计算机技术译林精选系列

XML & SGML 参考手册

Rick Jelliffe 著

潇湘工作室 译

人民邮电出版社

计算机技术译林精选系列

XML & SGML 参考手册

◆ 著　　Rick Jelliffe
译　　潇湘工作室
责任编辑　李　际

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街 14 号
邮编　100061　　电子函件　315@ pptph.com.cn
网址　http://www.pptph.com.cn
北京汉魂图文设计有限公司制作
北京顺义振华印刷厂印刷
新华书店总店北京发行所经销

◆ 开本:787 × 1092　1/16
印张:33
字数:818 千字　　2000 年 10 月第 1 版
印数:1 - 5 000 册　　2000 年 10 月北京第 1 次印刷

著作权合同登记　图字:01 - 1999 - 0430 号
ISBN 7-115-08727-X/TP·1778

定价:65.00 元

版权声明

译者序

1998 年 2 月，W3C 正式批准了 XML（扩展标记语言）的标准定义，XML 可以在文本文档中标记结构，也就是说，它可以对文档和数据进行结构化处理，从而使它们能在部门、客户和供应商之间进行交换。现在，尽管 XML 尚未广泛使用，但由于它能让公司共享和使用分布在不同平台上的信息，因而它正在引起人们越来越多的注意。

利用 XML，可以实现动态内容生成、企业集成和应用开发。现在，人们还要支持过时的旧系统，所以需要合并多种旧式系统中的数据，在此方面，XML 可以为我们减轻负担。

本书是介绍 XML 系列书中的一本，本套书共有 3 本，分别是《XML 用户手册》、《XML & SGML 参考手册》和《XML 应用实例——建立电子商务应用》。本套书适用于利用 XML 进行各种应用（如网上银行、“推”技术、Web 自动化操作、数据库发布、软件销售等）的人员。本套书的主要内容有：

- 在《XML 用户手册》中，介绍 XML 的定义和作用，它与 HTML 的区别以及各种概念。在这里可以学到在现实世界中成功的 XML 应用以及富有经验的开发人员取得成功的秘诀，了解他们是如何应用 XML 来进行电子商务、内容管理、结构化、创建和表达的。
- 在《XML 应用实例——建立电子商务应用》中，循序渐进地介绍如何建立完整的 XML 电子商务应用程序。读者阅读这本书，可以理解 DOM 的关键作用，并可以看到 XML 在高级电子商务应用程序中的成功案例。
- 在《XML & SGML 参考手册》中，给出了在各种常见的编辑结构中应用的元素，它可以作为参考资料使用。另外，还给出了 XML、SGML、HTML、TEI、CALS 和国际化应用程序的实用技巧、警告信息以及解决方案。

在本书中，有建立 XML Web 应用的各种技巧、代码、工具和资源，它是一本很全面的参考资料。

参加本书翻译的人员主要有胡斌辉、彭少熙、王晓鹏、陶涛、滕杰、肖展业、孟丽艳等，另外，孙俊峰、贺民、龚亚平等也做了大量的工作。全书最后由贺军统稿。

由于水平所限，翻译错误和疏漏之处在所难免，请不吝指正。

潇湘工作室　2000 年 6 月

原书序

每位厨师都有其独到的烹饪方法。不必说，那就是可以增加菜肴滋味的酱油、高汤和调料的某种组合。

Rick Jelliffe 也不例外。他是非常棒的“厨艺”大师，他烹饪的是结构化的信息，也就是支持企业运转，以及支撑着电子商务和 World Wide Web 的数据和文档。10 年来，他一直在帮助世界各地的客户掌握结构并赢得对自己的信息处理的控制。

Rick 的秘密配料就是 SGML（用于结构化信息的国际性标准）和 XML（为 World Wide Web 设计的 SGML 的子集）。事实上，Rick 一直在开发这些标准的委员会中与我共事。

多年以来，Rick 收集了许多令人惊异的“烹饪”秘诀，并将在本书中与你分享。当然，其中包括了所有的基础知识，但在诸如东亚语言和国际化这样的主题中也有相当深奥的专家经验。

Rick 不仅是个“美食家”，而且还是个善于讲故事的人。为了帮助理解和保持概念，书中还点缀着迷人的历史和文化背景。

所以，请读一读本书，和 Rick 一起加入到结构化信息技术的盛宴中来。或者，如同他说的那样：

“Come on in and we'll throw another structure on the barbie!”（快点，我们来烤另一种结构）。

Charles F. Goldfarb

Saratoga

加利福尼亚

前　言

当我今天完成这本书时，我从网上获得了两份激动人心的新出版物，新鲜得甚至还留着激光打印机的余温：一份是 XML 1.0 的最终文本（World Wide Web 联盟的 SGML 子集，即 Extensible Markup Language（扩展标记语言）），另一份是 Web SGML（对 SGML 标准——ISO 8879——的修正和改进，像 XML 一样用于 WWW）。

这两份出版物共同掀起了电子出版业的革命，它们使得从前大型团体出版才可获得的 SGML 的优势现在在桌面上也可以获得了。最重要的是，它们令你——数据的拥有者和数据系统的创建者——可以控制自己的文档：SGML 及其针对 Web 优化的子集 XML 代表了 Open Systems（开放系统）运动的胜利。我们不仅能够获得用于网络协议和语言的开放系统，还能够获得激活它们的数据所用的开放系统。

本书是这个全新世界的一本实践指南——SGML 和 XML 元素类型集合和实体集合的构想和声明，这些集合能实现重要而有用的文档结构。我尝试着放入许多资源，这些资源用其他方法很难搜寻到。并且我特别注意了设法说清楚 XML 中更深入的模型，该模型对从 HTML 或其他专有标记语言转来的读者来说可能会觉得陌生：处理指令和特殊的符号。

这不是一本语法指南：指南类的书可以找到很多不错的，包括本系列中先前的几本。我特别留心避免引用 SGML 的声明，避免引用 SGML 中那些很少实现的、可选的功能。我把对一个领域的特别详细的论述包括了进来，该领域至今为止还几乎没有被详细地论述过：字符、字形和国际化。

一、规律、结构、模式和格式

本书是针对规律、结构、模式和格式的。在本书中：

- 规律（order）是指潜在的，抽象的（有时是难以言说的）关系和事物的本质。
- 结构（structure）是指如何用有形的标记来捕获规律。
- 模式（pattern）是指用于创建结构的一类模式或方法（例如，制衣意义上的“模板”不同于文字处理意义上的“模式”）。
- 格式（form）指一个结构和另一个结构之间特殊的一致性（例如，具

体事物层面意义上的“格式”不同于形而上学意义上的“格式”)。

和所有的好食谱一样，除了说明“如何”生成结构外，本书还会设法解释“为什么”，会探究有没有别的选择，会给出各种正面和反面的意见。

像 SGML 这样非常通用的技术的自由度可能会令新文档系统的设计者感到不适。能够向任何方向移动并不是那么舒服的，如果你不能容忍走错方向的可能！幸运的是，在 SGML 作为国际性标准存在的 10 年中，适用于许多常用的文档模板的集成方法和解决方案已经显现出来。本书尝试列出并讨论它们中最好的、最有教益的和最有用的方法。

二、文档系统

对系统的局限性和除了文档类型声明之外的各种因素的考虑常常成为决定工程是否成功的关键。正因为如此，本书更关注于“文档系统设计者”(document-system designer)而不仅仅是“DTD 编写者”(DTD writer)。我们不可能在真空中管理信息，文档是作为文档系统的一部分而存在的。有时系统是封闭的，有时则是开放的。如果这些信息有足够的价值，可以为管理工作提供保证，那么在创建一个大型的 DTD 时，为整个文档系统做考虑就会很有用。

无论如何，本书对制作 XML 文档的人们以及那些甚至从未把 SGML 当作符号来创建正式的元素类型声明的人们也有用处。如果你是这些人中的一员，我强烈推荐你学习并至少尝试着在非正式的文档中使用 SGML 内容模型：SGML 可提供非常方便并已得到好评的符号，它们适用于许多种类的结构，你还可获得图形化和可视化工具的帮助。

所以，即使可以在文档的任何地方找到规律，许多情况下结构也是松散的，其中包含着例外或者是不完全的结构。因此，本书中用于元素类型集合的模板可以视为原型和标本，你可以利用并改造它们来满足自己的特殊需要，而不是那种你只能顺从地剪切和粘贴的模板。

文档系统设计者需要知道 DTD 简洁性的局限。设计者认为可以作为著者原型的模板可能实际上并没有太大的用处。只有在能够揭示出一些实际规律，而决不是强加给一种似是而非的规律时，模板的使用才是成功的。

文档系统设计者往往具有拒绝混乱、爱好整洁的癖好，有时会为想在没有规律的地方找到规律而付出代价：那只是从先前的文档中得到的“幻象”。所以，本书除了给出各种模板之外，也为选择模板提供了一些原则。必须用成功的欲望来节制对简洁的要求。我希望那些翻开这本书并期望能够得到快刀斩乱麻般干净利落的解决方案的读者能提高自己的能力，你会理解最适合于自己的个人需求的那些问题并权衡利弊。

三、术语

XML 正在把大批人员从各种学科和技术领域带入 SGML 的世界，所以，目前术语还有相当大的变异和重复。为了控制叙述的方式，我使用了一些常见的经过简化的术语，它们着重于已使用的 SGML 关键字，如下表所示：

本书	ISO-ese
ANY 元素	已声明的内容类型为 ANY 的元素
CDATA 属性	已声明的值为 CDATA 的属性
CDATA 元素	已声明的内容类型为 CDATA 的元素
CDATA 实体	CDATA 实体
CDATA 标记段	CDATA 标记段
容器元素	具有子元素的元素
EMPTY 元素	空元素
ID 属性	已声明的值为 ID 的属性
IDREF 属性	已声明的值为 IDREF 的属性
NDATA 实体	NDATA 实体
NMTOKEN 属性	已声明的值为 NMTOKEN 的属性
RCDATA 元素	已声明的内容类型为 RCDATA 的元素
SDATA 实体	SDATA 实体
SUBDOC 实体	子文档实体

在本书中，“一个属性 ID”是指“一个名称为 ID 的属性”；“一个 ID 属性”是指“一个已声明的值为 ID 的属性”；但“该属性 ID”是指“示例代码段中名为 ID 的属性”。好的用法是“一个属性 ID”应该是“一个 ID 属性”，同样，“一个属性 IDREF”最好也是“一个 IDREF 属性”。

我打算在 Prentice Hall PTR 的 Web 站点 www.phptr.com 上维护一个 Web 网页，内容是本书的勘误表。

Rick Jelliffe
澳大利亚，悉尼

目　录

第一部分　文档系统

第三部分　字符和字形