Learning XML

XML 入门

O'REILLY®

中国电力出版社

Erik T. Ray 著

卓小涛 译

XML入门

Erik T. Ray 著

卓小涛 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

O'Reilly & Associates, Inc. 授权中国电力出版社出版

中国电力出版社

图书在版编目（CIP）数据

XML 入门/（美）雷（Ray, E. T.）编著；卓小涛译．-北京：中国电力出版社，2001

书名原文：Learning XML

ISBN 7-5083-0842-5

Ⅰ.X… Ⅱ.①雷… ②卓… Ⅲ.可扩充语言，XML－程序设计 Ⅳ.TP312

中国版本图书馆 CIP 数据核字（2001）第 078073 号

北京市版权局著作权合同登记

图字：01-2001-4627 号

书　　名/ XML 入门

书　　号/ ISBN 7-5083-0842-5

责任编辑/ 夏平

封面设计/ Edie Freedman，张健

出版发行/ 中国电力出版社（www.infopower.com.cn）

地　　址/ 北京三里河路 6 号（邮政编码 100044）

经　　销/ 全国新华书店

印　　刷/ 北京市地矿印刷厂

开　　本/ 787 毫米 × 1092 毫米　16 开本　22.5 印张　330 千字

版　　次/ 2001 年 11 月第一版　2001 年 11 月第一次印刷

印　　数/ 0001-5000 册

定　　价/ 42.00 元（册）

XML 入门

O'Reilly & Associates 公司介绍

为了满足读者对网络和软件技术知识的迫切需求，世界著名计算机图书出版机构O'Reilly & Associates公司授权中国电力出版社，翻译出版一批该公司久负盛名的英文经典技术专著。

O'Reilly & Associates公司是世界上在UNIX、X、Internet和其他开放系统图书领域具有领导地位的出版公司，同时是联机出版的先锋。

从最畅销的《The Whole Internet Use's Guide & Catalog》（被纽约公共图书馆评为二十世纪最重要的50本书之一）到GNN（最早的Internet门户和商业网站），再到WebSite（第一个桌面PC的Web服务器软件），O'Reilly & Associates一直处于Internet发展的最前沿。

许多书店的反馈表明，O'Reilly & Associates是最稳定的计算机图书出版商——每一本书都一版再版。与大多数计算机图书出版商相比，O'Reilly & Associates公司具有深厚的计算机专业背景，这使得O'Reilly & Associates形成了一个非常不同于其他出版商的出版方针。O'Reilly & Associates所有的编辑人员以前都是程序员，或者是顶尖级的技术专家。O'Reilly & Associates还有许多固定的作者群体——他们本身是相关领域的技术专家、咨询专家，而现在编写著作，O'Reilly & Associates依靠他们及时地推出图书。因为O'Reilly & Associates紧密地与计算机业界联系着，所以O'Reilly & Associates知道市场上真正需要什么图书。

目录

前言

自从二十世纪九十年代后期可扩展标记语言（Extensible Markup Language，XML）问世以来，已相继发布了大量与XML相关的缩略语、标准和规则。Internet社团中的一些人怀疑XML是否真的必要。毕竟，HTML已流行了多年并已培育了一种全新的经济和文化，那么为什么要改变一种好的事物呢？实际上，XML并不是要取代Web上已存在的东西，而是要建立一种更坚实和更灵活的基础。XML是由某些组织和公司组成的联盟为了创建一种HTML无法胜任的二十一世纪信息框架、而做的前所未有的努力。

为了理解XML的重要性，我们有必要澄清一些概念。第一，与其名字不同，XML并不是一种标记语言，而是一种创建、设计和使用标记语言的工具集。这个事实同时澄清了第二个误解，即XML会取代HTML。实际上，HTML将融入XML，变为一个更清晰的版本，称为XHTML。这只是一个开始，因为XML将创建大量的新标记语言，以涵盖各种应用和文档类型。

标准化过程将在此信息革命的发展过程中起着极其重要的作用。竞争性技术和专有语言的未加控制的发展已威胁到Web的完整性，XML本身就是为了控制这种趋势的一种尝试。XML创建了一种平台，通过它结构化信息可以与应用程序很好地配合，它最大化了信息的可访问性，同时又不会牺牲表达的丰富性。

XML已被Internet社团广泛接受，这为许多XML相关标准的制定打下了坚实的基础。XML的一些新的相关技术包括针对显示的样式表、转换、链接资源的强大方

法、数据处理和查询工具、错误检查工具、强制结构工具和大量的开发环境。这些新的应用，确保了XML成为一个优秀的结构化信息工具集。

当然，XML仍然是一种新生事物，而且它的一些相关技术也不成熟。在本书中讨论的某些主题只是半推测性的，因为它们的规范现在还是工作草案。然而，与其以后对新技术感到惊奇、还不如尽早学习。如果你正从事Web开发或信息管理，那么你就有必要了解XML。

本书的目的是向你展示正在发展中的XML的全貌。为了充分利用本书，你应当事先熟悉一下结构化标记（如HTML或TEX）、和万维网（World Wide Web）的概念，如超文本链接和数据表示。然而，你不必是一位理解XML概念的开发人员。我们将集中讨论文档编辑的理论和实践，而不详细讨论如何编写应用程序或学习软件工具。有关XML编程的具体细节请参阅其他书籍，而工业界的迅速发展使我们永远难以跟上最新的XML软件。然而，本书提供的信息将作为你进一步学习XML的良好起点。

本书的内容

本书包括下述各章：

第一章是对XML及其用途的综述。它是本书其余部分的出发点，介绍了将在后续各章中详细解释的主要概念。

第二章描述了XML的基本语法，是理解XML应用和技术的基础。

第三章介绍了在文档和资源之间创建简单链接的方法，是XML的一个重要方面。

第四章通过层叠样式表语言介绍了样式表的概念。

第五章介绍了文档类型定义（DTD）和XML Schema。这些是确保文档质量和完整性的主要技术。

第六章介绍了如何创建转换样式表，以从一种XML格式转换为另一种格式。

第七章介绍了XML的可访问性和国际化的一面，包括Unicode、字符编码和语言支持。

第八章提供了关于编写处理XML的软件的综述。

另外，还有两个附录和一个词汇表：

附录一包含进一步学习XML的参考文献。

附录二列出了与XML相关的技术。

词汇表是对本书中所使用术语的解释。

排版约定

为与正常文本区分开，本书在英文字体上有如下约定：

斜体（*italic*）
: 用来说明对书籍和文章的引用，命令，电子邮件地址，URL，文件名，强调文本和对术语的第一次引用。

等宽字体（`constant width`）
: 用来说明直接量，常量，代码清单和XML标记。

等宽斜体字（*`constant width italic`*）
: 用来说明可取代的参数和变量名。

等宽粗体字（**`constant width bold`**）
: 用来强调代码清单中被讨论的部分。

示例

本书的示例可从本书的Web站点*http://www.oreilly.com/catalog/learnxml*上免费下载。

建议与评论

本书的内容都经过测试，尽管我们做了最大的努力，但错误和疏忽仍然是在所难免

的。如果你发现有什么错误，或者是对将来的版本有什么建议，请通过下面的地址告诉我们：

美国：

O'Reilly & Associates,Inc.
101 Morris Street
Sebastopol, CA 95472

中国：

100080 北京市海淀区知春路 49 号希格玛公寓 B 座 809 室
奥莱理软件（北京）有限公司

询问技术问题或对本书的评论，请发电子邮件到：

info@mail.oreilly.com.cn

我们为本书建立了一个Web页面，该页面列出了勘误表、例子或其他附加信息。你可以通过如下地址来访问：

http://www.oreilly.com/catalog/learnxml

最后，您可以在 WWW 上找到我们：

http://www.oreilly.com
http://www.oreilly.com.cn

致谢

本书能够出版，要感谢：优秀的编辑 Andy Oram, Laurie Petrycki, John Posner 和 Ellen Siever；制作人员 Colleen Gorman, Emily Quill 和 Ellen Troutman-Zaig；杰出的审稿人 Jeff Liggett, Jon Udell, Anne-Marie Vaduva, Andy Oram, Norm Walsh 和 Jessica P. Hekman；尊敬的合作者 Sheryl Avruch, Cliff Dyer, Jason McIntosh Lenny Muellner,Benn Salter,Mike Sierra 和 Frank Willison；为本书编写附录的 Stephen Spainhour；以及为本书投入编写提供必要预备知识的 Chris Maden。

另外，我还要感谢我的妻子Jeannine Bestine的鼓励和支持；感谢我家人（妈妈：Birgit，Helen，爸爸：Al，Butch，以及Ed，Elton，Jon-Paul，奶奶和爷爷：Bestine，Mare，Margaret，Gene，Lianne）所给予的始终不渝的爱心和美味的食物；感谢我的宠物鸟Estero，Zagnut，Milkyway，Snickers，Punji，Kitkat和Chi Chu；感谢我的好友Derrick Arnelle，J. David Curran先生，Sarah Demb，Chris "800" Gernon，John Grigsby，Andy Grosser，Lisa Musiker，Benn "Nietzsche" Salter和Greg "Mitochondrion" Travis；感谢给我提供灵感和勇气的Laurie Anderson，Isaac Asimov，Wernher von Braun，James Burke，Albert Einstein，Mahatma Gandhi，Chuck Jones，Miyamoto Musashi，Ralph Nader，Rainer Maria Rilke和Oscar Wilde；特别感谢Weber提供的美味的三明治。

本章内容：

- 什么是XML?
- XML的起源
- XML的目标
- XML的现状
- 创建文档
- 查看XML
- 测试XML
- 转换

第一章

概述

可扩展标记语言（Extensible Markup Language，XML）是一种各类信息的数据存储工具包和可配置载体，同时也是一种已被各界人士（从银行家至网站管理员）广泛接受的、正在发展中的开放式标准。在过去的几年中，它也引起了技术权威和工业专家的关注。那么XML成功的秘密是什么呢？

下面的XML特征列表可以说明一切：

- XML可以以一种符合用户需求的格式来存储和组织几乎任何信息。
- XML是一种开放式标准，既不属于任何独立公司，也不属于任何特定软件。
- XML使用Unicode作为标准字符集，它支持大量的书写系统（文字）和符号，包括北欧古文字字符和中国象形文字。
- XML提供了语法规则、内部链接检查、与文档模型比较和数据类型等多种方法来检验文档的质量。
- 由于XML具有清晰简单的语法和无歧义的结构，因此XML对于人类和程序同样易于阅读和解析。
- XML可以很容易地与样式表组合，以便以任何用户想要的样式来创建格式化文档。信息结构的单一性不会妨碍格式转换。

所有这一切意味着一个使用新的交流方式的时代即将到来。我们可触及的信息量是

惊人的，但是现有技术的限制使得它们难于访问。商务在Web上逐步拓展并打开了数据交换的通道，但是与传统数据系统的不兼容性严重阻碍了这一发展进程。开放源码运动带动了软件开发的蓬勃发展，但同时对一致的通信接口的需求也变得迫在眉睫。设计XML的目的是处理所有这些问题，以促进信息基础结构的发展。

本章将从多个角度介绍XML，包括XML是如何工作的，以及XML各部分是如何组合在一起的，这将作为进一步学习后续各章的基础。后续各章将更详细地介绍样式表、转换和文档模型。阅读完本书，你将会明白XML是如何协助管理信息的，以及该如何进一步学习。

什么是XML？

这个问题并不容易回答。从某个层次来看，XML是一个包含和管理信息的协议。从另一个层次来看，XML是可以完成从格式化文档到过滤数据等所有任务的一系列技术。从最高层次来看，XML是一种处理信息的方法，通过把数据提纯为最单一和最结构化的格式来获得数据最大的可用性和灵活性。对XML的透彻理解必须触及各个层次。

让我们从分析XML的第一个层次开始，即XML是如何使用标记来包含和管理信息的。这个通用的数据打包方案是下一个层次的基础，在下一个层次中，XML才是真正令人兴奋的，其中涉及各种附属技术，如：样式表、转换和自描述的标记语言。理解标记、文档和表示的基本概念有助于充分利用XML及其附属技术。

标记

注意，XML本身并不是标记语言，它只是一个创建标记语言的规则集。究竟什么是标记语言呢？标记（markup）是一些添加到文档中的信息，这些信息以某些方式来增强文档的含义，用于标识文档的各个部分以及各部分之间的关系。例如，当你阅读一份报纸时，可以通过各文章在页面中的间距和位置以及使用不同字体的标题来区分它们。标记的原理非常类似，只是标记使用符号，而不是空间。标记语言（markup language）是一个符号集，这些符号被放置在文档的文本中，以区分和标注此文档的各个部分。

标记对于电子文档来说是非常重要的，因为这些文档需要由计算机程序来处理。如果一个文档没有标注或边界，那么程序将无法把一段文本与其他文本区分开。实际上，该程序只好把整个文档当作一个单元来处理，这就严重限制了文档的用途。一份文章之间没有间距，而且只有一种文本样式的报纸本质上就是一个巨型的、令人厌烦的文本块。你或许可以区分出哪里是文章的末尾，哪里是文章的开头，但此过程的工作量是非常大的。计算机程序甚至无法这样做，因为它几乎缺乏最基本的模式匹配能力。

幸运的是，标记可以解决这些问题。下面是一个嵌有标记的文本段实例：

```
<message>
  <exclamation>Hello, world!</exclamation>
  <paragraph>XML is <emphasis>fun</emphasis> and
    <emphasis>easy</emphasis> to use.
    <graphic fileref="smiley_face.pict"/></paragraph>
</message>
```

此代码片段包含如下标记符号或标签（tag）：

- 标签 <message> 和 </message> 标记了整个 XML 代码片段的起始位置和结束位置。
- 标签 <exclamation> 和 </exclamation> 包围了文本“Hello, world!”。
- 标签 <paragraph> 和 </paragraph> 包围了一个更大的文本和标签区域。
- 几对 <emphasis> 和 </emphasis> 标签标注一些独立的单词。
- <graphic fileref="smiley_face.pict"/> 标签标记此文本中一个插入图形的位置。

从这个例子中可以看到一种模式：一些标签的功能类似于书挡，标记各个区域的起始位置和结束位置，而其他标签标记文本中的位置。即使如此简单的文档也包含有如下大量的信息：

边界（Boundary）

一段文本从一个位置起始，而在另一个位置结束。标签 <message> 和 </message> 定义了一段文本的起始位置和结束位置，同时定义了被标注为 message 的标记。