

数字图书馆技术与应用

TECHNOLOGY AND APPLICATION OF DIGITAL LIBRARY

田捷 编著

科学出版社

数字图书馆技术与应用

田 捷 编著

科学出版社

2002

内 容 简 介

本书是关于数字图书馆原理、方法、技术及应用方面的书籍。

全书共 11 章。介绍了国内外有关数字图书馆的概念、起源与发展,数字图书馆的构成,数字图书馆的技术原理与实现方法,数字图书馆的应用和典型案例,可以让读者对数字图书馆的实现方法与应用前景有一个直观的认识。

本书内容新颖,学术水平高,文字流畅,可读性和实用性强,是广大图书馆工作者、文献情报工作者、计算机软件开发以及互联网从业人员必不可少的工具书,也是高等院校信息管理专业难得的一本教学参考书。

图书在版编目(CIP)数据

数字图书馆的原理和方法/田捷编. —北京:科学出版社, 2002

ISBN 7-03-010378-5

I.数... II.田... III.数字图书馆—基本知识 IV.G250.76

中国版本图书馆 CIP 数据核字(2002)第 024989 号

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2002 年 6 月第 一 版 开本: 787×1092 1/16
2002 年 6 月第一次印刷 印张: 22 1/2
印数: 1—4 000 字数: 511 000

定价: 34.00 元

(如有印装质量问题,我社负责调换(环伟))

序（一）

1998年8月，在国务院、文化部、科技部领导的肯定和支持下，我有幸牵头组织了中国数字图书馆工程的研究论证工作，在工作中结识了汪成为、张效祥、朱高峰、石元春、胡启恒、李未、李国杰、杨芙清、戴汝为、唐世渭、张钹、王雨生等院士、专家，得到了他们对中国数字图书馆工程的支持，并与高文、樊建平、刘峰、饶戈平等青年学者一起组成了国家863计划中国数字图书馆发展战略组，专门对数字图书馆涉及的技术、法律、运营、管理等问题进行了深入的研究，从而对一个完全陌生的领域有了一定的了解和认识。

三年过去了，伴随着数字图书馆的实践，理论研究也取得了可喜的成绩，出现了一些有关数字图书馆的专著。田捷博士的《数字图书馆技术与应用》一书就是这样一本关于数字图书馆的最新专著。田捷博士是一位计算机领域的专家，他所带领的团队从事数字图书馆技术研发已多年，田捷博士本人还相继在北京大学等高校开设了数字图书馆技术课程。这本专著正是他多年对数字图书馆原理的思考与实践总结。相信这本书的出版，对从事数字图书馆建设和理论研究的同行会有一定的指导和借鉴作用。田捷博士的专著让我作序，有点勉为其难，但盛情难却，我就谈点自己对数字图书馆的认识吧。

1. 要从构建新世纪中国文化的高度重新认识中国数字图书馆的时代意义

要理解数字图书馆，首先要理解网络，那么，什么是网络呢？简单地说，网络就是计算机，网络就是电信、电视、计算机三网合一，网络就是宽带高速，也就是人们常常比喻的信息高速公路。

网络的发展速度是前所未有的，从它真正运用于民间到今天的无所不在，前后不止10年。由于网络已经渗透到了社会生活的各个领域，并正在彻底改变着人们的学习、工作、生活、娱乐方式，创造出了全新的数字化生存模式，在我们承认这一现实的前提下，网络这条高速公路上跑什么车，车里装什么货——信息内容就显得至关重要。

网络是一项人类文明的成果，但同时它又是一把双刃剑，需要人类运用充分的智慧去驾驭。不然，网络就会如同一个打开了的潘朵拉魔盒，散布出邪恶，危害人类社会的安全。目前由网吧现象引起人们对互联网的关注，不正是互联网散布的黄色、迷信、反动的内容，消极、漫骂、缺乏诚信的氛围，以及它如同毒品一样使人迷恋以至沉沦给人们造成的恐慌的反映吗？这正是网络不健康的内容和负面影响造成的。网络的核心是内容，网络文化的内涵由网络内容所决定。以什么样的信息内容去占领互联网这个新兴的思想、政治、文化新阵地是问题的关键。

所以，建设数字图书馆，就是开发网络信息资源，主要抢占互联网阵地的积极有效的措施，数字图书馆就是网络内容建设王牌中的王牌，因为它是根据国家、社会需要，对中文基础信息、知识再加工后建立起来的具有中国特色的强大的信息资源库。

我们中华民族有着光辉灿烂的历史,文化信息资源丰富,在信息领域通过跨跃式发展,赶超西方发达国家,从而依托互联网建立起强势的中华文化是完全可能的。因此,我们要站在“三个代表”的高度,从重新构建网络条件下 21 世纪中国文化的高度来理解数字图书馆的时代意义。

2. 正确认识数字图书馆

数字图书馆是一项系统工程,从三年多的实践来看,首先要弄清数字图书馆是什么。

现在还说这个问题,也许显得可笑,但就是对这个问题的理解的不同,会直接地影响到数字图书馆的设计和实施,美国数字图书馆建设早期,主要依靠计算机通信领域的专家和技术人员具体实施,科学研究与工程建设几乎完全脱节,所以其研究成果有些不切实际;法国国家图书馆在数字化建设中,由于混淆了传统图书馆与数字图书馆的区别,设计、实施全由图书馆自己承担,结果出现问题,只有推倒重来,造成巨大浪费。这两种倾向在我们的数字图书馆建设中也是存在的,我们必须极力加以克服。

早在 1998 年 8 月,我就从组织实施的角度指出,中国数字图书馆工程是一项跨行业、跨部门、跨地区的国家重大工程,单靠哪一个部门都是无法完成的。中国数字图书馆工程的实践更充分说明,我们必须充分理解科技与文化结合的重大意义,跳出部门和行业的局限,站在国家整体利益的角度来共同推进数字图书馆建设。

较之过去大家对于数字图书馆的认识还仅仅局限于书面的概念,那么,今天数字图书馆已经有了不同的模式可供比较了。较之专业化的定义,我对数字图书馆的通俗理解是:数字图书馆是数字化的信息资源库。信息资源库也就是信息数据库,它应有以下几个特性:①分布的、但在统一的标准下建设;②可以在统一的网络平台上运行;③可以不断扩展。从这些特性可看出,数字图书馆不是传统图书馆的数字化,但传统图书馆的数字化是数字图书馆重要的信息来源和组成部分。明白了这个道理,也就明白了图书馆不是图书馆的专项,信息资源的开发是全社会的事,传统图书馆只不过在占有图书资料方面占有先机。这也可以理解,为什么现在有不同的单位在搞数字图书馆。虽然具体的技术途径和解决方案不尽相同,但无论何种形式,都是基于网络平台的海量数据库,并以网站的形式表现出来,从这个角度来看,数字图书馆则是网站上支持跨库检索等多种搜索功能的强大的海量数据库。

3. 实施数字图书馆面临的问题

(1) 技术问题。高文博士已经将数字图书馆所面临的技术问题归纳为信息资源建设、存储与压缩、分类、索引和检索、安全性、用户界面等 10 个技术挑战。这都是亟需解决的,其中最关键的,我认为是统一技术标准。中国数字图书馆的资源库是分散建设的,为了达到全社会的共享并避免重复建设,必须有一个国家认可、大家共同执行的技术标准。在这方面,国家 863 计划中国数字图书馆发展战略组应该承担起责任,搞好示范试点工程,从中总结经验并加以推广。

(2) 法律问题。数字图书馆的法律问题,实质是版权和知识产权在网络时代的运用问题,这个问题在数字图书馆的建设中已日渐突出。王蒙等著名作家状告网站侵

权,以及文化部中华文化信息网的建设中,均碰到了版权的问题,如何做到既保护作者的知识产权,同时又可将各类文化、科技的文明成果纳入数字图书馆,以使其为更多人服务,创造出更大的价值,这是一个急迫的任务。目前国际上也没有解决。我们要抓住机遇,随着加入 WTO,主动参与网上法律规范。

(3) 人才问题。信息化人才是国家信息化体系的六大要素之一。数字图书馆是一项高科技与文化结合的重大工程,必须选择懂行,而且具有创新意识的年轻人来承担。在推进中国数字图书馆的过程中,要不断发现和培养既懂技术,又具有深厚文化底蕴的复合型的信息化人才。有关方面已经招收了数字图书馆方向的研究生,这是一个良好的开端。我们在实际工作中,还应进一步解放思想,打破传统的用人机制,以利于人才的脱颖而出和事业的发展。

(4) 运营模式问题。信息化是当代世界经济和社会发展的的大趋势,党的十五届五中全会指出,要加强信息资源的开发和利用,要用“信息化带动产业化”。

文化信息资源的开发利用是数字图书馆的核心,实现文化与科技的结合,强强联合,用市场化操作的方式推进数字图书馆建设。国家图书馆、中央党校数字图书馆建设正在进行积极的探索。

如何在既保证社会利益的同时,又保证经济效益,这是我们必须面对的一个问题。早在 1999 年,人大、政协组织专家考察国家图书馆数字图书馆建设时,许多专家就提出:国家图书馆的藏书是国家投资购买的,是公益事业,在不考虑版权的前提下,把这些图书上网并收费,如何体现公益呢?

这个问题对于一些依托传统图书馆建设的数字图书馆是非常重要的,必须给予重视。如何既保证图书馆的公益性,又通过适当的市场运作模式实现一定的经济效益,从而促进数字图书馆的良性发展,是构建中国数字图书馆运作模式的基础。我认为,进行数字图书馆建设,必须进行思维创新和体制创新,在具体的实践过程中不断完善,只有这样,才能回答各种疑问,保证数字图书馆的健康发展。

徐文伯

(国家 863 计划中国数字图书馆发展战略组
组长、文化部原副部长)

序（二）

在上世纪末全球信息高速公路热潮中，数字图书馆成为许多发达国家关注的焦点之一。这是国家信息基础设施建设不可缺少的重要内容，也是新形势下国家间资源争夺的一个新领域。对学术界来说，数字图书馆的教学和研究也逐步被纳入一部分高等学校和研究机构的计划，一些专业和课程开始涉及这个主题。例如，北京大学信息管理系除了在研究生层次开设数字图书馆专题以外，又在本科两个专业的教学计划中开始引入这方面的主题或课程。而且，自 2001 年起，在本科专业中专门设置了数字图书馆和网络信息资源管理方向，以便向学生提供较系统的数字图书馆方面的知识，为我国信息界和图书馆界培养亟需的具备数字图书馆建设和管理专业素质的人才。

本书的编写和出版就是为了满足数字图书馆教学和研究的需要。作者田捷博士等人近年来从事数字图书馆的研究开发工作，承担了多项重要的科研项目，并且把数字图书馆的研究与开发应用紧密结合在一起，取得了许多在同行中具有先进水平的研究成果。这部教材既是田捷博士的研究团队多年来出色的研究工作成果的一部分，也是他们与北京大学信息管理系合作的一项成果。其中不少内容都已经在教学中得到了检验。本书涉及面广，体系合理，内容丰富新颖，组织剪裁得当，既有宽度又有深度，理论、技术与实例融合为一体，很有利于学生全面、系统地学习掌握数字图书馆的知识和技能。我们相信，此书的出版不仅会促进数字图书馆教育的培训工作，也会吸引更多的人（尤其是青年学生）投身于这个领域，振兴和繁荣我国的数字图书馆的研究、开发和应用事业，使我国的信息资源开发利用水平有一个突破性的进展。

赖茂生

（北京大学信息管理系副系主任
教授、博士生导师）

前 言

两年前,数字图书馆在我国还只是停留在概念阶段,今天,数字图书馆的实施已成星火燎原之势。一些重点项目,如中国数字图书馆、中央党校数字图书馆等相继启动。在学术界,伴随着数字图书馆概念的研究、技术攻关以及工程实施的不断深入,涌现出了不少优秀的著作。本书是我们在开展数字图书馆教学、科研、技术攻关以及工程实施过程中的经验总汇,很久以来一直期盼能有机会与广大用户及学术同行共享,这是本书能和读者见面的最初动因。另外,要特别提到的是,北京大学开各院校信息管理专业之先河,专门为硕士研究生开设了数字图书馆专题课程,本书也是我两年来在北京大学授课内容的总结。

数字图书馆大致从20世纪90年代起被一些发达国家重视和大力推进。中国数字图书馆研究的起步基本与国际同步,但在发展上比发达国家落后好多年,这也是整个国内计算机普及程度和互联网大环境所致。90年代中后期,中国在互联网用户呈几何级数增长,网站的数量也不断增加。随着人们对互联网需求的增长和深入了解,其中的一些缺陷和瓶颈都一一浮出水面,最突出的就是网络上有价值信息数量缺乏与有用信息的封闭性之间的矛盾。如何让两者更好地结合以提升双方的价值成了迫在眉睫的问题。

在这种情况下,国家将数字图书馆的研究列入了“863计划”,相继启动了一些试验性的数字图书馆项目。到20世纪90年代末,这些研究陆续产生了成果,国家开始投入大量资金推广数字图书馆技术。近两年,提供数字图书馆全面和单项方案的公司越来越多,一方面是由于市场的需求,另一方面得益于研究成果的公开和推广。但我们也发现,市场上比较成熟的数字图书馆解决方案绝大多数是外国公司提供的,比如沈阳和上海数字图书馆都是应用IBM的解决方案,而国内自主研究开发的很少。数字图书馆采用的技术非常复杂,这或许是国内企业自主研究开发数字图书馆困难的原因之一;而一些学术著作将技术分割得过于零散,对国内数字图书馆的发展不能起到很好的指导作用。通过两年技术与实践总结,形成了本书的结构主线:海量信息的快速采集、海量信息的组织和检索、信息的网络传播与数字发布和信息安全。其中尤以信息的检索为重点,因为数字图书馆的最终目的是实现信息的共享,即为大众服务,而检索是信息使用的必备手段。另外,本书突出了信息的“海量”特性,这是数字信息时代的基本特点,是今后技术发展的重要方向。本书还提供了数字图书馆的参考方案,使得枯燥的技术可以和实践结合起来,不仅对学生有很好的指导作用,对企业 and 用户更是一本指南性的读物。

这本书能顺利出版,首先要感谢科学出版社的大力支持。另外还要感谢本书的合作者:浙江大学的庄越挺教授(第8章与第9章),中科院计算所的刘斌,中国科学院自动化所的陈宏、何余良,北京大学的罗丽丽。数字方舟信息技术有限公司的区咏梅、李小冬、何震也给我很大的支持和帮助。特别致谢区咏梅,在本书的规划、写作和出版过程中,她分担了大量繁重的工作。这本书是大家共同心血的结晶,是我们对我国数字图书馆事业的一点贡献。

编著者

目 录

第 1 章 数字图书馆概述	1
1.1 数字图书馆的概念	1
1.2 数字图书馆的产生和发展	2
1.2.1 数字图书馆产生的背景	2
1.2.2 世界范围数字图书馆的发展状况	3
1.2.3 中国数字图书馆的发展	10
1.3 数字图书馆的构成	13
1.3.1 数字图书馆的技术构成	13
1.3.2 数字图书馆的体系构成	13
第 2 章 信息采集	15
2.1 文本信息的特点和处理	15
2.1.1 文本信息特点	15
2.1.2 获得文本信息的方式	16
2.1.3 文本信息的处理	16
2.1.4 常见文件格式	18
2.2 图像信息的特点和处理	22
2.2.1 图形和图像	22
2.2.2 描述图像的技术参数	23
2.2.3 图像信息的主要特征	24
2.2.4 图像的产生途径	24
2.2.5 图像处理	25
2.2.6 图像格式	30
2.3 音频信息的特点和处理	32
2.3.1 音频信号的特点	32
2.3.2 音频信号相关技术参数	35
2.3.3 音频信号的处理	36
2.3.4 音频文件格式	46
2.4 视频信息的特点和处理	48
2.4.1 数字视频信息特点	49
2.4.2 视频有关技术参数	52
2.4.3 视频信息处理	53
2.4.4 常见的视频文件格式	53
第 3 章 文种的处理技术	58
3.1 ISO646	58
3.1.1 GBI988 的代码构成	58

3.1.2	GB1988 代码的应用	59
3.2	GB23121-80	60
3.2.1	七位编码字符集的扩充方法	60
3.2.2	GB2312-80 代码的结构	61
3.2.3	GB2312 代码表中汉字的选择、分级和排列	62
3.3	ISO/II10646	63
3.3.1	ISO10646 标准提出背景	63
3.3.2	ISO/IEC10646	64
3.3.3	GBK	67
第 4 章	海量信息的组织技术	69
4.1	MARC 数据和 Z39.50 标准	69
4.2	Dublin Core 标准的发展	79
4.3	XML	91
第 5 章	海量信息压缩技术	117
5.1	图像压缩	117
5.1.1	通用图像压缩方法	117
5.1.2	黑白二值图的压缩: JBIG 和 JBIG2	120
5.1.3	基于小波分析的图像压缩方法	123
5.1.4	新的图像压缩标准: JPEG 2000	132
5.2	视频动态压缩	133
5.2.1	MPEG1	134
5.2.2	MPEG2	142
5.2.3	MPEG4	146
5.2.4	MPEG7	152
5.2.5	H.261	152
5.3	音频动态压缩 (MP3)	156
第 6 章	海量信息的存储	161
6.1	海量信息存储结构	161
6.1.1	第三级存储器	161
6.1.2	三级存储器系统	165
6.2	海量信息的基本检索方式	168
6.2.1	TPO 查询优化处理器的结构	168
6.2.2	优化器	169
第 7 章	信息检索、过滤与挖掘	171
7.1	信息检索技术	171
7.1.1	计算机信息检索系统的历史与发展	172
7.1.2	计算机信息检索系统的构成	173
7.1.3	多媒体信息检索技术	176
7.2	全文搜索引擎	177

7.2.1	网络信息检索系统的分类	177
7.2.2	网络信息检索系统的性能指标	178
7.2.3	搜索引擎的核心技术	179
7.2.4	网络信息检索新技术	185
7.2.5	未来动向	187
7.2.6	搜索引擎 Google	189
7.2.7	发展展望	190
7.3	网络信息过滤	192
7.3.1	基于内容的过滤	192
7.3.2	协同过滤	193
7.4	网络信息挖掘	195
7.4.1	内容挖掘	195
7.4.2	结构挖掘	198
7.4.3	用户访问挖掘	199
第 8 章	多媒体检索技术	205
8.1	多媒体视频和音频特征	205
8.2	视频检索	207
8.2.1	视频的结构化	208
8.2.2	基于相似度的视频检索	222
8.2.3	视频检索中的相关反馈	227
8.3	音频检索	227
8.3.1	音频分割	228
8.3.2	音频识别模型	234
8.3.3	音频与音乐检索系统	242
8.4	多媒体检索的未来趋势	254
8.4.1	多媒体融合检索	254
8.4.2	互联网络多媒体信息检索	256
8.4.3	网络交互模型	258
第 9 章	数字图书馆的信息安全技术	261
9.1	网络安全	261
9.1.1	网络安全的定义	262
9.1.2	网络安全的目标	263
9.1.3	网络安全的层次划分	264
9.1.4	网络安全面临的主要攻击	265
9.2	访问控制技术及信息加密技术的应用	268
9.2.1	访问控制技术	268
9.2.2	信息加密技术	273
9.3	PDF MERCHANT	283
9.3.1	PDF 简介	283

9.3.2 PDF MERCHANT 技术.....	284
9.4 数字水印技术	288
9.4.1 数字水印理论	288
9.4.2 数字水印典型算法	295
9.4.3 数字水印所面临的攻击和水印测试.....	296
9.4.4 数字水印技术标准化.....	298
第 10 章 数字图书馆的应用	300
10.1 数字图书馆与传统图书馆业务流与数据流的区别	300
10.1.1 业务流的区别	300
10.1.2 信息流的区别	301
10.2 公众服务	302
10.2.1 中国数字图书馆	302
10.2.2 中国试验型数字图书馆.....	304
10.2.3 上海数字图书馆	307
10.2.4 中央党校数字图书馆.....	308
10.3 电子商务服务	311
10.3.1 网络出版和电子图书 (e-book)	311
10.3.2 e-book 阅读器	315
第 11 章 方案介绍.....	319
11.1 IBM 数字图书馆.....	319
11.1.1 DL 的功能	319
11.1.2 数字图书馆的拓展结构.....	321
11.1.3 视频加载服务器的主要功能部件.....	326
11.2 Adobe ePaper 解决方案.....	328
11.2.1 电子文档的制作.....	329
11.2.2 电子文档的发布.....	332
附录.....	334
参考文献	339

7.2.1	网络信息检索系统的分类.....	177
7.2.2	网络信息检索系统的性能指标.....	178
7.2.3	搜索引擎的核心技术.....	179
7.2.4	网络信息检索新技术.....	185
7.2.5	未来动向.....	187
7.2.6	搜索引擎 Google.....	189
7.2.7	发展展望.....	190
7.3	网络信息过滤.....	192
7.3.1	基于内容的过滤.....	192
7.3.2	协同过滤.....	193
7.4	网络信息挖掘.....	195
7.4.1	内容挖掘.....	195
7.4.2	结构挖掘.....	198
7.4.3	用户访问挖掘.....	199
第 8 章	多媒体检索技术.....	205
8.1	多媒体视频和音频特征.....	205
8.2	视频检索.....	207
8.2.1	视频的结构化.....	208
8.2.2	基于相似度的视频检索.....	222
8.2.3	视频检索中的相关反馈.....	227
8.3	音频检索.....	227
8.3.1	音频分割.....	228
8.3.2	音频识别模型.....	234
8.3.3	音频与音乐检索系统.....	242
8.4	多媒体检索的未来趋势.....	254
8.4.1	多媒体融合检索.....	254
8.4.2	互连网络多媒体信息检索.....	256
8.4.3	网络交互模型.....	258
第 9 章	数字图书馆的信息安全技术.....	261
9.1	网络安全.....	261
9.1.1	网络安全的定义.....	262
9.1.2	网络安全的目标.....	263
9.1.3	网络安全的层次划分.....	264
9.1.4	网络安全面临的主要攻击.....	265
9.2	访问控制技术及信息加密技术的应用.....	268
9.2.1	访问控制技术.....	268
9.2.2	信息加密技术.....	273
9.3	PDF MERCHANT.....	283
9.3.1	PDF 简介.....	283

9.3.2 PDF MERCHANT 技术.....	284
9.4 数字水印技术	288
9.4.1 数字水印理论	288
9.4.2 数字水印典型算法	295
9.4.3 数字水印所面临的攻击和水印测试.....	296
9.4.4 数字水印技术标准化.....	298
第 10 章 数字图书馆的应用	300
10.1 数字图书馆与传统图书馆业务流与数据流的区别	300
10.1.1 业务流的区别	300
10.1.2 信息流的区别	301
10.2 公众服务	302
10.2.1 中国数字图书馆	302
10.2.2 中国试验型数字图书馆.....	304
10.2.3 上海数字图书馆	307
10.2.4 中央党校数字图书馆.....	308
10.3 电子商务服务	311
10.3.1 网络出版和电子图书 (e-book)	311
10.3.2 e-book 阅读器	315
第 11 章 方案介绍.....	319
11.1 IBM 数字图书馆.....	319
11.1.1 DL 的功能	319
11.1.2 数字图书馆的拓展结构.....	321
11.1.3 视频加载服务器的主要功能部件.....	326
11.2 Adobe ePaper 解决方案.....	328
11.2.1 电子文档的制作.....	329
11.2.2 电子文档的发布.....	332
附录.....	334
参考文献	339

第 1 章 数字图书馆概述

1.1 数字图书馆的概念

随着计算机和互联网技术的发展,网络已快速地渗透到社会生活的各个方面,围绕网络展开的各种应用也层出不穷。20 世纪 90 年代起,人们对网络多媒体信息的需求迅速增长,而宽带技术打破了多媒体信息的网络传输瓶颈,人们可以不到传统图书馆、音乐厅和电影院,通过网络就可看书、听音乐、看电影。

技术可能性的提高却暴露了网络信息需求和供给的巨大缺口,在中文信息方面尤甚。目前互联网上只有不到 10% 的信息是中文的,这与中文用户的数量极不相称。信息供给不足已经成为制约网络发展的主要因素之一。

为了丰富网络内容,人们需要不断地搜集信息,对它们进行加工使之适用于网络(这一过程可称为信息的数字化),并最终将它们放到网上。在各种信息的搜集、加工、上网的过程中,人们发现了一系列问题:如何合理有效地对海量数字化信息进行组织、检索、访问和利用?如何通过网络向用户提供这些信息服务?在研究解决这些问题的过程中,产生了“数字图书馆”的概念。

20 世纪 90 年代初,美国科学家提出了 Digital Library 的概念,可直译为“数字图书馆”。其实,英文 library 不仅有“图书馆”的意思,也有“库”的意思。从 Digital Library 反映的内涵来看,“库”更接近其本义。那么,其中文含义应该是“数字化资源库”,这个“库”里不仅有传统图书馆里的图书,也包含了其他所有可以数字化的信息。

虽然 Digital Library 与传统图书馆没有必然联系,但是由于目前资源主要集中在图书馆,因此必然要借鉴图书馆的已有理论和信息管理,因此在中文名称上本书仍然沿用“数字图书馆”。

迄今为止,国内外对数字图书馆的定义有很多种。

“数字图书馆”的概念包含这样一些要素:

- 数字图书馆不是一个单一的实体。
- 数字图书馆要求有连接各种资源的技术。
- 数字图书馆和信息服务之间的连接对最终用户是透明的。
- 数字图书馆和信息服务的通用入口是一个目标。

比较典型的数字图书馆的定义包括:

数字图书馆的收藏不限于文献资料,“它们可延伸到不能由印刷格式显示或区分的数字人造物。”(Drabenstott, Karen M 1994 年)

“无论是什么形式的数字媒介(文本、动态视频、音频、图形或图像),数字图书馆都可以为它们提供一个媒体资产解决方案,包括充分的存储空间、升级、速度、多级权限管理、高级搜索技术以及互联网入口,以连通新市场,保护资产不受损失或贬值,

保障知识产权。”(IBM 1999 年)

“数字图书馆是采用现代高新技术所支持的数字信息资源系统,是下一代因特网上信息资源的管理模式,它将从根本上改变目前因特网上信息分散不便使用的现状。”(孙承鉴,刘刚 2000 年)

“数字图书馆是以电子格式去存储海量的多媒体信息并能对这些信息资源进行高效的,如插入、删除、修改、检索、提供访问接口和信息保护等。”(高文等 2000 年)

“数字图书馆是社会信息基础结构中信息资源的基本组织形式,这一形式满足分布式面向对象的信息查询需要。”(刘炜,刘年娣 2000 年)

“所谓数字图书馆,就是对有价值的图像、文本、语音、影视、软件和科学数据等多媒体信息进行收集、组织和规范再加工,通过网络提供高速横向跨库连接的多媒体信息存取服务,促进社会各类信息高效、经济地传递,从而极大地方便人们的学习、交流和生活。”(刘峰 2000 年)

还有一些定义如“数字化图书馆就是图书馆在线服务系统”;“数字图书馆就是以数字形式存储和处理信息的图书馆”;“数字图书馆是指图书馆所有的工作流程都基于计算机,而且馆藏资源都实现数字化”等,此处不再一一列举。

以图书馆业务为例,数字图书馆与传统图书馆、自动化图书馆就有很大区别,如表 1.1 所示。

表 1.1 数字图书馆与传统图书馆、自动化图书馆的区别

	数字图书馆	传统图书馆	自动化图书馆
工作中心	用户	馆藏	馆藏
馆藏形式	数字信息资源	印刷型	印刷型及少量电子出版物
工作方式	对文献内容进行自动化加工	手工作业	对书目数据及专题数据库进行自动化加工
检索手段	对文献内容进行自动检索	手工检索卡片	对书目数据及专题数据库进行自动化检索
服务对象	面向全球读者提供网上服务	为到馆读者服务	以到馆读者服务为主,在一定范围内提供文献传递服务
馆藏加工	加工,并使馆藏具有增值效应	不加工	基本不加工

图表来源:《国内数字图书馆研究述评》,刘炜,刘年娣,2000 年

本书将基于广义的数字图书馆定义,并对其技术和应用进行进一步的阐述。

1.2 数字图书馆的产生和发展

1.2.1 数字图书馆产生的背景

计算机与互联网络技术的发展使“地球村”由概念变成现实,全球信息一体化和全球经济一体化的进程大大加速。事实上,互联网触发了人类有史以来最广泛、最深刻的

变革。一方面互联网已融入到人们的日常生活之中, 另一方面人们对个性化信息提出了更高的要求。

我国的因特网市场规模正以每年超过 200% 的速度增长。但是, 网上中文信息的匮乏、组织的无序大大降低了其实际的应用效率。目前网上信息资源爆炸式地增长, 亟需新型信息管理模式加以组织, 才能避免人类被信息垃圾“淹没”的危险; 同时为了满足人们利用互联网提供的服务进行学习和创造, 迫切需要大力建设网上优质信息资源, 使得互联网真正成为人类进步的加速器。

1.2.2 世界范围数字图书馆的发展状况

1. 美国

美国 1993 年率先开始数字化图书馆研究。美国的数字化图书馆项目由白宫总统的信息基础设施特别工作组和竞争政务会决定纳入国家信息基础设施虚拟图书馆 (NII Virtual Library) 中, 列在美国全球资源项目 (GLP-us) 之下。

美国的 NII 计划中, 在其早期已经形成了采用数字图书馆为“下一代因特网”信息组织方式的主流方案。目前, 美国数字图书馆建设规模已相当大, 仅在 2000 年, 从获得批准资助并加以公开宣布的正在进行的资源数字化及数字图书馆项目来看, 达 400 项左右, 批准金额超过 1.3 亿美元。美国数字图书馆建设的管理机构主要是国家科学基金会 (NSF), 技术实施主要由国家研究创新公司 (CNRI) 进行协调。

美国现有分布于各地的八个数字图书馆研究中心, 六个国家级数字图书馆试验基地。此外, 圣地亚哥 SDSC 国家超级计算机中心的任务之一即是作为美国数字图书馆群与科学数据库集成服务的试验基地。

美国有五组十分引人注目的数字图书馆规划, 有的已经完成, 有的正在开始进行, 相关情况如下。

(1) 美国“数字图书馆首创计划”第一阶段

1994 年 9 月, 美国国家科学基金会 (NSF) 正式公布了一项为期四年、投入 2440 万美元的“数字图书馆首创计划” (Digital Library Initiative)。该计划由美国国家科学基金会、国防部先进技术局 (ARPA) 和国家宇航局 (NASA) 联合出资, 由科学基金会机器人学与智能系统信息分部负责协调。计划中包括六个研究项目, 分别由六所大学牵头, 开发数字图书馆所需的各种新技术。该项计划的目标是: “使收集、存储和组织数字化信息的技术手段得到极大的进步, 使数字化信息能通过网络查询、检索和处理, 并以用户友好方式实现。”该计划的实现途径是: “对所有项目的一个共同要求是要有研究伙伴。我们把建立研究者、应用开发者和用户之间的合作关系视为在产生新知识、推广新思维、加速技术转化过程中取得成功的先决条件。”每个项目都将先开发一个测试基地, 用来进行研究和建立数字图书馆的原型, 然后将原型放大以存储更多的信息, 应用更先进的信息处理工具, 容纳更多的用户。

该计划中的六个研究项目包括: 信息媒体、环境科学电子图书馆、密歇根大学数字图书馆研究、亚历山大工程、斯坦福集成数字图书馆项目和构造互联空间等。

这六个项目, 已于 1998 年 8 月前完成, 从中得到了许多宝贵的经验, 特别是研制