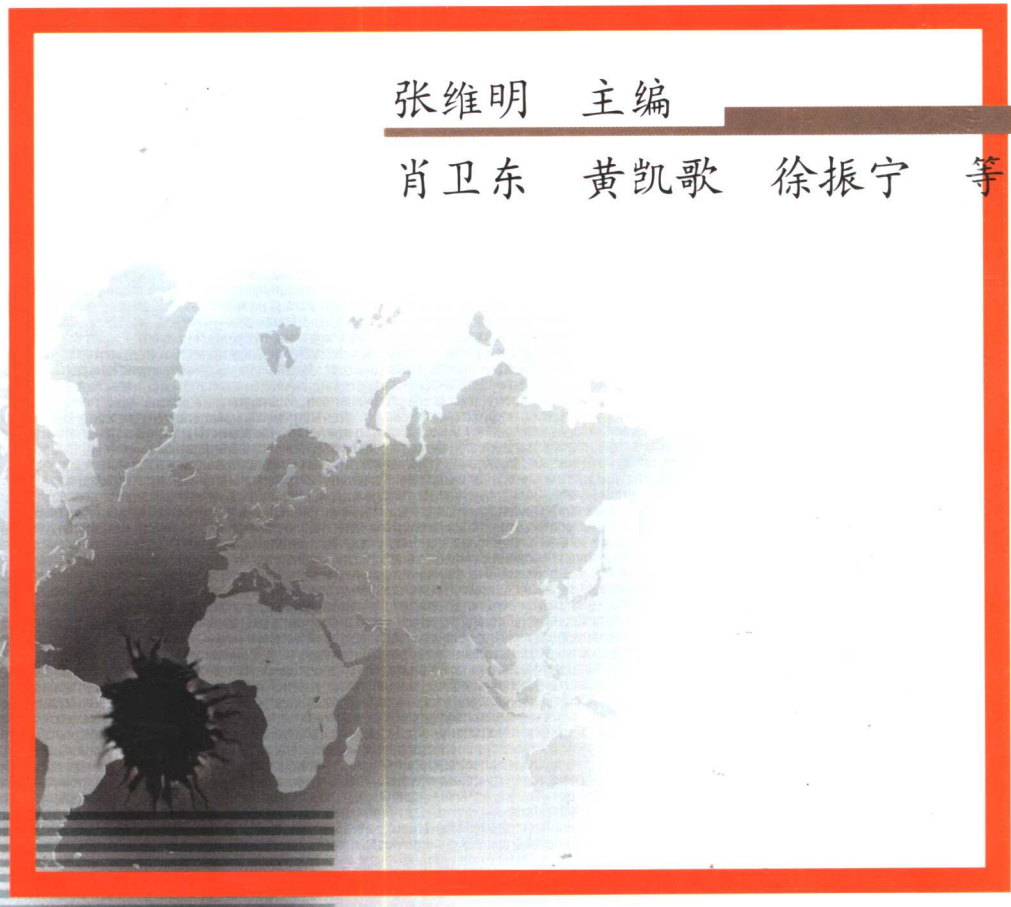
信息系统工程丛书

语义信息模型及应用

张维明　主编

肖卫东　黄凯歌　徐振宁　等编著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

http://www.phei.com.cn

信息系统工程丛书

语义信息模型及应用

张维明　主编
肖卫东　黄凯歌　徐振宁　等编著

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

内 容 简 介

随着 Internet 的发展和信息共享需求的不断提升，语义信息模型必将成为 Internet 上的主流信息模型。作为语义信息模型的具体实现手段之一，XML 将为许多问题提供崭新而有效的解决思路和方法。

本书试图从语义信息模型的基础知识入手，以 XML 为具体手段介绍语义信息模型及其应用。全书可分为两部分，第一部分由前 4 章组成，重点介绍语义信息模型和 XML 的基本知识，第二部分由后 4 章组成，介绍 XML 在人工智能、分布式计算、数据库和 Web 服务方面的应用。

本书既可作为高等院校信息管理相关专业硕士研究生课程的参考书，也可以作为广大科技工作者和研究人员的参考书。

图书在版编目(CIP)数据

语义信息模型及应用/张维明主编．—北京：电子工业出版社，2002.3
(信息系统工程丛书)
ISBN 7-5053-6896-6

Ⅰ.语… Ⅱ.张… Ⅲ.语义信息-模型 Ⅳ.G201

中国版本图书馆 CIP 数据核字(2002)第 009457 号

策划编辑：秦 梅
责任编辑：秦 梅 李 萌
印 刷：北京东光印刷厂
出版发行：电子工业出版社 http://www.phei.com.cn
北京市海淀区万寿路 173 信箱 邮编 100036
经 销：各地新华书店
开 本：787×1092 1/16 印张：13.25 字数：339.2 千字
版 次：2002 年 3 月第 1 版 2002 年 3 月第 1 次印刷
印 数：4 000 册 定价：27.00 元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系。
联系电话：(010)68279077

《信息系统工程丛书》编委会

《语义信息模型及应用》编写人员名单

主编 张维明

编著 肖卫东 黄凯歌 徐振宁 汤大权 曹泽文 李勇 李由 宋剑峰

丛 书 序

从现实世界的角度看，客观世界是由物质、能量和信息三大基本要素组成的，人类的社会生活每时每刻都离不开信息。从远古时代开始，人类就一直在同信息打交道，围绕着信息形成了不同的信息作业，包括信息的采集、存储、表示、传递、加工处理、检索利用和控制等，所有这些环节形成了信息系统，并作为客观世界每个系统的一个子系统或显式或隐式地存在着。

从科学技术的角度看，信息系统是在20世纪中叶由信息科学、计算机科学、管理科学、决策科学、系统科学等学科相互渗透交叉而发展起来的，经过多年的研究目前已经形成了比较完整的独具特色的体系。信息系统工程是20世纪80年代出现的以建立信息系统为目标的新兴学科，它是用系统工程的原理、方法来指导信息系统建设与管理的一门工程技术学科，主要研究各级各类信息系统建设和管理中的规律性的问题。它既不是“信息的系统工程”，也不是“信息系统的工程”，而是“信息系统的系统工程”。一般认为，信息系统工程的目标是为以计算机和其他信息技术为手段的各类信息系统提供科学的开发方法、管理手段及有关的工具、标准、规范，通常不包括通信工程、信号处理等具体学科领域的技术。

信息系统工程的研究范围主要包括：

(1)信息系统的基本理论。信息系统的基本观点、认识论和方法论等。

(2)信息系统建模。信息系统概念模型、逻辑模型和物理模型的描述、观察、试验与验证等。

(3)信息系统开发。信息系统建设与管理的概念、方法、评价、规划、工具、标准等一系列相关技术问题和工程问题。

(4)信息系统支撑技术在信息系统中的应用。数据库/数据仓库、网络通信、人机交互、分布计算、决策支持、人工智能等技术如何满足信息系统各层次用户的需求，实现业务管理、信息共享、分析决策等功能，并在组织和人的参与下最终达到信息系统的目标。

(5)信息系统集成。研究系统集成的原则、方法、技术、工具和有关的标准、规范，应用先进的相关技术，将支持各个信息“孤岛”的小运行环境，集成统一在一个大运行环境中，最终形成一体化的信息系统。

《信息系统工程丛书》是由国防科技大学管理学院组织多位专家和科研人员面向信息系统工程专业撰写的教材类图书。作者所在的单位是20世纪70年代末在钱学森院士的亲自倡导下建立起来的，在国内最早开设了信息系统工程专业。作者长期从事信息系统工程方面的教学、科研和开发，这套丛书是其多年学术研究和科技开发的成果总结，也是其多年教学工作中的实践积累，从丛书体系的设置到内容的安排，都基本体现了对当今信息系统工程领域前沿技术的把握。

这套丛书准备分批出版，第一批由《信息系统原理与工程》、《信息系统集成技术》、《信息系统建模》、《多媒体信息系统》、《智能协作信息技术》、《数据仓库原理与应用》、《语义信息模型及应用》7部教材和专著组成，再加上该单位近年已出版的《决策支持系统技术》和《智能决策支持技术》两部研究生教材，基本上已覆盖了上述的信息系统工程主要研究范围。其中：

《信息系统原理与工程》主要介绍信息系统的基本概念、基本原理、技术和设计开发方法。

具体包括信息系统与信息系统工程的基本概念，信息系统中的基础理论、开发方法，结构化系统分析、系统设计和面向对象的分析设计方法，信息系统战略规划，系统实施，信息系统对计划、控制、决策的支持，计算机辅助信息系统开发等。

《信息系统集成技术》主要介绍信息系统集成的基本概念、基本原理和设计开发方法。首先介绍信息系统集成技术的发展，然后从体系结构入手，分网络集成、数据集成和应用集成三个层次展开对信息系统集成的论述，并给出了系统集成的案例。

《信息系统建模》主要介绍信息系统建模的基本概念、基本原理、方法、工程技术与工具。具体包括面向信息系统建模的思想，需求建模，逻辑建模，对象建模，Agent 建模，数据建模，统一建模语言等，是国内第一部按照较完整的体系专门介绍信息系统建模技术的著作。

《多媒体信息系统》主要介绍多媒体信息系统的基本概念、原理、技术和应用，主要内容包括多媒体信息系统的体系结构和数据模型、多媒体数据库和信息管理、多媒体通信和网络、多媒体人机交互与表现技术、原型系统与应用等。

《智能协作信息技术》主要介绍智能协作信息技术及系统的基本概念、基本原理和设计开发方法。具体包括智能协作信息技术的发展概况，智能主体概念、性质、内部结构和实现方法，多智能主体协作的基本原理、实现技术等，还介绍了智能协作信息系统的开发方法和智能协作信息技术在工业、管理、办公自动化等领域的应用。它是国内第一部全面介绍智能协作信息技术和智能协作信息系统的专著。

《数据仓库原理与应用》主要介绍数据仓库的概念、基本原理、规划、开发方法以及相关算法，包括数据仓库的发展、技术体系、元数据管理、分析设计方法和开发工具，并对数据开采的主要理论和方法、联机分析等应用技术作了深入的阐述，是一本理论与实践相结合的教材，是国内较为全面地分析数据仓库、开发数据仓库的书籍。

《语义信息模型及应用》深入到目前信息管理领域的前沿，探讨了语义信息模型的基本概念，并以 XML 为具体实现手段介绍了语义信息模型在信息组织、信息处理、信息服务、信息交换等方面高级应用的原理与实现机制。

除《语义信息模型及应用》以外，丛书中所有教材都作为内印教材或讲义试用过多次，吸收了许多专家学者以及学生的意见。

这套丛书既能够使广大读者从整体上把握知识结构、理清相关技术领域的关系和分类，又能够从中找到每项具体理论、技术、方法、工具的介绍和例解，再加上融合了多项“九五”期间的高水平科研成果，应该使这套丛书具有较高的系统性和实用性。

《信息系统工程丛书》是一套理论与工程实践并重的著作，它不仅可作为相关专业的大学本科生和研究生的系列化教材和参考书，而且也可以为从事信息系统工程的科研人员提供参考。我们相信，这套丛书的出版，将对我国信息系统工程的全面、深入发展起到重要的推动和促进作用。

《信息系统工程丛书》编委会

2001 年 6 月

前　言

在整个人类文明发展的历史上，对话、交流与理解的历程一直伴随着信息技术的发展步伐。Internet 的产生和发展，使全人类的交流更加开放。从目前的状况来看，Internet 要想完全发挥其威力，不仅在于更快的 CPU 和更宽的带宽，还在于建立一种更有利于交流与对话的机制，在于发展一种更有利于相互理解的基础技术。这种技术必须从最基本的信息表示和交换开始，排除一切平台和语言的分歧，以自由、平等、开放为原则，以人类对现实世界的一致理解为基础，为全人类提供一种全新的高质量的信息服务。

在因特网发展过程中，有三大技术起了决定性的作用：第一是分组交换技术与中介信息处理器（Interface Message Processor）的发明，使分布式网络——APARnet（Internet 的前身）得以诞生；第二是 TCP/IP 协议的提出与实施，使 APARnet 扩展延伸，数据传输畅通无阻；第三是 HTML 与 WWW 的出现，使得一个全球最大的信息资源利用系统诞生了。我们仔细分析一下可以看出，分组交换技术与中介信息处理器的发明使得物理层的扩展成为可能，TCP/IP 协议的提出与实施在网络层统一了机器交互的语法。这两层属于信息基础设施，技术已趋成熟，不管怎样总是朝着超高带宽发展，但有一点是可以肯定的，那就是对应用层的透明性。目前，让 Internet 完全发挥威力的努力更多地针对应用层，而应用层要从根本上得到发展，只有走统一语义的道路。有了基于语义的信息表示、信息交换、信息处理技术，有了对标准化道路的一致共识，我们对于这个宏伟的目标就有了前所未有的信心。

全书共分 8 章，主要论述以 XML 为代表的语义信息模型及其在各个相关领域的应用。全书共分两大部分。第一部分由前 4 章组成。第 1 章概述语义、语义信息模型和 XML 的基本概念。第 2 章介绍语义信息模型的关键基础——资源描述框架 RDF。第 3 章探讨利用本体技术表示和处理语义的方法。第 4 章介绍语义 Web 的结构和思想，展现了语义信息模型在下一代 Web 中的应用。第二部分由后 4 章组成。第 5 章介绍 XML 在人工智能领域中的一些研究进展和成果。第 6 章介绍 XML 的语义消息在分布式计算中的应用。第 7 章分析 XML 与关系数据库的互相映射方法。第 8 章介绍 XML 在 Web 服务中的作用。

本书是作者多年潜心研究和开发工作的总结，是研究小组所有师生集体智慧的结晶。由于语义信息模型还正处在研究阶段，加之作者水平有限，难免在完整性、准确性等方面存在着问题。我们迫切希望能够得到广大读者特别是专家和同行的指点，共同探讨有关问题，交流研究经验，使我们的研究工作取得进步。

作　者

目　录

第1章　语义信息模型概述

在整个人类文明发展的历史上，对话、交流与理解的历程，就是整个人类文明发展的历程；对话、交流与理解的程度，代表着人类文明发展的程度。

纵观IT业的发展历史，有一条由封闭走向开放的明确主线，开放的程度代表着IT业发展的水平。尤其是Internet的发展，使IT技术开放的步伐加大、加快了，改变了整个IT界的思路，“开放”的号角与“标准化”的鼓声响彻IT技术的每一个角落。

Internet要完全发挥出其威力，不仅在于更快的CPU，不仅在于更宽的网络带宽，还在于建立一种更有利于交流与对话的机制，在于发展一种更有利于相互理解的基础技术。这种技术必须从最基本的数据表示和信息交换开始，排除一切平台、语言的分歧，以自由、平等和开放为原则，以人类对现实世界的一致理解为基础，为全人类提供一种全新的高质量的信息服务。

有了基于语义的信息表示、信息交换和信息处理的技术，有了对标准化道路的一致共识，这些宏伟的目标正逐步变为现实。

1.1　信息模型

在过去数年里，“信息模型”这一术语频繁地出现在各种与信息技术相关的文档中。美国科学家联合会对“信息模型”的解释为：信息模型用于描述组织内的信息资源以及这些资源的相互关系。信息模型用于支持数据建模、生成数据库和文档存储的设计需求，提供数据体系结构的信息资源管理者视图(TAFIM 3.0)。

信息作为一种资源是人们已经逐步达成的共识。物质、能量和信息都是人类可以利用的战略资源，它们都可以用来制造生产工具。物质资源可以被加工为材料，但是仅利用材料就只能制造出人力工具。能量可以被转换为动力，它与材料结合起来就可能制造动力工具。而信息可以被提炼成为知识，它与动力和材料相结合，则可以制造出智能工具。人力工具可以支持农业社会的生产力，动力工具可以支持工业社会的生产力，而智力工具则可以支持信息社会的生产力。

在信息社会中，各种组织为了从容地应对频繁变化的环境，需要准确而迅速地掌握组织自身基础结构和环境究竟发生了哪些变化，这就必须要了解反映自身基础结构以及采取相应行动所必需的信息以及它们之间的关联。为此，组织内部成立专门的部门，开发大量的信息系统来解决这些问题。那么究竟要管理哪些信息资源呢？

“信息”以及“数据”的定义可谓仁者见仁，智者见智，加起来不下数十种。具体到信息系统中，参考IDEF1X的观点应该最具有实际指导意义。在美国，政府强制要求信息系统建设时必须采用IDEF1X方法进行数据建模。IDEF1X方法认为“信息”是为了某个特定目的或在一定范围内聚合起来的数据集，而数据可以被认为是有含义事件的符号表示，单一的含义能够被用于多个不同的事件。

以上的定义实际上反映了对信息构成的一种理解，即信息是由数据构成的，而数据是由事实和其代表的含义构成的。同时也表明，建立信息模型管理信息资源的目的和重点应该放在

被管理数据应用在事件的含义上,或者说放在数据所代表的语义上。

1.2 语义信息模型

这一节我们对语义信息模型进行介绍。有些观点是我们在研究中逐步形成的。希望能和专家读者进行探讨。

1.2.1 什么是语义信息模型

我们认为数据库的三模式结构是外部模式、内部模式和概念模式。但是在过去,系统分析员们定义的模式结构大多只是外部模式和内部模式这两个模式,留下一个概念模式在自己的头脑中或不规范的记录中。

概念模式是数据的一个单一集成定义。在一个组织中,它不偏向于任何专门的数据应用,同时还独立于数据的物理存储和存取方式。这个概念模式的主要目标是提供一个数据的含义和相互关系的一致定义,从而用来集成、共享和管理数据的完整性。

从直观上讲,可能不会有人反对将数据库概念模式称为一种语义信息模型。但是同"信息模型"等基本术语一样,目前"语义信息模型"也没有一种公认的定义,甚至对它的定义也不多见。这种情况的出现并不意味着对"语义信息模型"的研究不重要,而可能是由于这样的一些考虑:任何信息模型都或多或少地描述或隐含了信息资源的语义,从而不需要再单独强调语义信息模型。但同时,我们都明确地知道概念模式、内模式和外模式是不同类型的模式,解决的问题不同,包含的语义不同等。实际上语义数据模型化技术的发展也开始于从概念视图定义数据的要求。

显然,语义信息模型有不同于一般信息模型的特征。归纳出这些特征,能够在一定程度上明确语义信息模型的内涵。语义信息模型必须具有以下的基本特征:

(1) 语义信息模型是与问题域的基础结构相一致,能够覆盖问题域当前应用范围的信息模型。

(2) 语义信息模型是明确地描述组织的基础结构,能够转化为多种视图和存储结构的信息模型。

(3) 语义信息模型是明确地记录对问题域基础结构进行概念化时所作的约定、假设和依据等的信息模型。

(4) 语义信息模型是可扩展的信息模型。

可以看出,这里所指的"语义"不同于语义学中的定义。从语义学的角度讲,语义是语言形式表达的内容。语义是思维的体现者,是客观事物在人们头脑中的反映,是人们交际过程的中心所在。从信息模型的角度讲,语义是构建在一定语法上,反映一定认知结果的数据对象,数据对象之间关系的描述与客观存在的一种对应关系。因此,信息模型中的语义与对客观存在的概念化以及描述认知结果的语言密切相关。定义这种语义的核心是在数据的相互关系中定义数据的含义。

1.2.2 语义信息模型的用途

针对不同目的,语义信息模型可以采用多种形式以及抽象层次。语义信息模型中包含的信息量要满足以下几种目的:

(1) 问题域基本结构的抽象说明。这种抽象的说明包括领域中最通用的信息及有关知识,以使各方面的利益相关者能够相互理解,并达成对问题域认识的一致。

(2) 全面把握复杂问题域的有关信息。语义信息模型是对问题域有关信息的查找、过滤、检查、重获、组织、重获以及抽象。通过模型研究系统内各组成部分之间的依赖关系,全面把握问题域的信息流。

(3) 对问题域全面或部分的描述。问题域作为一个独立系统不需要进行分解的情况在实际中是很少见的。较通常的情况是,问题域可以分解为相互区别、不连续的描述单元。这些单元作为整体描述的一部分被单独存储和操纵。其中的某些单元由于在不同的抽象层次上具有相似特征,而被认为是同一种类型。语义信息模型需要对反映认识结果的种类进行描述。同时模型要能够描述问题域中典型的事实范例。

从而能够指导信息系统的数据建模,系统的数据处理的设计与实现。

1.2.3 语义信息模型的内容

语义信息模型和任何其他模型一样也包括语法、语义和语境。语法和语境都是为模型所要表达的语义服务的。在具体的工程实践中,通常要根据信息模型所要表达的语义,选择或开发相应的语法规则。由于信息模型要描述数据资源以及数据资源之间的关系,语义信息模型中的语法一般要具有类、关联、状态、实例、规则和消息等语义模型元素。

语义信息模型自身是一个为了在计算机系统中处理、加工、利用信息资源而开发的人工制品,被应用于一个提供模型含义的大型语境中。这个语境包括模型的内部结构、模型建立时的假设、模型应用时的前提条件、模型分析、建立过程的依据和说明、缺省值集合以及模型与其所处环境之间的关系。

1.2.4 语义信息模型的开发

语义信息模型的开发是信息系统开发的一部分,现有的语义数据建模方法都是在从不同的角度努力构造着系统涉及问题域的语义信息模型。语义信息模型的开发方法和工具随着需求变化和时代发展不断地前进。

语义数据模型发展的最初动力是为了克服传统数据模型的缺陷,提供不受具体实现结构限制的方法,即与数据处理的物理实现无关和更多地面向用户的模型。语义数据模型提供了一种"自然"的机制来说明数据库的设计,同时更准确地表示数据及其之间的关系。

语义数据模型主要包括实体关系模型(Entity Relationship Model,简称 E-R 模型)、RM/T 模型、TAXIS 模型、SDM 模型、函数模型、SAM 模型以及 SHM+模型等。其中,以实体关系模型为代表。E-R 模型经过多年的发展、完善,逐渐形成了完整、统一的建模标准,同时有许多软件产品支持用 E-R 模型完成数据库的概念、逻辑和物理建模过程。

前面提到过的 IDEF 方法家族是这一方面应用最为广泛的开发方法。由于问题本身的复杂性,一种表达形式一般只可能适用于一类特征。IDEF 家族发展了各种不同的方法来描述不同的特征。其中的 IDEF5 被称为本体论描述获取(Ontology Description Capture),是一种正在逐渐兴起的语义信息建模方法。这种建模方法还不是很成熟,国内也没有任何机构应用它开发过大型的项目。

从知识共享的角度看,本体论可以被看做是一种概念化的显式说明或表示,是对客观存在的概念和关系的描述。从表面上看,本体论对问题域的描述与数据字典没有太大的区别,也是

许多名词的集合。人们经过进一步研究后，发现两者存在着很大的差距。本体论采用的文法和公理一般采用精确的形式语言、精确的句法和明确定义的语义。本体论的方法并不仅局限于此，这种方法努力使问题域中的概念与概念、概念与对象、对象与对象之间的关系以及在问题域中对象上所施加的约束明确定义，而不是隐含在分析者的头脑中或实现者的程序中，从而大大减小对问题域中概念和逻辑关系可能造成的误解。本书中将对这一方法进行探讨。

1.3 XML 与语义信息模型

许多文献都称 XML 是语义自描述的，是一种语义信息模型。那么 XML 与语义信息模型究竟是什么关系呢？这一节我们分析一下这个问题。

1.3.1 Web 与语义信息模型

Web 应用已经扩展到了人类活动的任何一个方面，远远超出面向文档系统的范围，尽管 Web 一开始被想像成一个文档系统。对于 Web 应用的巨大需求，包括全新的和源于对现有系统改造的需求，加上缺乏熟练的 IT 人才，导致了对更好的软件工程方法的强烈要求。这种情况有点类似于在软件领域采用的较成熟方法的发展，如数据库技术和面向对象方法的发展过程。

电子商务、数字图书馆和远程学习等 Internet 领域的应用具有多项特征混合在一起的特点。这一特点使得它们与以前的信息技术有着根本的不同：

(1) 对于具有有限计算机应用经验或是没有计算机应用经验的用户，通用存取技术需要一个新的人机界面，可以捕获用户的注意力，简化信息存储。

(2) 不同种类的信息资源的全球有效性需要对可能存储于不同系统(数据库、文件系统、多媒体存储设备)和分布存储于多个站点的结构化和非结构化信息进行一体化管理。

近几年，WWW 已被推选为开发 Internet 应用的理想平台，这归功于它强大的基于多媒体性能和浏览方法的信息交流样式，还要归功于它开放的构造标准。这些都促进了不同类型内容和系统的一体化。

现代 Web 应用通常被描述为超媒体和信息系统的混合。由于其混合特性，Web 应用的开发面临着一些实用性的要求：

(1) 需要处理结构化数据(例如数据库记录)和非结构化数据(例如多媒体数据)。

(2) 通过导航界面支持探索性的数据访问。

(3) 高水平的图形质量。

(4) 内容结构的用户化和可能的动态适应性，导航原语和表现风格。

(5) 支持前摄行为，例如，推荐和过滤。

为了提高效率，扩大网站开发所面临任务的覆盖度势在必行。因为市场上绝大多数的网站开发工具只是集中于设计和实现，而不注意需求分析和概念建模，因此，实现并维护一个大的网站是一项人力密集、易出差错的活动，这种活动很难从采用模型概念和 CASE 工具的规范开发过程中获益。

为了解决这些问题，研究人员提出了所谓的模型驱动网站设计方法。这类方法采用了以半结构化符号表达数据结构和网站的超文本拓扑的思想，同时使用概念级规范来驱动设计和实现的思想。

在这类 Web 开发方法中，完成需求分析之后，有一个被称为概念化(Conceptualization)的设计阶段。在这一阶段中应用被一组覆盖预期解决方案主要组成部分的抽象模型所表示。在 Web 环境下，概念化与信息系统设计中的类似行为有着显著的风格差异。在 Web 环境下的概念化将主要精力集中于获取展现给用户的对象和对象之间的关系，而不是这些对象和对象之间的关系如何在软件系统中表示。尽管它们使用的符号可能是相同的，但是由此概念化产生的 Web 应用视图却有别于一般的数据库应用。

1.3.2 XML 与语义信息模型

XML 对统一结构化语法和半结构化语法的承诺，又重新燃起了人们将一些几乎不可能的事变成切实可行的希望。那么，什么是 XML 的语义？因为语义这个词的特殊性，每个人对语义定义的观点都各有不同。一般来说，语义是构建在公用语法上的系统中 XML 数据的一层规范。按照 Uche Ogbuji 的想法，标记 XML 语义的概念如下，当然，在这三概念之间有一些重叠：

(1) 元素类型名称、属性名称和某些情况下内容术语的解释。

(2) 用于用有效文档引导事务的处理规则(也称作商业规则)。

(3) 一个文档中的结构化元素与另一个文档中的结构化元素之间的关系。

当我们提到 XML 时，通常并不单指 XML 协议本身，更普遍的情况是指以 XML 为基础的相关协议簇以及由此而带来解决问题的新方法。就 XML 本身而言，XML 目前有明确的语法和结构，但它没有提语义透明性。语义透明性可以使 XML 机器建立元素(比如，Purchase-Order 或 PO)和根据该元素执行专门操作的高阶处理之间的关系。总而言之，它意味着数据中的表达式如实地表示了相应概念的含义。语义透明性的最终测试是如果某个人只使用适用于 XML 处理软件的机制，它能否正确理解 XML 数据的含义。显然，单靠 XML 根本无法实现语义透明性，这正是众多 XML 技术专家关注语义透明性的原因。

那么为什么人们会对 XML 的出现而欢欣鼓舞并寄予厚望？原因是多方面的，其中最重要的就是：XML 是一种完全面向数据语义的标志语言，取消了 HTML 的显示样式与布局描述能力，突出了数据的语义与元素结构描述能力。具体地讲包括以下几个方面：

(1) XML 使用 ASCII 编码。

(2) 以内容为中心，符合 Web 的主旨。

(3) 构建于 Web 标准和概念之上。

(4) 沿袭 SGML 的传统。

(5) 与 EDI 兼容。

(6) 被广泛采纳和使用。

(7) 应用领域广泛。

我们将在本书中以 XML 为具体手段，以 Web 应用为背景介绍语义信息模型的多种应用，探讨 XML 在语义信息模型中的应用方法。下面首先介绍 XML 与数据的语义表示，作为进一步研究的基础。

1.4 XML 与数据的语义表示

早期的计算机使用的数据和处理它的程序都固化在穿孔卡片上，数据的语义和数据的处

理合二为一。后来,数据与它的部分语义被抽取出来,放到特殊格式的数据文件里,如 dbf 文件。文件头部包含数据的部分语义,另一部分语义依然保留在处理它的程序里。在这个发展过程中,数据的语义与处理它的程序的耦合程度由紧密走向分离。分离意味着灵活性和共享性。这一发展方向预示着数据的语义最终会独立出来,数据不再只是数据,而是带有语义的信息;处理数据的程序不再只是专有的、惟一的,而是共享的、开放的。在开放的环境中,软件的互操作才真正成为可能,这是一条"标准化"的道路。

W3C(World Wide Web Consortium)开发的 XML(eXtensible Markup Language,可扩展标记语言)正是实现这一目标的产物。W3C 创立于 1994 年 10 月,致力于领导万维网(World Wide Web),制定公共的协议,促进万维网的发展并确保其互操作性。W3C 在世界各地已有 400 多个成员组织,并在推进万维网的发展方面赢得了广泛的国际赞誉。

1.4.1 数据表示的第一次统一

什么是数制? 用一组固定的数字和一套统一的规则来表示数目的方法就叫做数制(Number System)。人们比较熟悉的是十进制,即 0,1,2,3,4,5,6,7,8,9 这十个数。在计算机领域中,还有另外 3 种数制:二进制、八进制和十六进制。八进制的基数是 8,十六进制的基数是 16。十六进制用 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F 来表示,这些数制原理与十进制原理相似。二进制是计算机与网络通信中采用的基本数制,而八进制和十六进制则是二进制的压缩形式。这些数制之间是可以相互转换的。

什么是数据? 数据指人们看到的形象和听到的事实。自从有了计算机,就存在两种形态的数据。一种形态称为人类可读形式的数据,简称人读数据,是人类使用的语言、文字、数字以及图像。另一种形态称为机器可读形式的数据,简称机读数据,是计算机能够识别的二进制数的形式。在二进制中,数据的最小单位就是二进制的一位数(0 或 1),简称为位,英文名称是比特(bit)。正是通过这些 0 和 1 的组合,可以将人读数据中的所有基本符号——字母、数字以及专门符号表示出来。在计算机中,8 位为一个字节(Byte),这是计算机进行数据处理和存储的基本单位。此外,我们还会遇到另一种单位:机器字长,表示计算机一次能够处理数据的位数,是衡量机器能力强弱的标志之一。最初的微型机只有 8 位,后来发展到 16 位机、32 位机,直到字长为 64 位的巨型机。

什么是编码? 编码就是用二进制表示字母、数字以及专门符号的一串数字。这是一个涉及世界范围内有关信息表示、交换、处理和存储的基本问题,一般都以国家标准或国际标准的形式颁布施行。目前国际上通用的编码系统为美国标准信息交换码(ASCII)。为了适应汉字信息交换的需要,我国于 1981 年颁布了汉字编码系统,这个系统包含 6763 个常用汉字,在国际上也适用。有了这两套编码系统,人们才真正可以用汉语与计算机"对话"。

通过编码把人读数据转化为机读数据的过程是由计算机硬件系统和软件系统完成的。硬件系统只能识别 0 和 1,软件系统分为计算机本身运行所需的系统软件和用户完成特定任务所需的应用软件。

任何文字、图形等数据,只要用计算机处理就变成了 0 和 1,即所谓数字化,这一进步把人类带入了新的时代。任何信息只要通过计算机处理就进入了新的信息管理状态,其共享、传递和储存都有了新的方式。正是在这个意义上,人们开始了信息编码化的历程。

在 Licklider 提出"电脑与人类交流"的思想之后,1963 年,一位在电脑发展史上做出重大贡献的人物终于制定出统一的信息表示方法,即 ASCII(美国信息交换标准码)编码。这为