高等院校“十二五”规划精品教材

XML
Jishu jiaocheng

XML技术教程

主　编　王占中
副主编　芦娜　许研　姬利娜

西南财经大学出版社
Southwestern University of Finance & Economics Press

高等院校“十二五”规划精品教材

XML
Jishu jiaocheng

XML技术教程

主　编　王占中
副主编　芦娜　许研　姬利娜

西南财经大学出版社
Southwestern University of Finance & Economics Press

图书在版编目(CIP)数据

XML 技术教程/王占中主编. —成都:西南财经大学出版社,2011. 12
ISBN 978-7-5504-0497-7

Ⅰ. ①X… Ⅱ. ①王… Ⅲ. ①可扩充语言,XML—程序设计—高等学校—教材 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字(2011)第 258101 号

XML 技术教程

主　编:王占中

副主编:芦　娜　许　研　姬利娜

责任编辑:李霞湘

助理编辑:庞光伟　李玉华

封面设计:杨红鹰

责任印制:封俊川

出版发行	西南财经大学出版社(四川省成都市光华村街 55 号)
网　　址	http://www.bookcj.com
电子邮件	bookcj@foxmail.com
邮政编码	610074
电　　话	028-87353785　87352368
印　　刷	四川森林印务有限责任公司
成品尺寸	185mm×260mm
印　　张	15
字　　数	340 千字
版　　次	2011 年 12 月第 1 版
印　　次	2011 年 12 月第 1 次印刷
书　　号	ISBN 978-7-5504-0497-7
定　　价	29.00 元

前 言

近二十年来，对于 Internet 的飞速发展，HTML 可以说居功至伟。HTML 使贩夫走卒到大学教授都能方便地使用 Internet。但是随着 Internet 的发展，海量数据的出现对网络技术提出了新的要求——从浩如烟海的数据中准确地提取出有用的数据。这就要求对数据准确地定义和表达，而这是 HTML 所不能胜任的。HTML 只是方便人们浏览网页，不具备数据定义的能力。作为 HTML 的补充，XML 应运而生。

XML 是由万维网联盟（W3C）定义的一种标记语言，是表示结构化数据的行业标准。利用 XML，各个行业或组织可以按照自己的需要定义数据标准，从而使得 Internet 上的数据相互交流更加方便。利用 XML，可以通过编程自动地处理网上数据，而不只是浏览网页。利用 XML，可以将不同来源的数据进行无缝集成，这成为 Web Service 和电子商务的支点。

本书结合实例详细讲解了 XML 的基本知识与应用。全书共分 9 章。第 1 章主要对 XML 作了简单的介绍，使读者对 XML 的来源、概念、相关技术有一个整体的了解。第 2 章是 XML 的语法部分，介绍了诸如 XML 的文件结构、元素规则、属性规则、名称空间、处理指令等内容，使读者能够创建格式良好的 XML 文档。第 3、4 章是关于有效性检测的内容，即是对 XML 文档数据结构进行约束的技术。其中第 3 章介绍文档类型定义（DTD）。DTD 以特定的方式说明元素约束、属性约束以及实体的声明与引用。第 4 章介绍 Schema 技术。Schema 所起的作用和 DTD 一样，其优势在于它是一个规范的 XML 文档且在完成元素和属性约束的同时极大地丰富了数据类型，其次 Schema 对名称空间有较好的支持。第 5、6 章是关于显示技术的内容，包括 CSS 和 XSL。CSS 是先于 XML 出现的技术，规定文档中各数据单元在网页中的显示样式，使网页更加生动、引人入胜；XSL 是用 XML 文档书写的样式语言，它能根据用户的需要将 XML 的树状结构转换成其他的树状结构，满足各种不同的显示需要。第 7、8 章是 XML 解析技术，解析技术是程序化处理 XML 数据的技术保障。第 7 章介绍的 DOM 接口技术是 W3C 定义的接口标准，这使对 XML 的解析规范化。DOM 的基本思想是将 XML 中各种内容映射成内存中的一棵树，可以随机地对数据进行读取、添加、修改与删除处理。第 8 章介绍 SAX 接口技术，它是将读取 XML 文档时遇到的各种数据都映射成事件，事件处理器捕

获事件进行相应处理。SAX 接口编程只能顺序处理 XML 且只能读取不能编辑，编程的重点在事件处理器的构建。本书的最后一章介绍 XML 与其他数据文件的转换。XML 是作为 Internet 上数据交换的标准出现的，欲融入到各式各样的处理系统，必须解决和现有的数据文件的转换问题。这里我们选取数据库和 Excel 表作了与 XML 的转换尝试。XML 和数据库表的转换见诸于其他的资料，XML 与 Excel 表的转换是本书的一个特色。

本书分工情况如下：王占中撰写第 1、7、8、9 章并统筹全书的整体构架；芦娜撰写第 5、6 章；许研撰写第 3、4 章；姬利娜撰写第 2 章。另外，李阳为本书提出一些合理建议并完成了部分章节的习题。

本书是我们教授 XML 的总结，如果能对读者学习 XML 有些许帮助，我们将无比高兴。XML 的内容极为丰富，绝不是一部教材可以涵盖的，我们只是选取了较为基本、实用的部分。读者欲在实际系统中使用 XML 文档，还需要参考其他资料特别是丰富的网上资源。我们在此也想给读者一条建议：学习 XML 技术的根本方法是实际的应用而不是对条文的记忆。由于我们水平有限，书中错误在所难免，敬请批评指正。

编者

2011 年 12 月

目 录

第一章 概述

主要内容

- ▶ XML 发展过程
- ▶ XML 的优点
- ▶ XML 设计的目标
- ▶ XML 编辑工具
- ▶ 知识体系介绍

难点

- ▶ 标记的理解
- ▶ 知识体系理解

1.1 XML 的发展史

XML 的全称是 Extensible Markup Language，其意为可扩展的标记语言，它是标准通用标记语言(Standard Generalized Markup Language，SGML)的一个子集。那么标记语言又是什么呢?

1.1.1 标记语言产生

首先解决的一个问题是:“什么是标记?”标记一般意义上的理解是记号。其实我们现实生活中用到记号的情景很多，例如我们阅读时在书中用自己熟悉的记号在感兴趣的部分或是重点内容处标出。所以，我们可以理解成记号的作用是将某一部分内容与其他的内容区分开来，只不过计算机所能处理的记号不是我们随手做的记号，而是文档中的电子记号。

标记语言就是使用某种“记号”来表示特殊信息的语言，如用一种“记号”来表示格式信息或表示数据信息。

表示格式的事例大量存在于 HTML 中，可以这样讲，HTML 中的绝大部分标记都是格式的指定。如 < table > 是要生成一个表格，< p > 是要生成一个自然段，< font > 是要指定字体字号等。为了更好地理解标记语言，我们从一些具体应用软件介绍。

1.1.2 RTF 标记语言

RTF(Rich Text Format)是在文字处理软件中广泛应用的一种标记语言，Windows 系

统自带的 Word 和写字板等软件都支持这种标语言。标记语言离我们很近，我们每天做文档资料时都要和它打交道。RTF 语言预先定义了许多标记，这些标记可以表示字体、排版等各种信息。下面通过具体的实例来感知这种语言的特点。

(1)打开 Word 2003，进入如图 1.1 所示的窗口。

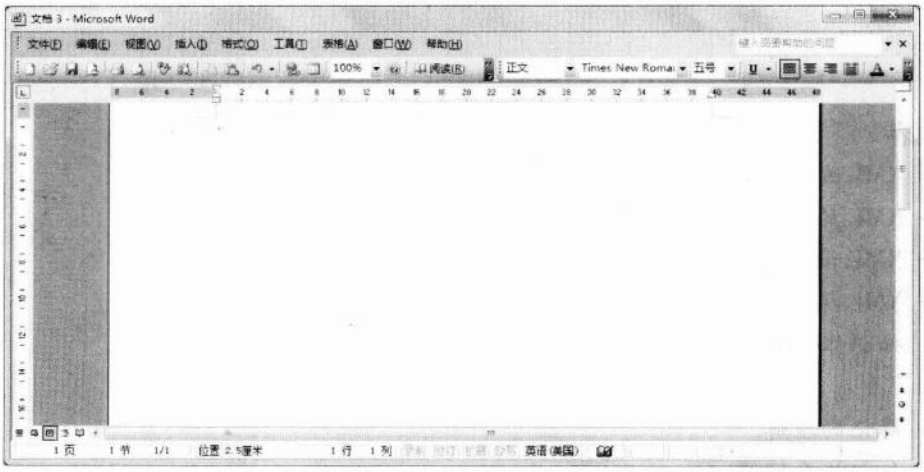

图 1.1 Word 2003 窗口

(2)在窗口中输入“去年今日此门中，人面桃花相映红”，将“桃花”改成红色，其最终结果如图 1.2 所示。

图 1.2 输入文字后结果

(3)将文档保存成 RTF 文件格式，保存格式如图 1.3 所示。

图 1.3 保存文件界面

该文件已经保存到桌面的文件中，使用记事本程序打开该文件的结果，如图 1.4 所示。

图 1.4 RTF 源码展示

Word 2003 只是这些代码的生成工具，是一种代码生成器。读者可以在记事本中手工创建该文件(但是相当不易！这也是代码生存工具存在的必要性)，其效果是一样的。

Word 文档也是标记语言生成的，但是当使用记事本打开任意一个 Word 文档时得到的结果却与 RTF 文档有很大不同，如图 1.5 所示。

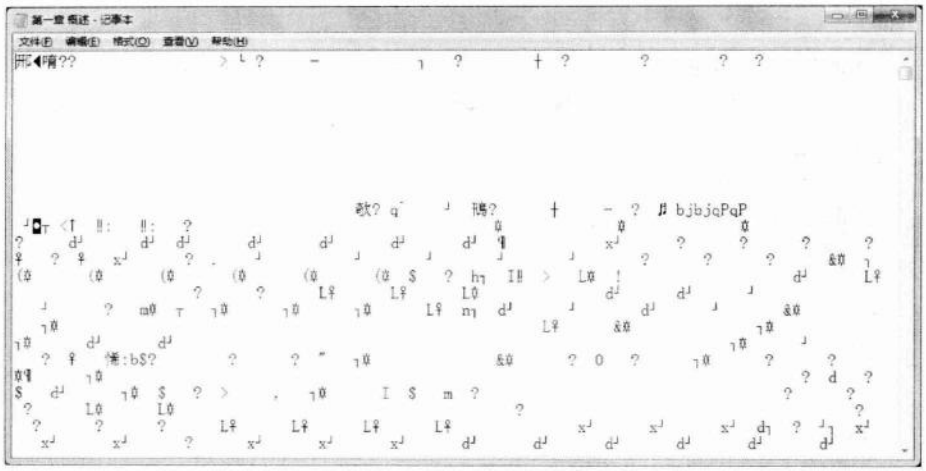

图 1.5　用记事本打开 Word 文档的结果

为什么会出现这种情况？这是因为 Word 应用软件在生成标记语言文件后又将其转换成二进制形式。

上面我们介绍了日常办公软件中存在的标记语言，说明了标记语言其实离我们很近。下面介绍在 Web 使用的一种标记语言 HTML。

1.1.3　HTML 标记语言

HTML(HyperText Markup Language，超文本语言标记)是为“网页创建和其他可在网页浏览器中看到的信息”设计的一种标记语言。HTML 被用来格式化信息，例如标题、段落、字体和列表等，由 IETF 用简化的 SGML(标准通用标记语言)语法进行进一步发展的 HTML，后来成为国际标准，由万维网联盟(W3C)维护。该语言最大的特点是简单明了，Web 技术盛行其道，HTML 可谓居功至伟。下面通过一个例子来认识一下这种标记语言的特点。具体见程序 1 - 1. html。

程序 1 - 1. html

```
<HTML>
<head>
<title>你好</title>
</head>
<body>
<p>这是一个 HTML 网页</p>
</body>
</HTML>
```

该文件的作用效果很简单，只在浏览器的窗口中显示“这是一个 HTML 网页”，在 IE 浏览器中的显示结果如图 1.6 所示。

图 1.6 HTML 标记语言在浏览器中显示结果

本书几乎所有的程序都可以在记事本中完成,调试工具可以使用 Web 浏览器和记事本。下面的例子显示了如何调试程序 1 - 1. html。

(1)打开记事本程序

(2)在该记事本中输入程序 1 - 1. html 代码,然后保存,保存的设置如图 1.7 所示。

图 1.7 文件保存成 HTML 的方法

保存时需注意:其一,在文件名文本框内输入“1 - 1. html”,即强制性地将文件命名为“1 - 1. html”,表示它的格式是 HTML 的;其二,保存的类型选择“所有文件”的。根据

笔者的实践，一般地说文件名可以不加引号。还是看最后形成的文件是否扩展名为 txt。

1.1.3.1 HTML 的特点

(1)预定义标记。HTML 文档的所有标记都是预先定义好的，允许用户使用的标记是极为有限的。HTML 中可以使用的标记不超过 100 个，而常用的也就几十个。这些标记是国际上公认的通用标记，用户只要记住这些标记的用法，就可以在 Internet 上发布网页。软件对某个标记的处理对所有的 HTML 文档都是一样的。

(2)语法要求宽松。HTML 在语法上要求极其宽松，如程序语句对大小写不敏感，标记不一定配对使用，允许标记的不合理嵌套等。

(3)制作 HTML 文档的应用软件很多。首先所有的文本编辑器都可以用来生成 HTML 文档，只要保存为 HTML 文档即可。其次还有大量的专用软件制作 HTML 网页，如著名的 FrontPage 和 Dreamware 等。第二类工具是制作 HTML 网页的主力军，其支持“所见即所得”，使网页制作的速度大大地提高，为 HTML 在 Internet 上流行奠定了坚实的基础。

1.1.3.2 HTML 的缺点

(1)标记固定，用户不能进行扩展。无论用户表达的是什么内容，数据使用的标记都是一样的，不能为特别的应用量身定制。

(2)HTML 的标记是为排版服务的。HTML 本质上是一种格式显示语言，和 RTF 标记所起的作用是相似的。它不能把数据和格式区分开来，这个缺点导致了 XML 的出现。

(3)HTML 标记语言的标准不统一。HTML 和浏览器的关系极为密切。HTML 的网页效果只有通过浏览器的解释才能体现出效果，而各厂商的浏览器产品对标记的支持不尽相同，导致同一个网页在不同的浏览器下效果有可能不一致，从而出现为适应不同的浏览器而编制 HTML 的现象。

1.1.4 标准通用标记语言

开篇就提到 XML 是 SGML 的子集，本节对其进行简单的介绍。在 20 世纪 80 年代早期，IBM 提出在各文档之间共享一些相似的属性，例如字体大小和版面。IBM 设计了一种文档系统，通过在文档中添加标记，来标识文档中的各种元素，IBM 把这种标识语言称作通用标记语言(Generalized Markup Language，GML)。经过若干年的发展，1984 年国际标准化组织(ISO)开始对此提案进行讨论，并于 1986 年正式发布了为生成标准化文档而定义的语言标准(ISO 8879)，称为新的语言 SGML，即标准通用标记语言。下面简单介绍其优缺点。

1.1.4.1 SGML 的优点

(1)稳定性高。SGML 从成为国际规范以来，已经发展了三十多年，可信度高，而且构架极为规范。

(2)功能完备。SGML 的功能极为强大，是一种元语言，可以对数据以及数据之间的关系进行描述。

(3)可移植性好。SGML 的设计一开始就注意到了通用性，能够跨越不同的软硬件平台。