Broadview
www.broadview.com.cn

Java Web
开发三部曲

SunXin's Series
孙鑫作品系列

国内第一本介绍XQuery和XQJ的图书

涵盖最新的XSLT 2.0、XQuery 1.0和XQJ规范

循序渐进的练习、检查真分的示例、先进的知识体系

XML、XML Schema、XSLT 2.0和XQuery开发详解

孙鑫 编著

- 国内第一本介绍XQuery和XQJ的图书
- 涵盖最新的XSLT 2.0、XQuery 1.0和XQJ规范
- 7种最常用和最新的XML技术：XML、DTD、XML名称空间、XML Schema、XPath 1.0和XPath 2.0、XSLT 1.0和XSLT 2.0，XQuery
- 4种Java解析XML文档技术：DOM、SAX、JDOM，dom4j
- 1种支持最新XQuery标准的查询API：XQJ（XQuery for Java API）

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Java Web 开发三部曲

SunXin's Series 孙鑫作品系列

XML、XML Schema、XSLT 2.0和XQuery开发详解

孙鑫 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书内容全面，详细讲解了目前最常用和最新的 XML 技术，包括：XML、DTD、XML 名称空间、XML Schema、XPath 1.0 和 XPath 2.0、XSLT 1.0 和 XSLT 2.0，以及 XQuery。此外，本书还介绍了如何使用 DOM、SAX、JDOM 和 dom4j 来解析和验证 XML 文档，以及使用最新的 XQJ API 来查询 XML 数据。

本书语言生动、通俗易懂、讲解细致，所有章节都提供了大量的例子，以帮助读者更好地理解所学的内容。

本书在内容的安排上独具匠心，在知识体系的讲解上由浅入深、循序渐进。

本书不仅可以作为 XML 开发的学习用书，还可以作为从事 XML 开发的程序员的参考用书和必备手册。

图书在版编目（CIP）数据

XML、XML Schema、XSLT 2.0 和 XQuery 开发详解 / 孙鑫编著. —北京：电子工业出版社，2009.1
（Java Web 开发三部曲）
ISBN 978-7-121-07737-1

Ⅰ. X… Ⅱ. 孙… Ⅲ. ①可扩充语言，XML—程序设计 ②可扩充语言，XSLT—程序设计
③Java 语言—程序设计 Ⅳ. TP312

中国版本图书馆 CIP 数据核字（2008）第 177417 号

责任编辑：李 冰
印 刷：北京智力达印刷有限公司
装 订：北京中新伟业印刷有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：787×1092 1/16 印张：30.25 字数：610 千字
印 次：2009 年 1 月第 1 次印刷
印 数：4000 册 定价：55.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

本书内容丰富，讲解了 7 种最常用和最新的 XML 技术：XML、DTD、XML 名称空间、XML Schema、XPath 1.0 和 XPath 2.0、XSLT 1.0、XSLT 2.0 和 XQuery；4 种 Java 解析 XML 文档技术：DOM、SAX、JDOM，dom4j；1 种支持最新 XQuery 标准的查询 API：XQJ（XQuery for Java API）。

本书面向的读者

- 毫无 XML 经验的初学者
- 有一定的 XML 经验，但没有从事过 XML 开发的读者
- 正在从事 XML 开发的初中级程序员
- 正在从事 XSLT 开发的程序员
- 正在从事 XQuery 开发的程序员

本书的内容组织

全书共分 5 部分，包括了 XML 篇、XML Schema 篇、解析篇、XSLT 篇和 XQuery 篇。本书在内容的安排上独具匠心，在知识体系的讲解上由浅入深、循序渐进。

本书第 I 部分是 XML 篇，包括以下几个主题：

- XML
- DTD
- XML 名称空间

本书第 II 部分是 XML Schema 篇，详细讲解了 XML Schema 这一功能强大的模式语言。这部分内容包括以下几个主题：

- XML Schema 初窥
- 内置简单类型
- 自定义简单类型
- 复杂类型
- 派生复杂类型
- 一致性约束
- 引入其他的模式文档

本书第III部分是解析篇，详细讲解了 DOM、SAX、JDOM 和 dom4j 解析 XML 文档。这部分内容包括以下几个主题：

- 使用 DOM、SAX 和 JAXP 解析 XML 文档
- 使用 JDOM 解析 XML 文档
- 使用 dom4j 解析 XML 文档

- 解析名称空间

本书第Ⅳ部分是XSLT篇，介绍了XSLT1.0/XPath 1.0，以及最新的XSLT 2.0/XPath 2.0。这部分内容包括以下两个主题：

- XSLT 1.0
- XSLT 2.0

本书第Ⅴ部分是XQuery篇，介绍了XML查询语言XQuery，以及在Java中执行XQuery查询的标准API —— XQJ。这部分内容包括以下两个主题：

- XQuery
- 在Java中使用XQuery —— XQJ

本书示例程序的下载

本书完整的示例程序可以从博文视点的网站（http://www.broadview.com.cn）和笔者的个人网站上下载（http://www.sunxin.org）。

最后，衷心地希望本书能够给读者带来知识，带来阅读上的快乐，读者对本书的肯定就是笔者最大的欣慰。由于本书的内容较多、牵涉的技术较广，错误和疏漏之处在所难免，欢迎广大技术专家和读者指正。作者的联系方式是csunxin@sina.com，读者也可以上作者的网站发表意见，网址是：http://www.sunxin.org/。

目　录

第 I 篇　XML 篇

第 II 篇　XML Schema 篇

第Ⅲ篇 解析篇

第Ⅳ篇 XSLT 篇

第V篇 XQuery 篇

第 I 篇

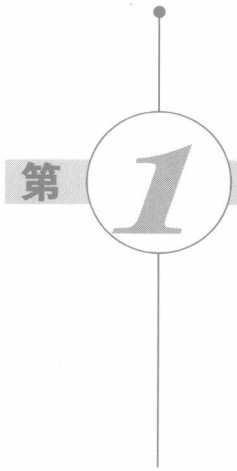
XML 篇

XML 已经成为大家耳熟能详的一个词汇，话说微软力推的少数几项非自己专有技术的标准中，就有大名鼎鼎的 XML。XML 一经推出，就得到了业界 IT 巨头的响应，一时间应者云集，XML 很快就在各行各业中展露了它的身影。XML 能够独立于计算机平台、操作系统和编程语言来表示数据，凭借其简单性、可扩展性、交互性和灵活性在计算机工业中获得了世界范围的支持和采纳。人们使用 XML 作为配置文件的首选格式，使用 XML 在不同的系统、不同的数据库，以及不同的软件之间传输数据，使用 XML 作为简易数据库来保存企业信息数据。

XML 的流行使得很多软件开发人员在开发项目的时候，不管是否需要，都在项目中使用了 XML，从而导致软件系统性能的降低，这种盲目追求新技术或流行技术的做法并不可取。

为了帮助读者对 XML 技术有一个清晰而全面的认识，本篇将重点讲述 XML、DTD，以及 XML 名称空间。本篇的内容包括：

- 第 1 章　XML
- 第 2 章　DTD
- 第 3 章　XML 名称空间



第1章 XML

本章要点

- 了解 XML 的起源
- 弄清楚 XML 与 HTML 的区别
- 学会使用 XMLSpy
- 了解 XML 的物理结构
- 掌握 XML 的逻辑结构
- 学会编写格式良好的 XML 文档

XML 对我们来说已不再陌生，其相关概念和知识在网络上随处可见，有关 XML 的应用也越来越多。本章的目的是帮助读者快速了解和掌握 XML，为后面章节的学习打下基础。

1.1 XML 的起源

XML 的全称是 Extensible Markup Language，意思是可扩展的标记语言，它是标准通用标记语言（Standard Generalized Markup Language，SGML）的一个子集。那 SGML 又是什么呢？

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 把这种标识语言称做通用标记语言（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标记语言标准（ISO 8879），称为新的语言 SGML，即标准通用标记语言。

SGML 功能非常强大，是可以定义标记语言的元语言，然而由于 SGML 过于复杂，不适合在 Web 上应用，因此 W3C 组织在 1996 年便开始设计一种可扩展的标记语言，以便

能将 SGML 的丰富功能与 HTML 的易用性结合到 Web 应用中。1998 年 2 月 10 日，W3C 发布了 XML 1.0 标准，其目的是为了在 Web 上能以现有的超文本标记语言（HTML）的使用方式提供、接收和处理通用的 SGML。XML 是 SGML 的一个简化子集，它以一种开放的、自我描述的方式定义了数据结构。在描述数据内容的同时能突出对结构的描述，从而体现出数据与数据之间的关系。

本书介绍的 XML 主要遵循 W3C 于 2006 年 8 月 16 日发布的 XML 1.0 推荐标准的第四版，读者可以在 http://www.w3.org/TR/REC-xml/了解到此标准的详细内容。

W3C 组织于 2004 年 2 月 4 日，发布了 XML 1.1 的推荐标准，并于 2006 年 8 月 16 日发布了 XML 1.1 推荐标准的第二版，这是最新的 XML 版本，不过由于目前大多数的应用还是基于 XML 1.0 的推荐标准，因此本书也将遵照 XML 1.0 规范来讲述。如果读者想要了解 XML1.1 规范的内容，可以参看网址：http://www.w3.org/TR/2006/REC-xml11-20060816/。

1.2 W3C 介绍

W3C 是万维网联盟（World Wide Web Consortium）英文的缩写，成立于 1994 年 10 月，它以开放论坛的方式来促进开发互通技术（包括规格、指南、软件和工具），以激发网络的全部潜能。万维网联盟（W3C）的会员包括软件开发商、内容提供商、企业用户、通信公司、研究机构、研究实验室、标准化团体，以及政府，会员中的一些知名 IT 企业包括：IBM、Microsoft、America Online、Apple、Adobe、Macromedia、Sun Microsystems 等。

W3C 自成立以来，已发布了 100 多份 Web 技术规范，领导着 Web 技术向前发展。

W3C 认为自身不是官方组织，因此将它正式发布的规范称为**推荐（建议）标准**，意思是进一步标准化的建议，但是由于该组织自身的权威性往往成为事实上的标准。一项技术要成为 W3C 的推荐标准，需要经过以下 7 个步骤。

（1）W3C 收到提交（W3C Submissions）

任何 W3C 的成员都可以向联盟提交希望成为 Web 标准的某项建议。如果建议的内容在 W3C 的工作范围内，W3C 将决定是否要对此开展工作。

（2）W3C 发布注释（W3C Notes）

通常，一项对 W3C 的提交会成为一份注释。注释作为一份公共的文档，是对建议的描述。

注释仅供讨论使用。一项注释的发布，并不表示 W3C 对其认可了。注释的内容由提交此注释的会员来编辑，而不是 W3C。注释可以在发布后随时被更新、替换或废弃。注释的发布，并不表示 W3C 已经开始任何与注释相关的工作。

（3）W3C 成立工作组（W3C Working Groups）

当一项提交被 W3C 认可之后，就成立由成员和一些对此感兴趣的团体参加的工作组。通常情况下，工作组将定义一个时间表，并发布一个提议标准的工作草案，描述当前的工作进展。

（4）W3C 发布工作草案（W3C Working Drafts）

W3C 通常会在其网站上（http://www.w3.org）发布工作草案，以及一个公众讨论的邀请。工作草案会说明当前的工作进展。由于工作草案的内容可随时被更新、替换或废弃，所以不应把它作为工作的依据。

（5）W3C 发布候选推荐标准（W3C Candidate Recommendations）

当规范比较复杂时，可能需要成员和软件提供商花费更多的时间来试用或测试。有时候，这些规范以候选推荐标准的形式发布。它与工作草案一样都是进展中的工作文件，所以不应把它作为工作的依据。该文档随时可以被更新、替换或废弃。

（6）W3C 发布提议的推荐标准

提议的推荐标准的发布标志着工作组的工作到了最后阶段。提议的推荐标准仍然是一个进展中的工作文件，仍然可以被更新、替换或废弃。虽然提议的推荐标准还没有被 W3C 正式认可，但是它在内容和时间上离最终的推荐标准已经非常接近了。

（7）W3C 发布推荐标准

推荐标准经过 W3C 的成员审阅，并由 W3C 的主任加盖正式批准图章，而最终成为规范。W3C 的推荐标准是一个稳定的文档，可以作为工作中的参考资料。

1.3 关于 XML 的两个问题

1. XML 是 HTML 的扩展吗

HTML 的全称是 Hypertext Markup Language（超文本标记语言），而 XML 的全称是 eXtensible Markup Language（可扩展的标记语言），这很容易让人联想到 XML 是通过增加新标记来扩展 HTML 的一种标记语言。实际上 HTML 和 XML 在标记语言中处于不同的层次。

下面我们通过 HTML 文档和 XML 文档的对比（分别如例 1-1 和例 1-2 所示），来弄清楚 XML 的一些概念。

例 1-1　HelloWorld.html

```
<html>
    <head>
        <title>这是一个欢迎的例子</title>
    </head>
    <body>
        你好！欢迎你！
    </body>
</html>
```

例 1-2 HelloWorld.xml

```
<?xml version="1.0" encoding="GB2312"?>
<欢迎词>
    <标题>这是一个欢迎的例子</标题>
    <内容>你好！欢迎你！</内容>
</欢迎词>
```

从上面的两个例子，我们可以看出：

① 在编写 HTML 文档时，所有的标记都已经固定下来（如<html>、<body>等），我们不能去创造新的标记；而在编写 XML 文档时，我们可以任意地创建新的标记，包括中文的标记（如<欢迎词>、<内容>等）。所以说 XML 是可扩展的标记语言。

② 在编写 XML 文档时，没有一套标准的标记供我们选择使用，需要我们自己去创建标记，所以我们说 XML 是创建标记语言的元语言。

XML 在设计之初，就考虑到了国际化的问题，同 HTML 4.01 一样，XML 也是基于 ISO/IEC 10646 字符集标准中定义的通用字符集（Universal Character Set，UCS），等价于 Unicode 2.0。

2. SGML、HTML 和 XML 之间是什么关系

SGML 是一种在 Web 发明之前就早已存在的使用标记来描述文档资料的通用语言。它是一种定义标记语言的元语言。HTML 和 XML 都是从 SGML 发展而来的标记语言，因此，它们有一些共同点，如相似的语法和标记的使用。不过 HTML 是在 SGML 定义下的一个描述性的语言，只是 SGML 的一个应用，其 DTD（参见第 2 章）作为标准被固定下来，而 XML 是 SGML 的一个简化版本，是 SGML 的一个子集，严格意义上来说，XML 仍然是 SGML。

HTML 不能用来定义新的应用，而 XML 可以，例如 RDF 和 CDF 都是使用 XML 定义的应用。事实上，XML 和 SGML 是兼容的，但又没有 SGML 那么复杂，它被设计用于有限带宽的网络，如 Internet。XML 规范的制定者之一 Tim Bray 说，“XML 的设计出发点是取 SGML 的优点，去除复杂的部分，使其保持轻巧，可以在 Web 上工作”。

下面我们用一个比喻来描述它们三者之间的关系。假如我们制定了一套设计图形元素的规则，称为 SGML。现在为了规范建筑行业建筑图的绘制，需要使用一套固定的图形元素，于是我们根据 SGML 这种设计图形元素的规则设计了一套专门用于绘制建筑图的图形元素，那么这套图形元素就称为 HTML。然而由于 SGML 过于复杂，不利于在其他行业推广，因此我们对 SGML 进行了简化，重新制定了一套设计图形元素的规则，这就是 XML。由于 XML 是 SGML 的一个子集，因此它也可以用来设计图形元素。

HTML、SGML 和 XML 将继续用于其适合的地方，它们中的任何一个都不会使其他一个废弃。对于像新闻、网络日记、论坛留言等大部分短期的数据，HTML 仍是在 Web 上快速发布数据的最简单的方法。如果数据要长期使用，并且需要更多的一些结构，我们更推荐使用 XML。不同于 HTML 和 XML，SGML 可能永远不会在 Internet 上被广泛接受，因为它不是为某个网络协议而设计的，也从来没有为某个网络协议的需求而优化过。对于高端的、复杂结构的发布应用，SGML 将继续使用。

1.4 XML 与 HTML 的比较

通过 XML 与 HTML 的比较，我们能够更好地理解和掌握 XML 的优点。

1.4.1 XML 将数据与显示分开

我们看附录中的一个例子。

例 1-3 SecondPage5.htm

```
<html>
    <head>
        <title>静夜思</title>
    </head>
    <body>
        <center>
                <h2><font color="red">静夜思</font></h2>
                    <b>作者：李白</b>
            <hr color="blue">
            <p>
            <b><i><font size=3 color="green">床前明月光，疑是地上霜。<br>
            举头望明月，低头思故乡。</font></i></b>
        </center>
    </body>
</html>
```

在上面的这个 HTML 文档中，“静夜思；作者：李白；床前明月光，疑是地上霜。举头望明月，低头思故乡。”这些是要显示的数据， 表示让内容居中显示，<h2>表示 2 号标题，<b>表示以粗体显示，<i>表示以斜体显示。

可以看到，HTML 文档将数据，页面的排版，以及页面的表现形式混合在了一起，如果要增加新的数据，相应地就要调整数据的排版与显示方式。当我们从其他的地方（例如，数据库、文本文件）得到数据后，一旦放入 HTML 文档中，整个数据就会被打乱，数据和 HTML 标记混合在了一起，数据本身就变得无法辨析了。另一方面，随着电子商务等网络应用的流行，不同系统、不同平台、不同软件的信息交换日益频繁，而 HTML 本身的这些限制，导致了 HTML 在日益广泛的 Web 应用中，显得捉襟见肘，为此，就有了 XML 的产生。

我们将例 1-3 中的 HTML 文档中的数据部分改为用 XML 文档来表示，如例 1-4 所示。

例 1-4 SecondPage5.xml

```
<?xml version="1.0" encoding="GB2312"?>
<poem>
    <title>静夜思</title>
    <author>李白</author>
    <content>
```

```
        <line>床前明月光</line>
        <line>疑是地上霜</line>
        <line>举头望明月</line>
        <line>低头思故乡</line>
    </content>
</poem>
```

在这个文档中，我们用<title>（此处的<title>是我们自己创建的标记）指明标题是“静夜思”，用<author>表明作者是“李白”，而内容是“床前明月光，疑是地上霜。举头望明月，低头思故乡”这四句诗。那么这个 XML 文档将如何显示呢？从文档中我们看不出这些数据将如何显示，这是因为：XML 文档不能描述页面的排版和表现形式，它只是用于描述数据和数据的结构。也就是说，XML 将数据和显示分开了，我们可以为这些数据设计不同的排版和表现形式，而数据本身不需要做任何的修改。

采用 XML 来表示数据，我们能够很容易地读懂 XML 文档，而计算机也能够很好地进行识别和处理。XML 表示数据的方式真正做到了独立于应用系统，并且数据能够重用，一份数据可以应用于不同的场合。有时候，XML 文档也被看作是文档的数据库化和数据的文档化。

1.4.2 XML 对文档的格式要求更加严格

由于 HTML 文档格式非常松散，导致了 HTML 文档解析的复杂性，也造成了浏览器兼容的问题，所以 XML 从一开始，就对文档的格式制定了非常严格的标准，凡是符合这一标准的 XML 文档就是格式良好的 XML 文档（Well-Formed XML Documents）。

（1）开始标签必须要有一个配套的结束标签

在 HTML 文档中，可以直接使用 、<tr>、<td>等标签，而不用加结束标签，在 XML 中，开始标签和结束标签必须配套，也就是必须写成 … 、<tr>…</tr>或<td>…</td>。

（2）空元素标签必须被关闭

在 HTML 文档中，可以使用
、<hr>、<img src="logo.gif" width="197" height= "81">等单标签，而在 XML 中，空元素标签必须被关闭。空元素标签采用斜杠（/）来关闭，例如：
、<hr/>、。

（3）所有的标签都区分大小写

在 HTML 文档中，标签是不区分大小写的，<tr>和</TR>是 tr 元素的开始标签和结束标签，但是在 XML 中，<tr>和<TR>是两个不同的标签，开始标签和结束标签的大小写形式必须一致。

（4）所有的标签都必须合理嵌套

在 HTML 文档中，…是允许的，但是在 XML 中，这是错误的。在 XML 中，所有的标签都要成对出现，合理嵌套，正确的形式是：…。