD, 只要文档是组织良好的, 语法分析程序也可以不对照 DTD 来检验文档的合法性。服务器可能已经执行了检查, 所以检验的时间和带宽将得以节省。

接下来, 我们介绍一下可扩展的样式语言 (XSL)。XSL 是用于规定 XML 文档样式的语言。XSL 能使 Web 浏览器改变文档的表示法, 例如数据的显示顺序改变, 就不需要再与服务器进行交互通信。通过变换样式表, 同一个文档可以显示得更大, 或者经过叠折只显示外面的一层, 或者变为打印格式。可以设想一个适合用户学习特点的技术手册, 它为初学者和更高一级的用户提供不同的样式, 而且所有的样式都是根据同样的文本产生的。

XSL 凭借其可扩展性能够控制无穷无尽的标记, 而控制每个标记的方式也是无穷尽的。这就给 Web 提供了高级的布局特性。例如旋转的文本、多列和独立区域。它支持国际书写格式, 可以在一页上混合使用从左至右、从右至左和从上至下的书写格式。

正如 XML 介于 HTML 和 SGML 之间一样, XSL 标准介于 CSS 和 SGML 的文档样式语义和规范语言 (DSSSL) 之间。DSSSL (Document Style Semantics and Specification Language) 定义格式化对象的全特征模式。由于 DSSSL 使用框架语法, 而且是很复杂的, 所以 DSSSL 未能得到推广应用。相比较而言, CSS 使用简单的模型 (例如它不能将元素再排序)。XSL 支持 DSSSL 流对象和 CSS 对象, 使用 XML 语法, 并对复杂的任务提供进入 ECMAScript 的转码, 而且允许扩展。实现从 CSS 到 XSL 的映射是可能的, 内容开发商无需学习这种语言的全部。作为一种技术预展, 微软最近发布了两种 XSL 处理器: 一个是可以从 XML 文档和 XSL 样式层产生 HTML 输出的命令行应用程序, 另一个是一种 ActiveX 控件, 用于在浏览器中显示 XML。微软的这种 XSL 处理器适合在 Windows95 和 WindowsNT 环境下通过 Internet Explorer 4.0 浏览器使用。

最后, 我们再来谈一谈可扩展链接语言 (XLL)。XLL 将支持目前 Web 上已有的简单链接, 并且将进一步扩展链接, 包括结束死链接的间接链接以及可以从服务器中仅查询某个元素的相关部分的连接符 “|”。

XML 工作组组长 Jon Bosak 说, 所谓超文本标记语言 HTML 只执行历来与超文本系统概念相关的极少功能, 只支持最简单的链接形式, 即指向硬编码位置的单向链接, 这与 XML 相比有很大的差别。在为 XML 所设想的真正的超文本系统中, 所有典型的超文本链接机制都将得到支持, 包括:

- 与位置无关的命名。
- 双向链接。
- 可以在文档外规定和管理的链接。
- n 元超链接 (如环路、多个窗口)。
- 集合链接 (多来源)。
- Transclusion (链接目标文档是链接源文档的一部分)。
- 链接属性 (链接类型)。

这些可以通过目前尚在开发中的 XLL 来实现。由于 XML 以 SGML 为基础, XSL 以 DSSSL 为基础, 所以 XLL 基本上是 Hytime (超媒体/基于时间的结构语言, ISO10744) 的一个子集。它还遵循文本编码倡议 (Text Encoding Initiative) 规定的链接概念。

1.4.5 XML 的设计目标

根据可扩展标记语言 (XML) 1.0 的 W3C 建议 (1998 年 2 月 10 日), XML 的设计目标如下:

- XML 应该可以直接用于因特网 (Internet)。
- XML 应该支持大量不同的应用。
- XML 应该与 SGML 兼容。
- 处理 XML 文件的程序应该容易编写。
- XML 中的可选项应无条件地保持最少, 理想状况下应该为 0 个。
- XML 文件应该是人可以直接阅读的, 应该是条理清楚的。
- XML 的设计应快速完成。
- XML 的设计应该是形式化的, 简洁的。
- XML 文件应易于创建。
- XML 标记的简洁性是最后考虑的目标。

1.4.6 XML、SGML、HTML 之间的关系

SGML 是一个语言模板, 它可以描述其他具有具体语义的语言。HTML 是 SGML 的一个实例, 它的 DTD 作为标准被固定下来, 因此, HTML 不能作为定义其他置标语言的元语言。它的语言遵照 SGML 规范, 并且定义其中每个标记的具体含义。XML 是 SGML 的一个精简子集, 它也是一个语言模板。严格地讲, XML 还是 SGML。与 HTML 不同的是, XML 有 DTD, 因而也可以作为元语言来定义其他文件系统 (置标语言)。

1.5 XML 标准

W3C 的 XML 工作组正致力于确保 XML 的互用性, 以及为开发人员、处理人员和不同系统和浏览器的使用者提供支持, 并进一步发展 XML 的标准。

XML 包括的标准如下:

- 可扩展标识语言 (XML) 标准: 这是 W3C 正式批准的, 该标准是稳定的, 可用于 Web 和工具的开发。
- XML 名域 (namespace) 标准: 这用来描述名域的句法, 支持能识别名域的 XML 解析器。
- 文档对象模式 (DOM) 标准: 这为给结构化的数据编写脚本提供了标准, 这样开发人员就能够同计算机在基于 XML 的数据上进行交互作用。
- 可扩展类型语言 (XSL) 标准: 这是一个工作草案。XSL 有两个模块, 即 XSL 转换语言和 XSL 格式化对象。转换语言可用来转换 XML 以满足显示要求。由于 XSL 的两部分是模块, 转换语言能够独立地用来进行多用途的转换, 包括把 XML 转换成结构完整的 HTML。CSS 可应用于结构简单的 XML 数据, 但不能以不同于信息如何传递来的方式显示