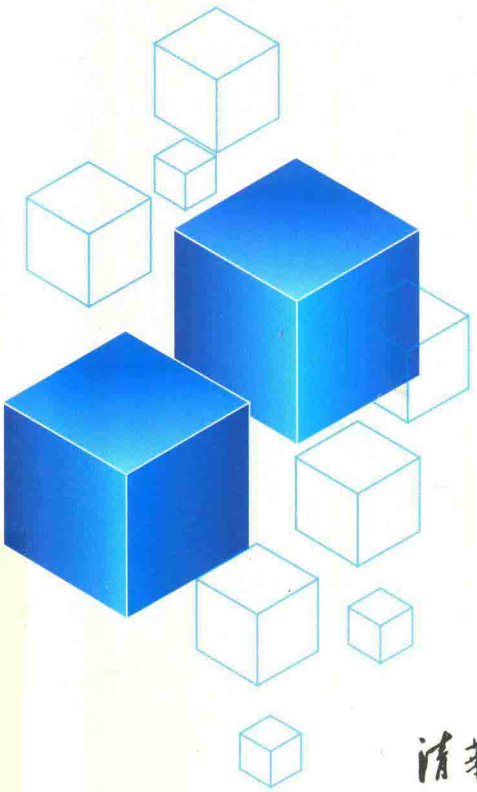
开放文档格式标准与互操作基础

李 宁 田英爱 侯 霞 等 著



清華大學出版社

国家社会科学基金项目资助

开放文档格式标准与互操作基础

李　宁　田英爱　侯　霞　等著

清华大学出版社
北京

内 容 简 介

本书以可扩展置标语言 XML 为核心，全面介绍了开放的文档格式标准与文档互操作理论及实践。其内容涵盖了置标语言技术、文档格式与相关标准、文档处理技术、文档互操作模型以及互操作性度量理论。此外，本书为方便读者学习与研究，还介绍了相关领域的基本知识、技术现状、研究成果以及未来发展方向等。

本书可作为高等院校计算机和信息技术相关专业的高年级本科生、研究生或教师的教材或学习资料，也可作为文档格式标准、文档处理技术、办公自动化、信息系统建设等相关领域科研工作者的参考用书。

图书在版编目(CIP)数据

开放文档格式标准与互操作基础/李宁等著. —北京：清华大学出版社，2014
ISBN 978-7-302-36467-2

Ⅰ. ①开… Ⅱ. ①李… Ⅲ. ①规范文档—介绍 Ⅳ. ①G254.36

中国版本图书馆 CIP 数据核字(2014)第 099798 号

责任编辑：付弘宇　薛　阳
封面设计：迷底书装
责任校对：白　蕾
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：三河市中晟雅豪印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：28.25　　**字　　数**：685 千字
版　　次：2014 年 10 月第 1 版　　**印　　次**：2014 年 10 月第 1 次印刷
印　　数：1～500
定　　价：79.00 元

产品编号：058043-01

前　言

文档作为信息和知识的载体，已经深入到人类社会的每一个角落：私人信件、政府公文、商业信函、合同契约、历史档案、科学论文、产品手册、媒体网页、电子读物……无一不是文档。随着数字化和网络的发展，如今信息处理已经逐渐从面向数据为中心的模式逐步发展到面向文档为中心的模式。文档信息处理技术涵盖文档的表示、编辑、转换、存储、检索、发布、印刷、出版、数字版权、数据流整合以及智能文档等各个方面。近年来，人们尤为关注以 XML 为基础的文档信息的结构化表示、文档格式的标准化与互操作、网络化办公、文档信息挖掘、语义网与智能文档的结合等新的技术发展动向。

如今，办公软件已经成为人们日常最常用的软件，产生了大量的办公文档。一些国家的政府组织发现，如果这些文档以某个私有公司的文档格式保存，那就意味着政府所有的公文都必须由该公司提供的软件才可以打开，而且在向社会提供政府服务的时候，还要求被服务者只有购买某个特定公司的软件才可以提供这种服务。作为政府公文，可能要保存很长时间，如果仅依赖一家厂商的产品，谁也无法保证这些文档可以获得持续的支持。这是很不合理的。

因此，美国马萨诸塞州首先通过了必须采用开放文档格式的政策，要求政府采购的软件必须支持开放的文档格式国际标准。这一政策得到了很大反响，包括美国其他州政府以及欧盟、亚洲、非洲等很多国家都认识到这个问题的严重性，纷纷发布政策，确保政府公文应当采用开放文档格式标准。

如今，有关文档的开放标准在全世界越来越受到重视，联合国、世界贸易组织、欧盟以及其他国家和地区都制定了相应的政策支持开放的格式标准。例如，欧盟 20 多年来一直致力于使用开放标准实现欧盟各国间文档信息的互联互通，曾经资助开发旨在方便应用程序间传递文档的开放文档体系结构标准(Open Document Architecture，ODA)，该标准 1985 年成为 ECMA 标准，1994 年前后成为国际标准化组织(International Standard Organization，ISO)标准(ISO/IEC 8613)。2003 年，欧盟又开展了对开放文档格式标准和市场发展走向的调查，并提出一些评价开放文档格式的准则。联合国也在 2005 年指出，所有成员国和有关人员都应该有权访问联合国各机构以电子格式提供的公共信息，任何人都没有义务为享有该权利而安装特定的软件。各机构应通过使用开放标准促进各 ICT(Information & Communication Technology)系统的互操作性，使人们无论使用何种软件都可以打开相应格式的文档。他们还应确保数据的编码可以满足公共电子记录的持久性，而不依赖于特定的软件提供商。一些国家(如英国)在 2005 年的电子政务互操作性框架中确定了以办公应用 XML 规范为核心的互操作性需求。2006 年，挪威就开始考虑在政府公文中采用开放的文档格式标准 ODF。2007 年 10 月，丹麦政府向 WTO(World Trade Organization)正式通告《政府基于开放标准软件采购的 WTO 声明》，指出目前的研究表明，开放标准对于产业的发

展和市场的竞争有积极的推动作用，要求从2008年起，所有的政府采购都必须基于ODF（Open Document Format）或OOXML（Office Open XML）标准。其他国家，包括荷兰、比利时、日本、南非和美国麻省州政府等国家和地区，都积极推动类似政策的制定和实施。

近年来顺应开放文档格式标准的呼声，出现了一系列开放的国际国内标准。例如，2007年成为国际标准的ODF（ISO/IEC 26300），2008年成为国际标准的OOXML（ISO/IEC 29500）。

我国文档信息处理的标准化工作开展得很早。2001年就开始了办公软件文档格式国家标准的研制工作。2007年“标文通”（Uniform Office Format，UOF）正式成为我国国家标准，并得到各主流办公软件产品的广泛支持。随后“标文通”标准一直在不断发展。“标文通”2.0草案已经于2010年年底完成。此外，我国正在制定面向电子公文的电子文件存储与交换格式标准，标准工作组已经确定将“标文通”2.0作为电子文件的文书类流式文档格式规范。

上述这些标准为文档互操作创造了条件，但是离真正达到互操作的目标还有很长的路要走。

本书的作者是一批长期从事文档信息处理的专家和学者，亲身参加了近年各类主要的与文档格式相关的国家标准的制定和应用推广工作。尽管文档信息处理是一个由来已久的研究方向，但是近年来文档处理技术的研究又得到前所未有的重视。我们在研究和教学实践中却很少见到一部能够全面反映近年文档信息处理技术发展成就的专业书籍，我们深切感到出版这样一部书的迫切性。恰逢2009年IBM公司将Faculty Award授予我们，给予我们宝贵的支持，于是我们几位作者自不量力，开始了本书的编写过程。经过4年的努力，终于得以付梓杀青。

本书的重点是开放的文档格式与互操作。本书以可扩展置标语言XML为核心，围绕文档格式与互操作，介绍文档信息处理的相关技术及其标准化状况。本书的目的是使读者了解相关领域的基本知识、技术发展以及最新成果。本书的读者可以是大学与研究机构中信息技术相关专业的本科生、研究生和教师，以及从事文档信息处理技术、办公自动化、信息系统建设的技术人员和管理人员。

本书共分为9章。

第1章为概述，主要讲解文档技术的相关概念，文档的分类，文档相关标准的概况以及文档处理技术的发展。

第2章为置标语言，主要介绍置标语言的概念与历史，置标语言在开放文档格式与文档互操作中的作用，关于“所见即所得”与式样内容分离的讨论以及基于XML的文档格式的设计方法。

第3章为文档格式基础，主要介绍流式文档与版式文档的主要区别，并分别介绍常见的流式办公文档及版式文档格式，重点介绍流式办公文档的基本构成要素。

第4章为主要的办公文档格式标准，剖析主要的办公文档格式标准ODF、OOXML以及UOF，并对这3种主流的格式标准进行比较和分析。

第5章为其他文档格式标准，较为详细地介绍EPUB电子书格式标准，并简单地对EPUB与其他电子书格式标准进行比较和分析；此外，还介绍一些其他的文档格式规范，包括HTML（XHTML、HTML5）、XSL-FO、MathML、SMIL、SVG等。

第6章为办公软件与文档处理，介绍主要的办公软件产品、文档的编辑处理与文档编

程，并简单介绍文档的自动化处理。

第7章为文档处理关键技术，主要介绍文档的一体化处理技术、跨媒体复合出版与交互、智能文档技术、文档安全与数字版权技术、国际化与本地化技术、信息无障碍技术、基于文档格式的文档内容理解等技术；此外，还介绍了文档处理技术的最新发展。

第8章为文档互操作技术，主要介绍文档互操作的概念、文档互操作的需求、影响文档互操作的因素、文档互操作的实现途径、文档格式转换工具及方法、文档互操作模型、文档互操作度量模型以及文档互操作的保障。

第9章为当前标准化工作与未来发展，主要介绍ISO/IEC SC34、OASIS、ECMA、W3C、标文通工作组、OpenOffice社区以及IBM等各大组织、机构、社团所开展的标准化工作，并勾勒出文档信息处理和互操作技术的未来。

本书第1章主要由李宁编写，第2章和第3章主要由侯霞、李宁和冯雪编写，第4章主要由侯霞编写，第5章主要由施运梅、刘旭红编写，第6章主要由田英爱编写，第7章主要由田英爱、李宁、冯雪、施运梅、刘旭红、梁琦、张伟编写，第8章和第9章主要由李宁编写。本书最后由李宁和田英爱进行了审校。此外，还有以下同志也参加了本书的撰写：罗文甜、张钰晗、王彦美、宋昊苏、李东明、高晓光、李杨、李娟、刘寅、高瑜蔚、许振伟、李秋玲、刘鹏等。中国电子技术标准化研究院、IBM中国公司和永中软件公司为我们提供了很多宝贵的资料和各方面的帮助，在此一并表示衷心感谢。

特别感谢IBM公司通过Faculty Award基金支持本书的编写出版。感谢贾明飞和田忠两位先生全程参与本书的讨论并提出中肯的建议。

本书的出版还得到了北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519)、北京市教委“计算机应用技术”重点建设学科专项项目(PXM2013_014224_000025)和网络文化与数字传播北京市重点实验室建设科研基地建设项目的支持。

我们深知无论我们如何努力，我们的才识都不足以完全达到本书预期的目标，但我们仍然希望把本书奉献给各位读者，疏漏或不当之处请读者批评指正。

编　者

2014年6月

目　录

第1章
概述

文档是信息的载体。文字的出现产生了书写体系，同时产生了文档。

显示在计算机屏幕或印在纸面上的信函、合同契约、教科书、报告、表单、备忘录、电子邮件、通讯录、电子书以及数以亿计的网页，都是文档，但是本书主要讨论的是电子文档。人们为了信息传递的目的，通常将数字化后的数据记载到文档之中，这样便形成了电子文档。

传统的文档以人类用户的信息交流为主要目的，主要记载人们可以感知的文本、图表等内容。文档主要以表现纸面介质上的内容为主。随着多媒体和人机交互技术的发展，文档记载的内容已经扩展到声、图、文、像等多媒体数据。近年随着互联网的发展，文档已经突破了人类用户信息交流的局限，文档不但可以被人理解，也可以被机器理解，一台机器上的多个程序可以通过文档传递信息，机器和机器之间也可以通过文档传递信息。

从计算机信息处理的角度看，文档表现为各种各样的数据。一般认为，可以把数据分为结构化数据，半结构化数据和非结构化数据。有时也称为规格化数据，半规格化数据和非规格化数据。结构化数据，就是其数据类型为人们所公知的那些数据，也就是那些具有基本数据类型的数据。例如，整型、实型和字符串类型等。这些数据一般均可以很容易地被程序语言或数据库所支持。半结构化数据，就是那些数据类型比较复杂，需要通过形式化模型加以说明的数据。例如，文档的章节结构并没有简单的数据类型与之对应，但是可以通过树状或层次模型进行定义，通过模型，一般可以把复杂的数据结构通过简单的数据类型表示出来。这种基于模型的数据类型说明也称为大纲(Schema)。计算机通过大纲便可以解析同类的数据。常见的文档多为半结构化数据。非结构化数据，就是那些数据类型未经形式化模型定义的数据。这类数据通常难以简单地映射到基本的数据类型之上，需要用特定的逻辑进行处理。需要指出的是，非结构化数据并不是杂乱无章的数据，一般都可以找到相应的规格说明解释其结构，只是没有形式化的大纲存在而已。声音、影像等多媒体数据多为非结构化数据。

在这几种形式的数据中，结构化数据最容易被机器处理，但是离人们对现实世界的概念最远，非专业人员不易理解；非结构化数据最接近人们对现实世界的理解，但是机器处理起来最为困难。半结构化数据介于这两者之间，起到了减少人机隔阂的作用。文档就是这样一种半结构化数据。

今天的文档，除了人类用于可以感知的内容之外，甚至还包括用于机器理解的文档的处理逻辑和用户界面的约定，例如，办公文档中的 VBA(Visual Basic for Applications)程序或

宏，以及网页中的脚本等。因此广义的文档可以包含现实世界所有的信息内容，可以预见，未来的文档、程序或数据之间不会再有明显的界限。

1. 文档的性质

一般来说，文档具有如下一些性质。

(1) 可描述性。文档必须能够充分地记录所要表达的内容。文档的描述方法常常分为从前向后的线性描述方式或以超文本为代表的非线性描述方式。

(2) 可保存性。信息必须有其存在的物理形式，例如，计算机文件系统中的文件就是常见的电子文档的保存形式。此外，文档也可以保存在数据库或 Web 之中。当对文档进行保存时，要考虑如何将文档的内容映射成计算机记录的数据，这就是文档格式。可以说，每个保存下来的文档都有一种特定的文档格式。

(3) 可理解性。文档的用户既可以是人类用户也可以是机器，或同时为这两者。不管是哪一种用户，都应该能够根据需要全部或部分地理解文档的内容，从而获得文档所表达的信息。文档的可理解性受多种因素影响，除了语言表达本身之外，还与阅读者的知识背景、阅读语境(Context)甚至排版式样相关。

(4) 可演化性。文档从创建之后可能经过反复修改，不但使内容得到丰富和完善，也可能演化出新的文档或新的表现形式。例如，维基百科(Wiki)从一个词条分化出新的词条，或将一部纸质的百科全书制作成为可检索的光盘(CD-title)。

(5) 功用确定性。文档一般用于有限的特定的目的。不同的使用目的对于文档的排版格式乃至描述方式等均有特殊的要求。一个文档虽然不能同时满足所有的目的，但是可以通过转换使之用于多种目的。

除上述性质之外，文档可能还具有一些其他性质，例如，可操作性，这是通过形式化的编程语言记录的文档所具有的性质，这些程序语言可能描述了该文档的一种处理流程，或者进行签名和验证的方法，这样使得文档的处理程序可以按照这种自描述的内容完成文档的流转或签名验证。除了上述的物理的、功能的和操作的性质之外，文档还留有丰富的社会和文化的印记。文档一般是用语言表达的，而语言是社会发展的产物，与民族和地区的文化密切相关。反映在文档之中有很多方面，例如，对于多民族文字符号的支持，纵向或横向的排版方式以及各种排版规则等。

2. 文档的生命周期

计算机是处理电子文档的工具。计算机对文档的处理包括文档的创建、编辑、排版、检索、获取、保存、发送、打印等，构成文档的生命周期。一般来讲，一个文档的生命周期主要包括[1,2]：

(1) 文档的准备。包括起草文档内容、设计版面、指定访问权限、命名文件等。

(2) 文档的生成。指的是从键盘输入文档内容，或从其他已有的电子文档导入内容。文档的输入可以通过友好的人机界面提高效率，例如，语音识别或光字符识别(Optical Character Recognition，OCR)。这个阶段的文档一般被称为草稿。

(3) 文档的编辑。指对文档的内容进行添加、修改、删除等操作。一个好的编辑工具往往会提供高效的、符合文档结构特点的编辑方法，例如，对文档的整块内容进行移动或复制，

对列表进行自动编号，对特定的内容进行全文替换等。文档编辑的最后还会涉及一个保存的过程。

(4) 文档的排版。将文档内容编排成清晰的显现样式，可以帮助读者更好地理解文档。文档的排版往往也作为文档编辑的一部分。一个好的编辑工具，也往往会提供高效的文档排版方法，例如，将所有标题或子标题按照特定的标题式样排成统一的风格，同时改变同类内容的显现式样等。

(5) 文档的审阅。文档初步编辑完成之后，一般要经过多次修改，在一个组织机构中，往往会送给不同的人员对文档的内容提出修改意见，或对文档中的错字错句进行检查。文档的审阅往往需要留下批注或修订信息。

(6) 文档的分发。文档编辑完成后，发送给使用者阅读。文档的分发可以通过网络进行，也可以通过交换存储介质来进行。文档的分发往往需要考虑文档的访问权限，一般借助办公自动化等业务系统进行管理。

(7) 文档的归档。当文档的使命结束后，文档或者被销毁(删除)，或者作为档案资料保存起来。往往需要将文档转换成适合长期保存的格式存储在可靠的介质中。

在整个文档的生命周期中，有多种角色参与其中。这些角色有些是个人，有些是组织。一些主要的角色如下。

(1) 创建者。文档的原始起草人。文档一般只有一个创建者。

(2) 作者。文档的编写人。文档可能由多个作者合作完成。

(3) 编辑。对文档进行编辑、加工和排版，使文档结构清晰，字句通顺，去除字面的错误。有时作者也是文档的编辑者，但作者主要关注文档的内容，编辑人员主要关注文档的可读性。

(4) 评审人。阅读文档，对文档进行评价，并对是否出版或分发提出建议，或对文档提出修改意见。

(5) 发布者。负责文档发布。将纸质或电子版文档发送到读者手中。

(6) 出版商。如果文档需要正式出版，需要向主管部门申请书号，筹措经费，确定印刷厂，对印刷质量进行监督，以及负责销售推广等。

(7) 消费者。文档的使用者，可能是人类用户也可能是计算机程序。使用者或者阅读文档，或者对文档进行解析和处理。

3. 文档处理

文档信息处理(简称“文档处理”)指的是采用计算机等信息处理工具，对文档进行分析获得所需的信息，并通过计算机对文档的生命周期的各个阶段进行管理。在学科分类上，文档信息处理和文本信息归属于计算机学科下的计算方法(Computing Methodology)[3]。在文档与文本处理之下还有下列分支[4]。

(1) 文档与文本编辑。

① 文档管理。

② 多语种。

③ 拼写。

④ 版本控制。

(2) 文档准备。

① 桌面出版。

② 格式和记法。

③ 超文本、超媒体。

④ 文档索引。

⑤ 多语种体系。

⑥ 置标语言。

⑦ 多媒体、混合媒体。

⑧ 排版。

⑨ 脚本语言。

⑩ 标准。

(3) 文档检索。

(4) 电子出版。

(5) 文档获取。

① 文档分析。

② 图形识别和理解。

③ 光字符识别。

④ 扫描。

2010 年,第一次全国文档信息处理学术会议将文档信息处理的范围定义如下。

(1) XML 技术与文档结构化表示:XML 相关技术;开放文档格式与标准化;文档模型与文档互操作技术。

(2) 文档处理技术:文档编辑技术;文档呈现、印刷技术;文档交互技术;国际化与本地化处理;多语种呈现和表达技术;文档处理软件体系结构设计;网络化办公与云计算;文档数据存储、共享与交换;文档与其他数字内容集成;文档压缩技术;自适应文档技术;文档工程生命周期;文档信息安全;文档处理系统性能评价。

(3) 文档智能:数据集的标注、检索和存储;智能文档与智能标签;基于语境的处理;文档协同处理;操作语义的描述方法;语义网技术;文档分类与文本挖掘技术;版面理解与逻辑结构提取;文字识别技术;信息无障碍技术。

(4) 文档应用:办公自动化;电子政务应用;电子商务应用;数字图书馆和档案馆应用;在云计算和移动计算等环境下的网络文档应用(Blog、Wiki 等);文档管理系统。

对文档信息进行处理的自动化或半自动化工具被称为文档处理器,由于文档处理的范围很广,文档处理器的种类也非常丰富。文字处理器(Word Processor)就是一种最常见的文档处理器。

在发明打字机之前,文档是手工书写的,文档在各种商业活动中使用,逐渐有了特定的写法和式样要求。19 世纪 70 年代诞生了手动打字机,20 世纪 60 年代初期,IBM 制造出电动打字机(Selectric Typewriter),使文档的撰写变得快捷。1964 年,IBM 又制造出磁带电动打字机(Magnetic Tape Selectric Typewriter),可以通过电子方式把击键的信息保存在磁带上,这样使文本内容得以编辑。有人认为磁带电动打字机是第一代自动化的文档处理器。之后众多的办公设备厂商便开始研制和销售文字处理或文档处理的硬件和软件。这类设备

可让用户借助CRT显示器浏览文档，在打印之前很方便地对文本进行管理、拷贝和排版格式设置。随着电子计算机的产生和发展，文档的存储介质变得更加丰富，如硬盘、光盘和记忆卡等。

真正意义上的文字处理器是在计算机产生之后才有的，主要是指对文档进行编辑等处理的专门的计算机程序。需要说明的是，虽然早期的文档是由文字组成的，但文字处理器并不直接处理文字，而是基于文档的基本组成部分对文档内容进行组织和加工。文档的基本部分一般包含段落、列表、页眉和页脚等，文字处理器可以对文档的这些成分进行统一处理，如统一设置所有段落的字体或改变所有页眉的内容等。另外，它还可能在一系列预设的规则下自动化地创建文档结构、对列表或页码自动编号，按照禁排规则进行分行，以及自动地产生与显现介质匹配的显现式样等，这些过程都不需要用户的过多介入。

1.1 文档的概念及分类

为便于读者理解本书后续各章节的内容，下面介绍文档的概念与分类。

1.1.1 文档的概念

前文提到，文档是信息的载体。计算机所处理的电子文档主要记录给人类用户和机器阅读和理解的信息，文档的内容按照一定的结构表示为数据，并按一定的格式保存成文件。

从计算机文件系统来看，文档与文件是等同的，对文档的处理与对文件的处理一样，包括创建、修改、删除、重命名等。

最早的电子文件是随着1952年计算机存储器出现而出现的，为与穿孔卡片上记录的形式相区别，第一次使用了文件(File)一词。但是那时的文件概念主要指的是硬件，比如人们称IBM 350计算机系统的硬盘驱动器为“磁盘文件”。1962年出现的分时操作系统(Compatible Time-Sharing System)在存储设备上能存储多个文件，从此文件一词才有了今天的概念。

从计算机的角度看，文档无论表示的是文本、图像还是声音，都是由一系列字节数据组成的。经过程序或浏览器的解释，用户才能真正理解这些数据的含义，获得文本、图像或声音的原貌。

从应用的角度看，文档按哪种顺序或结构记录信息并保存成数据文件很值得探讨。比如，有的记录方式很难修改，改变一个字节的数据，所有其他的数据都要重新计算；有的记录方式很难扩充，不允许随便增加其他的数据内容；有的记录方式与软硬件平台直接相关，换一种不同字长的机器或操作系统就无法正确理解这些数据了。另外，一个精心组织的文档，不但能够避免上述问题，还可能大大提高数据处理的效率。例如，存在索引信息的文档，可以从大量的数据中迅速找到所需要的文本或图像。

文档格式(或文档类型)是指计算机为了表示和存储信息而使用的对信息的特殊编码方式，是用于识别文档内储存的资料。每一类信息，都能够以一种或多种文档格式保存在计算机文件系统中，并通过文件名后缀或特殊的文件头部数据(简称“头信息”)加以标识，使得计算机系统可以调用适宜的文档处理器进行处理。

除了保存在文件系统中的文件格式之外，文档格式还有逻辑格式和内存格式等不同的含义。文件格式对应信息在文件系统中的存储方式(也称为“物理模型”)，例如，与文档数据相关的文件和目录的组织方式、打包压缩方法等，另外从内容和编码的角度也可把文件格式分为纯文本格式或二进制格式等(所谓纯文本，即不包含除文本内容之外任何其他信息的文档格式)；逻辑格式对应信息的概念模型，位于更高的层次，逻辑格式关心各种信息是如何表示并组织起来的，如线性的、层次的或网状的结构，以及压缩打包的算法和容器结构等；此外还有文档的内存格式，对应于数据模型，当文档在程序内部处理时，需要文档信息转化为内存的数据结构，如树或链表、类与对象等。

在本文中，若非特别说明，文档格式一般指文档的逻辑格式。此外，为了与文档存储格式相区别，文档排版中所用的格式，如字体设置、行距设置等称为排版格式。

1.1.2 文档的分类

文档作为信息交换的媒介，走过了长期的发展历程，文档承载的内容不断丰富；文档从单一用途转为多种用途；文档的利用方式和显现手段日趋多样。文档可以按多种方式进行分类。按记录内容分，可分为文本文档、图像文档和声音文档等；按用途分，可分为文字处理文档、电子表格文档和演示文稿等；按编码方式分，可分为二进制文档、纯文本文档；按开放程度分，可分为开放格式文档和私有格式文档等。

还可以按文档承载信息的抽象程度把文档分为如下层次，见表 1-1。

表 1-1 文档信息抽象层次

层 次	典型格式	抽象度
结构化数据(字符串)	DB(数据库)	很高
半结构化数据	XML	高
流式(办公)文档	ODF/OOXML/UOF	中
固定版式文档	PDF/CEBX/SEP	较低
非结构化数据	TXT/BMP/JPG	最低

作为记录语言的书面文字，一般采用纯文本方式来记录逻辑内容。从书面文字到数据模型，是为了信息处理的目的对文档信息进行抽象、提炼的过程。这一步主要是从书面文字中获得有价值的信息，进而转变为计算机可以理解的数据模型，进行更深层次的加工和理解。当前，从书面文字到数据，广泛采用置标语言来实现(参见第 2 章)。例如，将书面文字标注出分词和语法成分，可以通过规则进行自然语言理解；通过对信件的标注，获得收信人、发信人、发信日期、主题内容等信息，进而转换成关系模型存储到数据库之中，以便进一步利用。

文档处于半结构化数据层次。书面文字加入式样之后，将有助于清晰而直观地显现内容，利于人类阅读。这个层次的文档，还包含反映语义的式样信息，例如，章节、表格和列表，它们可以看作是书面文字的一种特殊表达形式，而不仅是感观上的式样。另外，这类文档还包含复杂的编辑语义，这是供编辑工具理解而使用的一种特殊信息模型，例如，章节等内容的嵌套层次，索引和链接关系，内容的前后阅读顺序，正文与批注的划分等，这些信息为文档编辑提供了很大方便，如果改变一个章节的字体只需要设置章节的式样，而不需要对章节的