Guoji Jingjixue

高等学校“十二五”应用型经管规划教材

国际经济学

梁 军　马 宁　赵效娟　编著

Guoji Jingjixue

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

本书由中国工程科技咨询知识服务系统项目资助出版

数字时代的元数据实践

主　编　孙晓菲
编　著　韩子静　曹玉霞　熊健敏

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

数字时代的元数据实践 / 孙晓菲主编. —杭州：浙江大学出版社，2013.3
ISBN 978-7-308-11226-0

Ⅰ. ①数… Ⅱ. ①孙… Ⅲ. ①元数据—研究 Ⅳ. ①G250

中国版本图书馆 CIP 数据核字（2013）第 037880 号

数字时代的元数据实践

主　编　孙晓菲
编　著　韩子静　曹玉霞　熊健敏

责任编辑　杜希武
封面设计　刘依群
出版发行　浙江大学出版社
（杭州市天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　杭州好友排版工作室
印　　刷　浙江省邮电印刷股份有限公司
开　　本　787mm×1092mm　1/16
印　　张　20.25
字　　数　492
版 印 次　2013 年 3 月第 1 版　2013 年 3 月第 1 次印刷
书　　号　ISBN 978-7-308-11226-0
定　　价　59.00 元

前　言

20世纪末以来，信息技术的发展、数字图书馆的兴起及网络环境的提升，逐步改变了人们利用信息资源以及阅读信息资源的习惯。网络传播的普及，使数字资源不但成为科研、教学的重要信息来源，甚至变成普通人信息生活中不可或缺的重要组成部分。迈入21世纪后，数字资源增速迅猛，如何组织、管理和利用数字信息成为社会关注焦点和学术研究热点。面对海量信息，信息资源收藏机构逐渐达成共识，只有在共建共知共享基础上，数字资源收集、管理及利用成本才会降低，效用才能增大。要达到这一目标，数字资源的制作和描述必须标准化和规范化，以方便不同机构之间共享成果和交流数据。而元数据在数字资源的制作和描述过程中起着关键作用，无论是数字资源的出版，还是书目信息加工及网络服务，都离不开元数据。元数据不仅是信息收藏机构的重要管理标准，也是数字资源出版的重要生产标准。

本书主要内容就是围绕数字信息资源的组织与管理，介绍元数据的相关知识以及编著者近年来的元数据实践。本书编著者熟悉国内外主要元数据标准，有多年资源组织及管理实践经验，曾参与多个项目的元数据设计，探索并完成了不同类型项目的元数据标准及规范制订。

本书第1章至第3章是元数据概述、常用元数据及元数据应用与构建，主要介绍元数据的基本知识以及元数据设计原则和方法等。第4章至第7章则以项目为背景，详细介绍了大学数字图书馆国际合作计划(CADAL)、非物质文化遗产、敦煌壁画数字资源以及科技咨询报告等项目的元数据设计原则、设计过程及元数据规范与实例等。第1章至第3章由孙晓菲撰写，第4章至第7章由孙晓菲、韩子静、曹玉霞、熊健敏等共同撰写。全书由孙晓菲统稿，熊健敏协助统稿。

目 录

第1章　元数据概述

近二十年来，随着数字技术在信息组织管理中的应用与发展，从用户信息检索利用的全过程来重新审视信息组织的要求和功能，从信息交流的整体高度来建立新的组织机制和工具已成为信息资源组织与管理的共识。从数字资源的制作加工到出版信息发布，从资源采购、组织与管理到资源服务，整个生产链、服务链都有海量信息产生，要实现海量信息的有序化传播与标准化规范化管理，必须建立符合现代技术环境的数据管理机制。这种管理机制须依赖统一的数据标准，对数据的数据进行规范。

元数据作为数据的数据，在数字资源的制作和组织管理中起着关键作用，数字资源的出版、信息加工及网络服务，都离不开元数据。元数据不仅是信息收藏机构重要的组织管理与利用标准，也是数字资源出版的重要生产标准。可以说，元数据的应用已超越了传统意义的信息对象揭示与描述，数据管理者已将元数据的概念和方法运用于信息系统和信息过程中的各个层面。

元数据除了应用于以 Marc 和 DC 为代表的信息资源著录与描述、发现、确认、检索与服务；还广泛应用于资源保护与保存、权限管理及资源评估等。本章将详细介绍元数据定义、体系结构、标准、类型以及资源描述框架等内容。

1.1　元数据定义

元数据(metadata)的定义是“关于数据的数据”(data about data)。这个定义被认为过于宽泛和简练，对元数据的内涵与外延缺乏清晰地描述。因此，不同的机构和组织根据自己的理解对元数据给出了不同的定义。以下列举几种：

中国中文元数据标准研究项目组在 2000 年 12 月的《国外元数据标准比较报告》中对元数据的定义是：“元数据是描述某种类型资源(或对象，object)的属性、并对这种资源进行定位和管理、同时有助于数据检索的数据。”[1]

美国图书馆协会对元数据的定义是：“元数据是数据的数据。它可以描述一个数据集、数据对象或资源，包括它的格式、收藏时间及收藏者。元数据一般指网络资源，但它也能描述实体资源及电子资源。”[2]

国际图书馆协会联盟(IFLA)对元数据的定义是：“元数据是数据的数据。用于帮助对任一网络电子资源的辨识、描述及定位。”[3]

国内出版的有关元数据研究的著作中，较有代表性的定义有：张晓林在《元数据应用与研究》一书对元数据比较简捷的定义为：“元数据是关于数据的数据，是对数据进行组织和处理的基础”[4]。肖珑在《中文元数据概论与实例》一书对元数据的定义进行了扩展，他认为：

“元数据是关于数据的数据，是专门用来描述数据（数字对象）的内容、特征和属性，并对数据进行管理和结构化的数据，是数字图书馆信息组织的基础”[5]。

以上几种定义阐述了元数据的本质和作用，即元数据是提供关于信息资源或数据的一种结构化的数据，是对信息资源的结构化描述；它的作用是描述信息资源或数据本身的特征和属性，规定数字化信息的组织结构，具有定位、发现、证明、评估和选择等功能。因此，元数据是实现数字资源共建、共知、共享的核心内容之一，不仅用来规范描述数字资源，而且用于组织、管理和挖掘数字资源，是数字信息体系的重要基础。

1.2 元数据体系结构

数字图书馆中，元数据结构体系由内、外两部分组成。内部系统主要指数字图书馆系统本身的元数据处理方法和体系结构，即元数据管理系统，它是整个数字化图书馆系统的重要组成部分，其基本功能是为数字化图书馆系统的运行建立基础；外部系统指数字图书馆外部的元数据环境，即各种独立于具体系统的、通用的元数据标准的总和。

为了实现数字图书馆和外界信息环境的沟通，元数据内部系统和外部系统必须是同构的，这种同构关系实际是将外部元数据系统映射到数字图书馆内部体系中的方法。为了建立同构关系，元数据管理系统的结构包括五个组成部分。

（1）通用元数据系统，是指某个数字图书馆标准的元数据系统。它的作用是作为基准元数据，组织标识数字化图书馆中的数字化信息资源；以标准形式描述用户的查询提问；为各种网络信息发掘工具提供数字化信息。

（2）元数据字典，是一种用于各种元数据体系到系统基准元数据系统相互转换的对照表，它描述了各种元数据的基本特征，构建了各种元数据与基准元数据系统的对应关系。它的基本作用是为系统的转换模块提供转换依据。

（3）数据属性集，是指数字图书馆存储数据的属性总和。元数据管理系统可通过数据属性集将数字化图书馆的数据结构和基准元数据相对照，保障它们之间的可互换性。

（4）数字化信息资源集，它的描述对象是信息源。数字图书馆系统可以通过信息源特征集来确定各个信息源所采用的元数据体系，将用基准元数据表达的查询式转换成各个信息源所采用的元数据表达式，从而决定各个信息源的检索方法并解释检索结构。

（5）元数据转换和维护模块。该模块提供系统内、外各种元数据之间相互转换和翻译方法，并可以对系统内各种对照表进行添加、删除、修改等动态管理，保证元数据管理系统的可扩展性和可维护性。

数字图书馆通过元数据实现对信息资源的描述与检索，选择与定位，整合与维护，管理与评估等功能，通过内、外系统满足用户对信息资源利用的各种需求。

1.3 元模型标准

元数据对机构或项目来说，除了用于机构或项目数据存储及使用外，最大的贡献或最高

境界就是实现数据的共建、共享与共知。因此，元数据行业需要标准的“元模型”，用于存储数据的对象或关系型物理数据模型，标准的元模型是元数据共享及互操作的基础。

当前，关于元模型标准有两大体系：元数据联盟(Metadata Coalition，MDC)的开放信息模型(Open Information Model，OIM)和对象管理组织(Object Management Group，OMG)的公共仓库元数据(Common warehousing Metadata，CWM)。下面将介绍元模型两大标准体系的开发及发展情况。

1.3.1 开放信息模型 OIM

元数据联盟(MDC)成立于1995年，成员有150多家，包括了微软和IBM等著名软件厂商。MDC是一个致力于建立与厂商无关、不依赖于具体技术的企业元数据管理标准的非营利技术联盟。1997年，MDC曾发布自己的元数据标准—元数据交换规范(Metadata Interchange specification，MDIS)，但未获成功。1998年，微软将开放信息模型(OIM)授权MDC开发；1999年，MDC接受了微软的建议，将OIM作为元数据标准。MDC将OIM的发展目标定为独立于具体技术和生产商的元数据标准，通过共享的元数据模型支持开发商工具间的互操作。

OIM的目的是通过公共的元数据信息来支持不同工具和系统之间数据的共享和重用。它涉及了信息系统(从设计到发布)的各个阶段，通过对元数据类型的标准描述来达到工具和知识库之间的数据共享。OIM所声明的元数据类型都采用统一建模语言UML(Universal Modeling Language)进行描述，并被组织成易于使用、易于扩展的多个主题范围(Subject Areas)，这些主题范围包括：

- 分析与设计(Analysis and Design)主要用于软件分析、设计和建模。该主题范围又进一步划分为：UML包(Package)、UML扩展包、通用元素(Generic Elements)包、公共数据类型(Common Data Types)包和实体关系建模(Entity Relationship Modeling)包等。
- 对象与组件(Object and Component)涉及面向对象开发技术的方方面面。该主题范围只包含组件描述建模(Component Description Modeling)包。
- 数据库与数据仓库(Database and Warehousing)包括数据库模式管理、复用和建立数据仓库提供元数据概念支持等。该主题范围进一步划分为：关系数据库模式(Relational Database Schema)包、OLAP模式(OLAP Schema)包、数据转换(Data Transformations)包、面向记录的数据库模式(Record-Oriented Database Schema)包、XML模式(XML Schema)包和报表定义(Report Definitions)包等。
- 业务工程(Business Engineering)是为企业运作提供一个蓝图。该主题范围进一步划分为：业务目标(Business Goal)包、组织元素(Organizational Elements)包、业务规则(Business Rules)包、商业流程(Business Processes)包等。
- 知识管理(Knowledge Management)涉及企业的信息结构。

上述主题范围中的包都是采用UML定义的，可以说UML语言是整个OIM标准的基础。虽然OIM标准并不是专门针对数据仓库的，但数据仓库是它的主要应用领域之一。

1.3.2 公共仓库元数据 CWM

对象管理组织(OMG)成立于1989年，是由Oracle、IBM、Unisys、NCR以及Hyperion发起，拥有500多会员的国际标准化组织，著名的CORBA标准即出自该组织。它的目标是

建立一种行业标准和对象管理规范来为实际软件开发提供一个通用的构架。2000 年，OMG 年采用 CWM，目的在于推动数据仓库、智能商务和知识管理方面元数据的共享和交换。2001 年 3 月，OMG 颁布了 CWM 1.0 标准[6]。

CWM 涵盖设计、建立和管理数据仓库应用的整个生命周期，并支持生命周期管理；同时，CWM 支持模型驱动架构（Model Driven Architecture，MDA）进行元数据交换，是迄今为止将 MDA 方法用于具体应用领域（数据仓库和商业智能领域）的最完美的例子。

CWM 完整地描述了数据仓库元数据交换的语法和语义以及用于异质平台之间的元数据交换机制，并在数据仓库系统中定义了一套完整的元模型体系结构，用于数据仓库构建和应用的元数据建模。CWM 元模型由一系列子元模型构成，包括：

- 资源数据元模型：用于为对象型的、关系型的、记录型的、多维的和 XML 等数据源建模；
- 数据分析元模型：用于为数据转换、联机处理分析（OLAP）、数据挖掘、结果信息可视化等分析处理结果建模；
- 仓库管理元模型：用于为数据仓库处理流程和操作功能进行建模。

OMG 组织不但发布了 CWM 规范，描述了一套完整的数据仓库元模型对象及对象定义，还为数据仓库和商业智能（Business Intelligence，BI）工具之间共享元数据，制定了一整套关于语法和语义的规范。这四个规范分别为（参见图 1-1）：

- CWMUML 类图：描述数据仓库系统的模型；
- CWMXML 文件：用 XML 形式描述 CWM；
- CWMDTD：数据仓库和商业智能工具共享元数据的交换格式；
- CWMIDL：数据仓库和商业智能工具共享元数据的应用程序访问接口（API）。

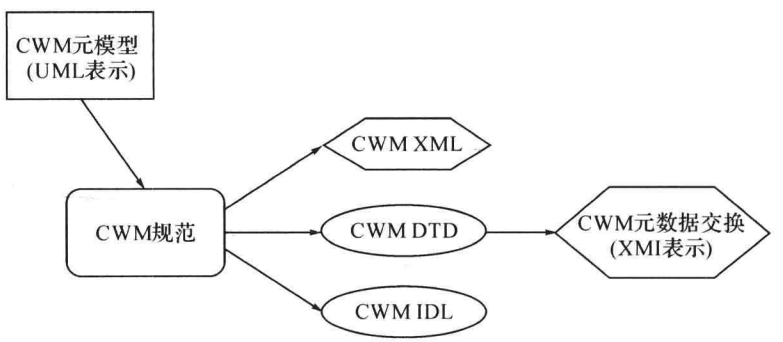

图 1-1　CWM 规范

1.3.3　OIM 与 CWM 的统一

前面介绍的 CWM 实际上是专门为数据仓库元数据而制定的一套元模型标准，而 OIM 并不是针对数据仓库元数据的。OIM 所关注的元数据的范围比 CWM 要广，CWM 只限定于数据仓库领域。

OIM 与 CWM 标准在一段时间内同时存在。显然，一个行业两个标准对元数据发展来说弊大于利。尽管，OIM 与 CWM 的分歧并不严重，软件商只需创建两个元模型的接口，就可以解决标准不一的问题。但是，OIM 与 CWM 合并将使整个 IT 行业受益，而机构可以高

效地管理数据仓库。因此，创建一个独立于平台和软件的元数据标准是一个明智选择。

为了推动标准并使业界广泛认同，MDC 成员决定合并进 OMG，并由 OMG 发布同一套标准 CWM，MDC 不再研究自己独立的标准。同时，CWM 从 OIM 中借鉴和学习了很多设计，致力于解决数据仓库和商业智能的元数据问题。随着 MDC 和 OMG 组织的合并，为数据仓库厂商提供了统一的标准，从而为元数据管理铺平了道路。

元数据标准的统一过程也给元模型标准是否成功提供了判定因素，而技术独立、生产商的立场、现实性与可广泛使用等是构成判定元数据标准是否成功的主要因素。这些因素具体如下：

(1) 技术必须独立

完善的元模型标准技术上必须独立，不受制于任何特定技术或特定平台，只需对元数据标准做一些针对具体环境的小调整，标准就可以执行。

(2) 生产商在技术上保持中立

元模型标准必须由所有主要的软件生产商协作开发，在设计标准时，不偏向任何特定的生产商，使元模型可以最大限度地应用于各软件平台。

(3) 标准具有现实性

任何标准并不能包罗万象，好的标准应该是获取机构所需通用元数据的 95%，剩余的 5%可以留在以后完成。为了制作大而全的标准付出的时间和精力，并不能让标准在现实中能得百之百分的实施。

(4) 标准需要广泛实施

标准的意义在于实施，停留在纸面上的标准没有任何价值。因此，标准得到业界认可，拥有广泛的使用人群，才是标准的意义所在。

1.4 元数据表达法则与编码体系

1.4.1 元数据表达法则

元数据的表达法则在计算机应用系统中包括描述语言和语法结构两个部分。元数据的描述语言是 XML，描述框架是 RDF，语法结构是 XML Schema 或 DTD(文献类型定义)。元数据格式则通过三层结构来完整定义：

第一层为内容结构，对其构成元素及其定义标准进行描述。元数据内容结构的构成元素有：描述性元素(Descriptive elements)、技术性元素(Technical elements)、管理性元素(Administrative elements)、复用元素(Reused elements)等[7]。

内容结构定义中，还需要制定元素的选择与使用规则，如元素的必备性、重复性、子元素组成等，一般这些规则参照成熟的信息处理框架或标准来执行，譬如 MARC 依据 ISBD 来执行著录规则，EAD 依据 ISAD(G)来处理信息等。

第二层为句法结构，定义元数据整体结构及描述方法，包括元素的分区分层分段结构、结构描述方法、DTD 描述语言、复用方式等；同时也可定义元数据与被描述数据对象的捆绑方式。

第三层为语义结构，定义元数据元素的具体描述方法，包括三个层次：元素定义、元素内

容编码规则定义、元素语义概念关系。

其中，元素定义遵循 ISO 11179 标准(Specification and standardization of data elements)，规定了元素的 10 个属性：

(1) 名称(name)

(2) 标识(identifier)

(3) 版本(version)

(4) 注册机构(registration authority)

(5) 语言(language)

(6) 定义(definition)

(7) 限定(obligation)

(8) 数据类型(datatype)

(9) 最大使用频率(maximum occurrence)

(10) 注释(comment)

元素内容编码规则可以是特定标准或是自定义的描述要求，也允许一个元素使用多个编码规则。在具体实施时，执行单位可根据客观需要选择其中的一种作为实行标准，甚至可以选择编码外的规则作为实行标准。如在中国使用都柏林核心元数据(Dublin Core)时，主题词描述规范为《汉语主题词表》或《中国分类主题词表》。

元素语义概念关系指把元素放在一个概念体系中来说明它与其他概念的关系，说明它的上、下文关系，如：part_of、used_by、interact_with 等。

1.4.2 元数据编码体系

元数据是资源在计算机科学的一种映射，是物理世界与数字世界的对话形式之一。它通过对资源各种属性的描述来揭示资源本体，从而为数字到资源的映射提供途径。这里的资源可以是实体对象或概念、关系，如一本书、一台仪器、一个网页等。

元数据编码体系是用元数据的形式实现资源目录组织，便于资源管理和检索利用的一系列解决方案。传统的图书馆卡片、花名册、账本、药典等，都是元数据。随着计算机技术在资源管理领域的应用，元数据不仅面临着人类的语义表达问题，还面临着机器的规范化处理问题，这些问题的解决本质上是计算机网络环境下的元数据编码问题。在领域内把元数据编码也叫做元数据方案形式化处理，主要解决元数据的语义表述、结构描述，以及句法规则[8]。

元数据编码通过元数据编码语言得以实现。所谓元数据编码语言(Metadata Encoding Languages)指对元数据元素和结构进行定义和描述的具体语法和语义规则，常称为定义描述语言(DDL)。在元数据发展初期，人们常使用自定义的记录语言(例如 MARC)或数据库记录结构(如 ROADS 等)。但随着元数据格式的增多和互操作的要求，人们开始采用一些标准化的 DDL 来描述元数据，例如通用标识语言标准(SGML)和可扩展标识语言(XML)。

通用标识语言标准 SGML(Standard Generalized Markup Language)、超文本标识语言 HTML(HyperText Markup Language)和可扩展标识语言 XML(Extensible Markup Language)都是标记语言，其中 SGML 是 XML 和 HTML 的母语言，它的覆盖面很广，几乎涉及了人们生活的每一个领域。以下将简要介绍 SGML、HTML 和 XML 三种标记语言。

(1) 通用标识语言标准(Standard Generalized Markup Language，下简称 SGML)

SGML是一种通用的文档结构描述标记语言，主要用来定义文献模型的逻辑和物理类结构，是ISO组织于1986年发布的ISO 8879国际标准。它的前身是通用标记语言(GML)，1969年由IBM公司研究人员Goldfard、Mosher和Lorris创建，并在20世纪70年代成为出版行业中的一个重要标准。1978年Charles Gold fard出任美国国家标准协会(ANSI)文本处理计算机语言委员会的主持人。在他的主持下，美国国家标准协会文本处理计算机语言委员会在1980年公布了SGML的第一个工作草案。在1984年，这个委员会发展成为一组协作共事的子委员会，为国际标准化组织和美国国家标准协会开发标准。1986年，SGML成为国际标准化组织的标准(ISO 8879:1986)。

SGML的文档由三部分构成：结构(structure)、内容(content)和样式(style)，主要处理结构和内容之间的关系。SGML语言程序由三部分组成：SGML声明、文件类型定义(简称DTD-Definition Type Document)和文件实例。SGML声明，定义了文件类型定义和文件实例的语法结构；文件类型定义，定义了文件实例的结构和组成结构的元素类型；文件实例是SGML语言程序的主体部分。SGML在实际使用中，每一个特定的DTD都定义了一类文件。不同类型文献将有不同DTD，DTD是SGML文件的核心部分。例如，所有的电子商务文件都可以使用同一个DTD。所以，习惯上把具有某一特定DTD的SGML语言，称为某某标记语言，例如用于国际互联网的HTML语言，SGML就成为那些派生语言的元语言。

SGML的标准可分三个层次：第一层次是元语言标准：SGML标准；第二层次是基础标准，如：HyTime、DSSSL等，是该体系的基本标准；第三层次是具体的应用标准，如：Internet上已广泛应用的HTML、VRML等标准。

而第二层次的基础标准又可分为三类：信息描述标准、信息表现标准和信息关联标准。信息描述标准是与SGML标准本身直接相关的标准，如：SGML公共标识符注册标准(ISO 9070)、SGML技术报告(ISO TR 9573)、SGML一致性测试系统标准(ISO/IEC 13673)、文档处理APIs标准等等。

信息表现标准是关于组合文档、超媒体文档的描述与处理的标准。如：文档处理框架与逻辑文档格式化、多语种字体信息、信息交换与服务等等。具体标准包括：文档样式语义与规范语言标准DSSSL(ISO/IEC 10179)、标准页面描述语言SPDL(ISO/IEC 10180)、字体标准Fonts(ISO/IEC 9541)、字体注册标准Font Registration(ISO/IEC 10036)等等。

信息关联标准是关于基于SGML进行信息管理与交换的标准，如：信息的链接与定位、基于时间的信息管理、知识结构与索引的表示法、交互式文档中的动作管理。具体的标准包括：超媒体/基于时间的结构化语言标准HyTime(ISO/IEC 10744)、基于主题的地图导航标准(ISO/IEC 13250)、可更改交互文档交换标准ISMID。

在三层协议中，信息表现部分最重要的标准是DSSSL(Document Style Semantics and Specification Language)，信息关联部分最重要的标准是HyTime(Information processing—Hypermedia/Time-based Structuring Language(HyTime)-2d edition)。这两个标准也是对XML标准体系中影响最为深刻的两个。DSSSL是一种与平台无关的进行SGML文件处理的语言，其中主要包括转换语言、样式语言、表达式语言和标准文档查询语言(SDQL)，XML相关标准中的CSS、XSL、XSLT、XPath，就有很多内容是从DSSSL中继承过来的。HyTime则定义了一个元素类型的集合，以便使用者可以利用这些类型，以一种标准的方式，在已有的SGML文档中提供超链及其他功能。

SGML 相当完备同时也相当复杂，它的标准超过了 500 页，十分庞大且难于学习和使用。因此，SGML 并不直接应用，而通过派生产生出不同的标识语言，如人们熟悉的 HTML、XML 语言。

(2) 超文本标识语言(HyperText Markup Language，下简称 HTML)

HTML 是 SGML 的一个实例，作为第一代网络语言，其精简的语法以及通用性、易用性很快为大众掌握，互联网的蓬勃发展 HTML 功不可没。作为标记语言，它本身不能显示在浏览器中，需经过浏览器的解释和编译，才能正确反映 HTML 标记语言的内容。

HTML 从 1.0 到 5.0 经历了巨大的变化。HTML 的第一个官方版本是由 IETF(互联网工程任务组)推出的 HTML 2.0。后来，W3C 取代 IETF 的角色，成为 HTML 标准制订的组织。上个世纪 90 年代的后半叶，HTML 的版本被频繁修改，直到 1999 年的 HTML 4.01才被广泛使用。

HTML 在 HTML 4.01 之后的第一个修订版本是 XHTML 1.0，XHTML 1.0 基于 HTML 4.01，它没有引入任何新标签或属性，唯一的区别是增加了严格的语法限制。后来，W3C 又推出了 XHTML 1.1。然而，来自 Opera、Apple 以及 Mozilla 的代表不满意 W3C 的工作，他们自发组织成立了 WHATWG 超文本应用技术工作组，致力于 HTML5 规范。后来 W3C 在 XHTML 方面的工作慢慢地陷入困境，最终选择了将 WHATWG 的成果作为 HTML5 基础，并于 2008 年发布了 HTML5 第一份正式草案，目前 W3C 正在对它进行进一步完善。

HTML5 的提出是为了解决网络应用的需求，大多数需要插件和扩展来完成的功能，原生的 HTML5 语言已经能过全部提供，另外还包括图像功能的增强、Web 数据存储和离线数据存储，加强了让浏览器处理本地数据的能力，使得浏览器可以〈部分代替操作系统。

a) HTML 文件的基本标签

HTML 定义了三种基本标签用于描述页面的整体结构，以及浏览器和 HTML 工具对 HTML 页面的确认。这三种基本标签为：

〈HTML〉标签：是 HTML 文档的第一个标签，它通知客户端该文档是 HTML 文档，结束标签〈/HTML〉则出现在 HTML 文档的尾部。

〈HEAD〉标签：出现在文档的起始部分，标明文档的题目(或介绍)，该部分包含的是文档的无序信息。文档标题部分可以包含题目和主题信息。结束标签〈/HEAD〉指明文档标题部分的结束之处。

〈BODY〉标签：HTML 文档中的〈BODY〉标签用来指明文档的主体区域，该部分通常能够包容其他字符串(如标题、段落、列表等)，读者可以把 HTML 文档的主体区域简单地理解成标题以外的所有部分。结束标签〈/BODY〉指明主体区域的结尾。

b) HTML 文件的基本组成

一个标准的 HTML 文件由 3 个基本部分组成：HTML 元素、元素的属性和相关属性值。

HTML 的基本元素包括标题、段落、换行和注释语句，前 3 种是组成一篇文章最基本的要素。

● 标题由〈h1〉到〈h6〉元素来定义，〈h1〉代表最高级别的标题，〈h6〉级别最低，其中字母 h 是英语 headline 的简称。它们的语法格式为：

〈h1〉一级标题〈/h1〉

〈h2〉二级标题〈/h2〉

……

标题的首尾标记必须成对出现，结尾标记不能忽略。

● 段落元素〈p〉的英文全称是 paragraph，用来起始一个段落，它是一个块级元素，不能再包含其他的任何块级元素。它的语法格式为：

〈p〉段落正文内容〈/p〉

它的起始标记必须有，结尾标记是可选的。也就是〈p〉段落正文内容〈/p〉和〈p〉段落正文内容所显示的内容是相同的。

● 换行元素〈br〉的英文全称是 line break，在不另起一段的情况下，将当前文本强制换行。它的语法格式为：

〈p〉段落正文内容〈br〉另起一行

它的起始标记必须有，而结尾标记是禁止出现的，即〈br〉元素只能单独出现，不能成对出现。

● 注释语句元素〈！————〉，注释语句的作用是对标记语言内容进行必要的说明，方便设计者日后修改代码和维护工作；对其他设计者来说，通过注释语句可以很快理解原设计者的内容。它的语法格式为：

〈！——注释语句内容——〉

c) HTML 的局限性

HTML 是一种预定义的标识语言，只是在一类特定的文件中定义了一种描述信息的方法。它的 DTD 作为标准被固定下来，因此，HTML 不能作为定义其他置标语言的元语言。HTML 自身的特点使它蕴藏了许多危机，随着它的不断发展，这些危机不但没有减弱，反而越来越突出，甚至成为 HTML 继续发展的障碍。HTML 存在的问题有以下几个方面：

第一，HTML 在网页制作中的局限性。HTML 无法描述数据内容，而这一点恰恰是数据检索、电子商务所必需的。近两年来，电子商务、远程教育等全新的领域在这个网络时代迅猛发展，并成为了互联网中重要的组成元素，当然随之而来的便是网站页面的复杂化、多样化、智能化，可这些恰恰是 HTML 所欠缺的因素，它的简单灵活让使用者受益匪浅，可是对如今网页中复杂多变的具体应用却显得无能为力。

第二，HTML 应用范围的局限，对数据表现的描述能力是十分不够的，如 HTML 还不能描述矢量图形、科学符号等对象，目前只能通过图像来表现这些对象。而当前的浏览器功能越来越丰富，甚至已经超出了 HTML 的设计范围，单向式页面描述格式的，HTML 语言就无法再适宜表达越来越丰富的页面需要了。况且现在出现了很多 HTML 无法识别和处理的专业格式，比如音乐乐谱、化学方程式、数学公式、财务报表以及工程应用等等，而在 XML 上它们都可以得到非常好的支持。

第三，HTML 与浏览器之间的矛盾，使 HTML 实例标记语言已不能适应对新标记需求的发展需要。如果想要使 Web 页面更加丰富多彩，或者是想在页面中添加特效，标准的 HTML 标记就显得力不从心了，需要通过动态 HTML 以及一些 JAVA 程序来实现。不过在加上这些非标准的新技术之后，就不能保证所制作出来的页面在不同的浏览器之下都会有同样的效果，甚至有一些用非标准的。HTML 语言所制作成的页面，用一种浏览器浏览

不会出现任何问题，而换用另外一种浏览器往往就不能得到正确的页面效果。这使得网页制作人员不得不在两者之间不停地测试，或者干脆做出两套适应不同浏览器使用者的页面来。

第四，HTML 链接方式的局限性。虽然现在 HTML 提供的超文本链接取得了巨大的成功，但它还是暴露出了相当的局限性：超链接在它的源端定义；超链接确定了它的目的端；用户只能从源端出发走到目的端；超链接的效果由浏览器而不是由超链接本身来决定。而 XML 的超链接方式则要比 HTML 多出了一些新的特性，无需手写很多的 JAVAScript 代码就可以创建出智能型的超级链接，利用 XML 中的 Xpointer 我们可以直接“取址到”其他文本的任何部分，而不再是“链接到”。

以上所列出的局限性中，有一部分随着 HTML 及相关技术的不断发展正在不断地改善与提高。值得关注的是，HTML 继承了 SGML 的许多重要的特点，比如结构化、实现独立和可描述性，具有通用性、易用性等优势，曾经 HTML 在表述文件结构化从而方便数据的交换等方面的应用中显得捉襟见肘，相信随着 HTML5 的应用普及，HTML＋浏览器的跨平台优势，必将为互联网提供更生动丰富的应用[9]。

(3) 可扩展标识语言(Extensible Markup Language，下简称 XML)

XML 是 W3C 组织于 1998 年 2 月发布的标准，W3C 组织制定 XML 标准的初衷是，定义一种互联网上交换数据的标准。W3C 采取了简化 SGML 的策略，在 SGML 基础上，去掉语法定义部分，适当简化 DTD 部分，并增加了部分互联网的特殊成分。因此，XML 也是一种标记语言，基本上是 SGML 的一个子集。因为 XML 也有 DTD，所以 XML 也可以作为派生其他标记语言的元语言。

在互联网上，服务器与服务器之间、服务器与浏览器之间有大量的数据需要交换，特别在电子商务中，用于交换的数据都被要求对数据的内容和表现方式有所说明。SGML 对互联网应用来讲太复杂了，因此，需要一种既能象 SGML 那样作为元语言使用，又能比较简单地进行处理的置标语言来担此重任。在这种背景下，XML 就应运而生了。因此，在互联网世界 XML 的用途主要有两个：一是作为元标记语言，定义各种实例标记语言标准；二是作为标准交换语言，担负起描述交换数据的作用。

a) XML 的优势

XML 最大的特点是以一种开放的自我描述方式定义了数据结构，并在描述数据内容的同时能突出对结构的描述，从而体现出数据之间的关系。与 HTML 相比，XML 具有以下一些优势：

● XML 具有易扩展性。XML 和 SGML 一样都是元语言，可以定义其他的语言。使用 XML 设计自己的文档类型，可以定义自己的一套标记，而且这些标记不必仅限于对于显示格式的描述。XML 允许用户根据各种不同的规则来扩展制定标记，比如根据商业规则，根据数据描述甚至根据数据关系来制定标记，典型的数学标记语言 MATHML、财经标记语言 FPML、电子商务标记语言 EBXML 等。

● XML 具有灵活性。XML 提供了一种结构化的数据表示方式，使得用户界面显示分离于结构化数据。所以，Web 用户所追求的许多先进功能在 XML 环境下更容易实现。

● XML 语法简单，可以被所有的机器解读，又可以在各种平台上使用，通行性更强。

● XML 编程简单，是交换语言的首选。XML 舍弃了 SGML 的复杂性，因此编写处理

XML的应用程序会很容易。

● XML信息易于存储，可重复使用。XML使用非常简单的数据格式，可以用纯ASCII文本来书写。ASCII文本几乎不会“磨损”，可以重复读取和使用。

● XML文件在SGML环境中也可使用，不一定局限于在WEB中使用。因为，XML本身就是数据，可以由程序任意控制。同样的数据既可以在浏览器中显示，也可以交给一个代理进行后台处理。

b) XML中的主要概念

● DTD(文档类型定义)：描述了包含在任何XML词汇中的部件和准则。包括：组成词汇的元素、元素的属性、包含在用DTD写成的文档中的实体以及所有这些部件相互影响的规则。

● 元素：XML文档中的关键结构。XML文档中的任何内容都必须用文档DTD的元素描述。

● 属性：元素的修饰部分，为元素及其内容提供附加的描述性信息。

● 实体：数据单元。如：二进制数据、图形、声音文件等，它是XML中最有用的特性之一。

● 内容模式：用来描述标志是怎样嵌套的规则。

● XLink：XLink定义简单链接和扩展链接。XML使用多指向超链接，链接不存在于任何包含链接的文档中的文档(也称作不一致链接)，即一个XML文档中的任何元素都可以成为某种链接。XLink为独立于XML的新规范。

● Xpointer：Xpointer定义在URL中嵌入文件节点路径标记的方法，详细描述链接如何在文档里的指向各个方向，还可提供引用子资源的其他方法，包括定位和串值。

● XSL：格式单定义语言，具有将某种标记语言描述的文档转换成另外一种标记语言描述的文档的能力。

● 文档：包含由一个用来声明实体和其他内容事务的内部DTD、标记所描述的基于文本的内容以及标记本身。

c) XML的主要规范

XML Schema规范是XML的主要规范之一，分为三部分：总体介绍、结构和内容约束文档和数据类型定义等。其中，第一部分(XML Schema Part 0：Primer，http://www.w3.org/TR/xmlschema-0)是对Schema的总体介绍，目的是帮助读者快速理解如何利用Schema语法创建Schema文档[11]；第二部分(XML Schema Part 1：Structure，http://www.w3.org/TR/xmlschema-1)和第三部分(XML Schema Part 2：Datatypes，http://www.w3.org/TR/xmlschema-2)是对XML Schema语法规范的完整描述，其中前者为描述XML1.0文档的结构和内容约束提供了文档，而后者则为Schema及其他XML规范定义了数据类型。目前，XML Schema尚处于草案阶段。

XML的其他规范还包括：DOM、JDOM、SAX、XSL、XSLT、XML链接与引用等。

d) XML在数字图书馆中的应用

XML是一种将信息内容与信息描述分开的标记语言，因此，它非常胜任描述元数据的功能，特别是图书馆的目录组织。图书馆的目录是一个信息源，它同时又提供了另外一个信息资源的信息。另一个信息通常是一本图书、一本期刊或它们的电子产品。XML在图书