XML
编程与应用开发教程

主编　汤华茂　王璐烽

电子科技大学出版社
University of Electronic Science and Technology of China Press

·成都·

前　言

XML 可扩展标记语言是标准通用标记语言的子集，是一种用于标记电子文件使其具有结构性的标记语言。在电子计算机中，标记指计算机所能理解的信息符号，通过此种标记，计算机之间可以处理包含的各种信息，比如文章等。XML 可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言，非常适合万维网传输，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。XML 是 Internet 环境中跨平台的、依赖于内容的技术，也是当今处理分布式结构信息的有效工具。

本书以技术需求为导向，以技术应用为核心，以任务驱动为主线，以应用开发为重点，以能力提升为目标，结合教学规律按照由浅入深、循序渐进、理论够用、实践为主的原则，精心设计、合理安排教学内容，全书共分为 11 章，课程内容及学时安排如下表所示。

章节	主要内容	课程学时	实践学时
第 1 章	XML 概述	2	4
第 2 章	XML 语法基础	4	
第 3 章	文档类型定义	4	4
第 4 章	XML Schema	6	4
第 5 章	XPath	2	2
第 6 章	XSLT	4	4
第 7 章	XML DOM	4	4
第 8 章	Java XML 编程	4	4
第 9 章	Web Service 基础	4	4
第 10 章	Java Web Service 开发	4	
第 11 章	基于 Web Servicer 的在线投票系统	6	10

本书的主要特点：

1. 以实践项目为依托，以任务驱动为主线，在设计、分析、完成任务的过程中使学生掌握相关理论及实践技能；

2. 图表结合、文字点睛、案例诠释，多年教学经历积累的大量实用案例，使学生能够即学即用；

3. 大量习题及上机实验，能够比较全面地检测学生相关理论及技能的掌握情况。

由于作者经验不足、水平有限，且时间较为仓促，书中不妥之处在所难免，恳请广大读者多加指正、不吝赐教，并将宝贵的意见反馈至作者的电子邮箱：cqtanghuaomao@163.com。

作　者

2017 年 12 月

目　录

第1章 XML概述

1.1 XML语言概述

1.1.1 XML语言简介

1.什么是XML语言

XML(eXtensible Markup Language),中文简称可扩展标记语言,是一种用于标记电子文件使其具有结构性的标记语言,方便信息或数据的传输和存储,它是SGML(标准通用标记语言)的子集。XML可以用来标记数据、定义数据类型,是一种允许用户对自己的标记语言进行定义的元语言,同时非常适合万维网传输,并提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。它既是Internet环境中跨平台的、依赖于内容的技术,也是当今处理分布式结构信息的有效工具。

以下是一个简单的XML文档实例,主要描述一本书的信息,这些信息包括书名、作者、出版日期和价格。

```
< ? xml version= "1.0"? >
< book>
  < title> XML教程< /title>
  < author> 张三< /author>
  < date> 20160101< /date>
  < price> 30< /price>
< /book>
```

2.什么是标记语言

标记语言,又称为置标语言,是用一系列约定好的标记来对电子文档进行标记,以实现对电子文档的语义、结构及格式的定义。这些标记必须很容易将内容区分,并且易于识别。

标记语言最早用于出版业，是作者、编辑以及出版商之间用于描述出版作品的排版格式所使用的。当今广泛使用的标记语言是超文本标记语言(HTML)和可扩展标记语言(XML)。标记语言广泛应用于网页和网络应用程序。

3. 标记语言的起源

为了促进数据交换和操作，在 20 世纪 60 年代，通过研究人员的杰出工作，IBM 公司得出了重要的结论：要提高系统的移植性，必须采用一种通用的文档格式，这种文档的格式必须遵守特定的规则。这也就是创建 GML(Generalized Markup Language，通用标记语言)的指导原则，从人们所产生的将文件结构化为标准格式的动机出发，IBM 创建了 GML。

在对标记语言的概念达成共识的基础上，IBM 公司的研究人员 Charles Goldfarb 带领的开发团队完善了 GML，将其称为 SGML(Standard Generalized Markup Language，标记通用标记语言)。SGML 成为 IBM 内部格式化和维护合法化文件的手段，后来经拓展和修改，作为一种全面的信息标准以适应工业范围的广泛应用。1986 年，SGML 被国际标准化组织(ISO)所采纳。

1989 年，欧洲粒子物理实验室(CERT)的研究员 Tim Berners-Lee 和 Anders Berglund 共同创建了一种基于标记的语言 HTML，它可看作 SGML 的简单应用。后来由 IETF(The Internet Engineering Task Force，国际互联网工程任务组)用简化的 SGML(标准通用标记语言)语法进行进一步发展，最终成为国际标准，由 W3C(World Wide Web Consortium，万维网联盟)维护。

1996 年，人们开始致力于描述一个新的标记语言，它是一种在 WEB 中应用 SGML 的灵活性和强大功能的方法，W3C 成立了专家小组从事这项工作。1998 年 2 月，W3C 批准了 XML1.0 规范。XML 具备 SGML 的核心特性，但更加简洁，它的内容甚至不到 SGML 的十分之一。

1.1.2 HTML 与 XML 的区别

HTML 取得了令人难以置信的成功，但是它的应用范围受到限制，只适用于在浏览器中显示文档。HTML 文档里的标签并没有提供与标签之间的内容有关的信息，提供的只是告诉浏览器如何显示标签之间的内容的指令。

1. HTML 被设计用来显示数据

HTML 文档中的所有标记对于人来说没有任何意义，其语义信息仅仅是告诉浏览器以何种方式显示信息，HTML 文档所表达的信息和含义只有通过浏览器转换并显示出来人们才能理解(图 1.1)。HTML 关注的重点是：显示数据以及如何更好地显示数据。

```
<!doctype html> <html lang="en"> <head> <meta charset="UTF-8"> <title>Document</title> </head> <body> <table border="1"><tr> <td>收件人</td> <td>张三</td> </tr> <tr> <td>发件人</td> <td>李四</td> </tr> <tr> <td>主题</td> <td>开会提醒</td> </tr> <tr> <td>内容</td> <td>别忘记了下午3点钟开会哦！</td> </tr> </table> </body> </html>
```

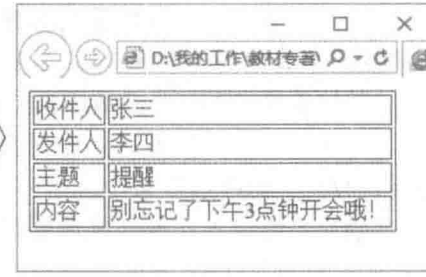

图 1.1　HTML 语义

2. XML 被设计用来描述数据

XML 文档中的数据是通过自定义标签以一种有意义和自描述的方式进行描述的(图 1.2)。自定义标签经领域专家精心选取，体现了人们的共识。例如，标签＜收件人＞对于人意为信件接收者，这样就可以推断标签中包括的数据是关于消息或信件接收者的信息。因此 XML 标记本身和 XML 的文档结构蕴含着一定的含义，这些标记对人来说是有意义的。XML 关注的重点是：什么是数据，如何描述或存放数据。

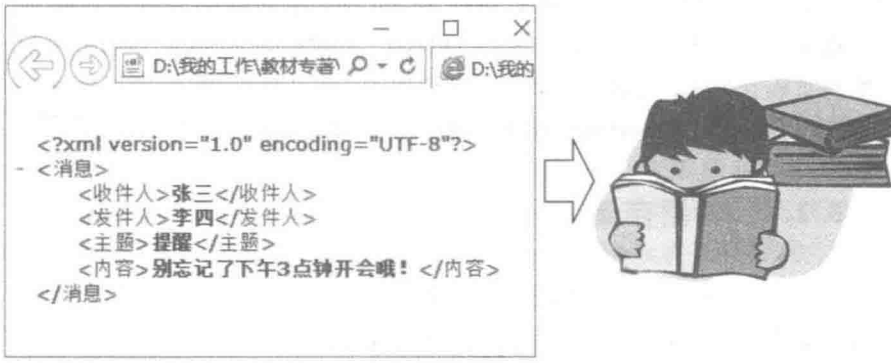

图 1.2　XML 语义

XML 不是 HTML 的替代，XML 和 HTML 的设计目的不同，主要差异为：

- XML 被设计用来传输和存储数据，其焦点是数据的内容；
- HTML 被设计用来显示数据，其焦点是数据的外观；
- HTML 旨在显示信息，而 XML 旨在传输信息。

1.1.3 XML 语言的特点

(1)易用性：XML 可以使用多种编辑器来进行编写，包括记事本等所有的纯文本编辑器。

(2)结构性：XML 是具有层次结构的标记语言，包括多层的嵌套。

(3)开放性：XML 语言允许开发人员自定义标记，这使得不同的领域都可以有自己的特色方案。

(4)分离性：XML 语言将数据的显示和数据的内容分开保存，各自处理。这使得基于 XML 的应用程序可以在 XML 文件中准确高效地搜索相关的数据内容，忽略其他不相关部分。

1.2 XML 应用

1.2.1 数据分离

XML 可以从 HTML 中分离数据。通过 XML，你可以在 HTML 文件之外存储数据。在没使用 XML 时，HTML 用于显示数据，数据必须存储在 HTML 文件之内；使用了 XML，数据就可以存放在分离的 XML 文档中。这种方法可以使人集中精力去使用 HTML 做好数据的显示和布局，同时也确保了数据改动时不会导致 HTML 文件也需要改动，这样可以方便维护页面。XML 数据同样可以以“数据岛”的形式存储在 HTML 页面中，使人可以将精力集中到使用 HTML 格式化和显示数据上去。

1.2.2 数据存储

XML 独立于硬件、软件以及应用程序，具有很强的跨平台可移植性，并且数据无须转换，所以 XML 也常用来存储数据。不仅仅在 HTML 页面中能访问 XML 数据，应用程序也可以从 XML 数据源中进行访问。XML

常被作为应用系统的配置文件供应用程序使用。通过 XML，我们的数据可以供各种阅读设备使用（如手持计算机、语音设备、新闻阅读器等），还可供盲人和其他残障人士使用。

1.2.3　数据交换

数据交换是指数据在不同的信息实体（如硬件平台、操作系统、应用软件）之间相互发送、传递的过程。实行数据交换的不同信息实体必须统一建立一种数据传输的标准格式，因此在数据交换过程中会涉及不同数据格式之间的转换和适配。XML 的很多特性使其成为数据交换领域事实上的标准。通过 XML，可以在跨平台、不兼容的系统之间轻松地交换数据。

首先，XML 使用元素和属性来描述数据。在数据传输过程中，XML 始终保留了数据之间的关系和结构。应用程序之间可以共享和解析同一个 XML 文件，不必使用传统的字符串解析或拆解过程。使用 XML 交换数据可以使应用程序更具有弹性。

另外，XML 还能够简化数据共享。在真实的世界中，计算机系统和数据使用不兼容的格式来存储数据。而 XML 数据以纯文本格式进行存储，因此提供了一种独立于软件和硬件的数据存储方法。这让创建不同应用程序可以共享数据变得更加容易。

1.2.4　系统集成

在计算机高速发展的进程中，企业和政府为提高办公和生产效率、简化办事和处理流程，建立了众多的业务系统，例如，企业资源管理系统、产品数据管理系统、办公自动化系统、保险业务系统、税务系统、公安人口管理系统等。在系统建设完成的初期，这些系统确实为办公效率的提高、日常业务处理的便利发挥了重要的作用。由于建设初期各种资源和技术上的限制，各单位、各系统各自为政，虽然越来越多的业务系统被开发和应用，人们可获取的信息越来越多，这些数据的价值也越来越为人们所认识，但是，数据以不同的格式分散存放在不同的业务系统和不同的数据库系统中，这些资源还是不能被有效地利用，这样就形成了众多的“信息孤岛”。

随着信息化进程的逐步深入和社会的不断进步，政府需要各个部门的协同配合以提供更灵活、方便的综合信息服务，企业之间也需要协作完成产

品设计和生产制造。于是，这种最初建设的“信息孤岛”式业务系统也就慢慢地不能完全满足企业和政府的需要了。“信息孤岛”现象现在已经成为信息化建设的瓶颈，要解决“信息孤岛”现象，就必须实现各个业务系统间的互联互通、信息共享和系统集成，而解决这些问题的关键在于如何在各系统间进行有效的数据交换和共享。

XML 语言具有适宜异构应用间的数据共享，可以进行数据检索，XML 文档本身的节点就是一种由若干节点组成的数据结构，这种特点有利于高级语言通过调用 XML 编程接口访问 XML 节点，而且 XML 能通过网络进行传输。此外，XML 的 DTD 是 XML 词汇形式和完整性定义的理想描述技术，可以提供系统的一致性约束和正确性验证。所有这些优点使得 XML 成为目前绝大多数信息集成框架的首选方法。

1.2.5 内容管理

随着 IT 应用的深入普及，各行各业都积累了大量的信息资源。科学管理和合理开发这些内部和外部信息资源已经成为企业正确决策、增强竞争力的关键。研究部门调查发现，在企业存储的大量数据中，传统关系数据库管理系统(RDBMS)处理的结构化数据仅占数据信息总量的 15%，而全球 85%的信息是非结构化的，包括纸上的文件、报告、视频和音频文件、照片、传真件、信件等。如何管理这些非结构化信息是传统结构化数据管理的一大难题。

企业内容管理就是随着数据管理的发展而为客户提供的一种应用软件，它管理、集成和访问从音频、视频到扫描图像的各种格式的商业信息。内容管理处理的对象范围比传统关系数据库管理系统(RDBMS)处理的结构化数据更广，除了一般文字、文档、多媒体、流媒体外，还包括 Web 网页、广告、程序(如 JavaScript)、软件等一切数字资产(Digital Asset)，即所有结构化的数据和非结构化的文档。内容管理解决方案重点解决各种非结构化或半结构化的数字资源的采集、管理、利用、传递和增值，并集成到结构化数据的信息系统中，如 ERP、CRM 等，从而为这些应用系统提供更加广泛的数据来源。

内容管理的顺畅有赖于内容的结构化，因为只有结构化，才能对内容分类、索引、排序和搜寻。利用 XML 相关工具来制作结构化的内容，正是内容管理的基础建设。通过数据转换方式，把原底层原有数据转换成 XML 格式，作为与别的系统衔接沟通的共同语言，再由一个信道把这些系统串联起来，把内容连接起来管理。这样，下层的原有运行的系统与数据，不论分

散在什么地方、也不论什么格式，都可以维持不动、继续运行。概括来说，内容管理是将现有各个底层的数据，建为一个共同的目录控管机制，各个系统处理数据的软件，都依此目录配送，使数据流动横跨各系统，既不必制造一个集中的庞大数据库，也不必更改现有系统的运行。

1.2.6　电子商务

电子商务是经济全球化和贸易自由化的重要手段，也是传统产业变革和企业实现技术跨越的关键推动力，已成为各国政府为增强国家竞争力、赢得市场资源配置优势而大力推进的战略性任务。电子商务不是一个单纯的技术问题，而是一个跨国界、跨地区、跨行业、跨学科、跨领域的系统工程。标准化在其中起着协调和统一有关技术问题、更新经营观念、确立市场运营的技术规则、连接电子商务的各个环节的作用，确保其协同工作，使之有序、高效、快速、健康发展的作用。

电子商务包括两大类：一类是电子数据交换(EDI)，另一类是基于 XML 的电子商务。联合国贸易便利化与电子业务中心(UN/CEFACT)将 EDI 定义为在增值网上一种电子数据传输方法，用这种方法，首先将商业或行政事务处理中的报文数据按照一个公认的标准形成结构化的事务处理的报文数据，然后将这些结构化的数据经由网络，从一个计算机传输到另一个计算机。

由于基于 XML 的电子商务是在互联网上进行，因此，必须为它的运行建立一套规则，即标准。这些标准包括基于 XML 的电子商务的网络标准、处理标准、数据标准和语义语法标准等。从当前工作重点上看，基于 XML 电子商务标准主要解决数据共享、业务协同、安全保密三大问题，即着重于以互联网为主要通信设施，以 XML 为信息描述语言，以业务交易数据语义、电子文档格式、业务过程、消息服务等为核心内容，并面向特定电子商务模式的综合性标准化解决方案的研制方面。

1.2.7　创建新语言

很多新的 Internet 语言是通过 XML 创建的，例如：

XHTML，最新的 HTML 版本，XHTML 是更严谨更纯净的 HTML 版本。它的可扩展性和灵活性将适应未来网络应用更多的需求。

WSDL，网络服务描述语言，是 Web Service 的描述语言，它包含一系列描述某个 Web Service 的定义。

WML，无线标记语言，WML 是专门为手持式移动通信终端（手机）设计的标记语言。

RSS，简易信息聚合，是一种描述和同步网站内容的格式。RSS 目前广泛用于网上新闻频道，blog 和 wiki，主要的版本有 0.91，1.0，2。使用 RSS 订阅能更快地获取信息，网站提供 RSS 输出，有利于让用户获取网站内容的最新更新。

SVG，可缩放矢量图形，是一种描述二维图像的语言。它主要是一种向量图形语言，提供了一种实用、灵活、使用 XML 表示的图像格式，可以以文本的方式，轻松、实时地创建各种图形。

VoiceXML，语音扩展标记语言，是一种基于 XML 的因特网标记语言，用于开发语音用户界面。它是“语音 Web”使用的语言，使得用户可以使用电话来访问因特网的内容，可将其视为用于电话的 HTML。利用 VoiceXML 可以建立基于 Web 的语音应用和服务。VoiceXML 为语音应用领域展现了一个广阔的未来，在语音门户、语音呼叫中心（Call Center）、语音信息服务、语音电子商务等领域有着广泛的应用。

1.3 XML 开发工具

Altova XMLSpy 由 Altova 公司开发的符合行业标准的 XML 开发环境，是业内最畅销的 XML 编辑器和开发环境，用于建模、编辑、转换并调试所有与 XML 相关的技术。该开发工具提供全球领先的图形图解设计工具、代码生成器、文件转换器、调试器、剖析器以及完整数据库集成，支持 XSLT、XPath、XQuery、WSDL、SOAP、XBRL 和 Office Open XML（OOXML）文档，并提供 Visual Studio 和 Eclipse 插件等。

1.3.1 认识 Altova XMLSpy 2010

Altova XMLSpy 2010 的图形用户界面主要由菜单栏、工具栏、主窗口、Project 窗口、Info 窗口、Message 窗口以及输入助手窗口等组成（图 1.3）。

主窗口：显示正在编辑的文档的窗口，可用的文档视图数目与正在编辑的文档类型有关。可以根据需要在各种视图间进行切换。

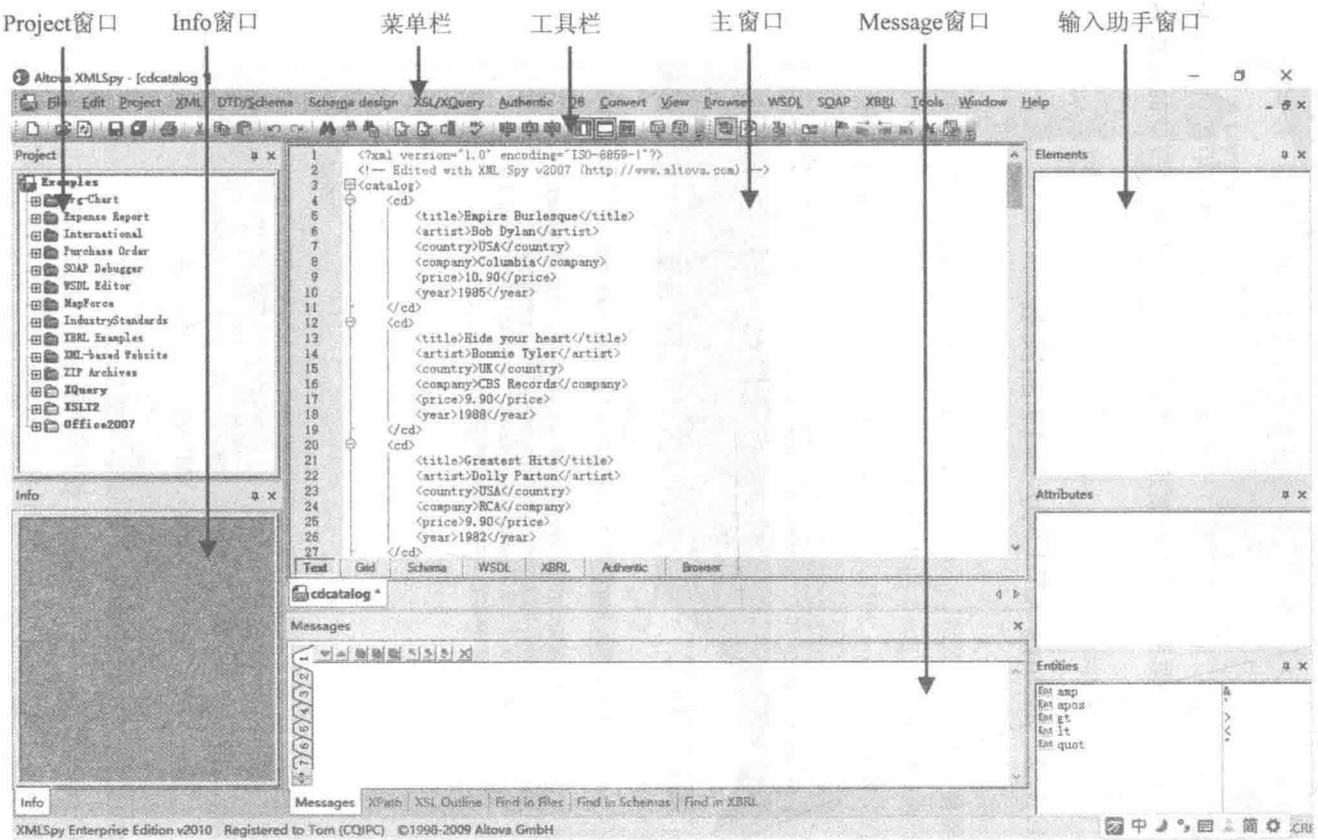

图1.3 Altova XMLSpy 2010的图形用户界面

Project 窗口：在该窗口中将文件组织为工程，并可对这些文件进行编辑。

Info 窗口：在该窗口中显示当前编辑项的信息。

Message 窗口(Entry Helper)：显示当前文件在语法检查、有效性验证等时的错误信息。

输入助手窗口：输入助手窗口泛指那些在文档编辑过程中提供帮助的窗口，可用的输入助手窗口将根据正在编辑的文档类型和主窗口的文档视图的不同而变化。

1.3.2 Altova XMLSpy 2010 的安装

要使用 Altova XMLSpy 2010，首先必须将它安装到本地计算机上。安装 Altova XMLSpy 2010 的步骤如下。

(1)双击“XMLSpy 2010.exe”的文件图标，系统将打开安装向导(图 1.4)。

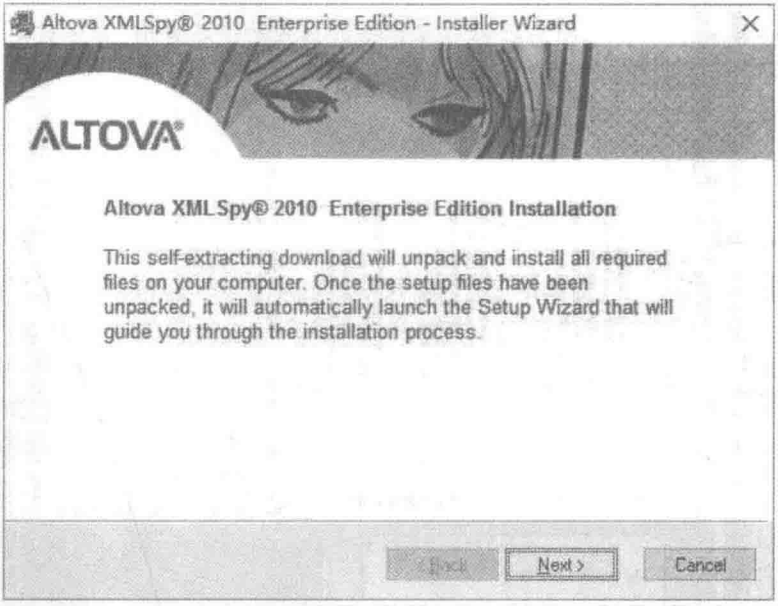

图 1.4 Altova XMLSpy 2010 安装向导

(2)单击“Next”按钮，将显示安装对话框(图 1.5)。

(3)单击“Next”按钮，将显示软件许可协议对话框(图 1.6)，在该对话框中显示出了许可协议的全文，用户必须同意该协议的所有条款才可以使用该软件。选中“I accept the terms in the license agreement and privacy policy”的单选按钮。

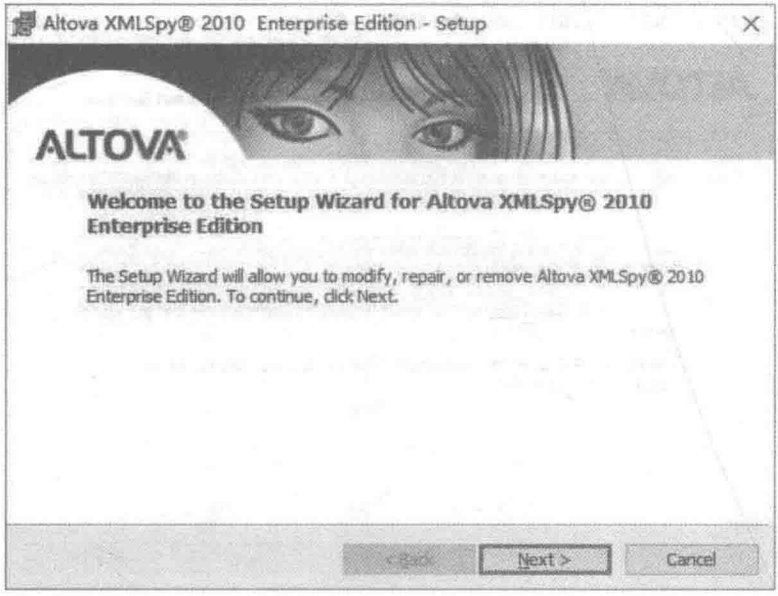

图 1.5　Altova XMLSpy 2010 安装对话框

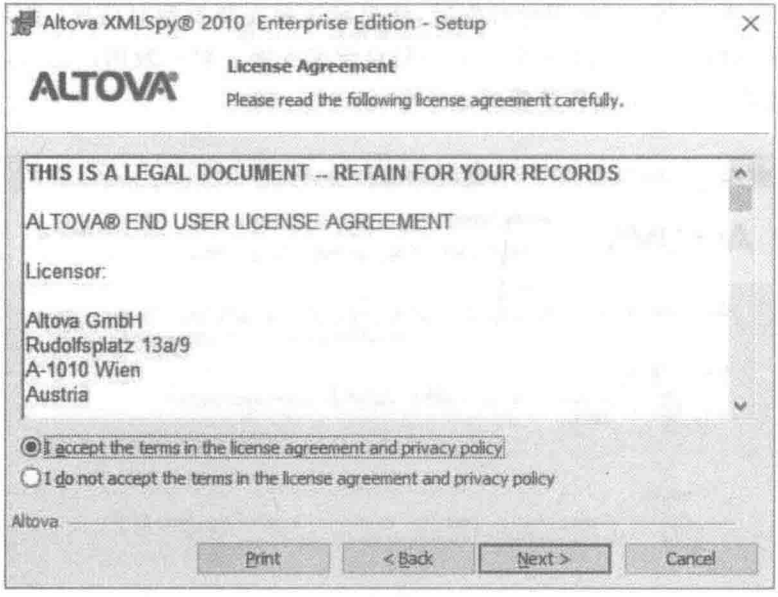

图 1.6　软件许可协议对话框

(4)单击“Next”按钮继续安装,将显示打开和编辑的文件类型关联对话框(图 1.7),在该对话框中可以选择与 Altova XMLSpy 关联的文件类型。