国家『十一五』重点图书出版规划项目

数字图书馆理论、方法与技术

张晓林／著

当代中国图书馆学研究文库
DANGDAI ZHONGGUO TUSHUGUANXUE YANJIU WENKU
（第一辑）

Studies on Digital Library
Technologies and Development

北京图书馆出版社

国家“十一五”重点图书出版规划项目
当代中国图书馆学研究文库(第一辑)

数字图书馆理论、方法与技术

Studies on Digital Library Technologies and Development

张晓林 著

北京图书馆出版社

图书在版编目(CIP)数据

数字图书馆理论、方法与技术/张晓林著. —北京:北京图书馆出版社,2007.7

(当代中国图书馆学研究文库. 第一辑)

ISBN 978-7-5013-3443-8

Ⅰ.图… Ⅱ.张… Ⅲ.数字图书馆—文集 Ⅳ.G250.76-53

中国版本图书馆 CIP 数据核字(2007)第 030792 号

书名 数字图书馆理论、方法与技术

著者 张晓林 著

出版 北京图书馆出版社 (100034 北京西城区文津街 7 号)

发行 010-66139745 66151313 66175620 66126153
66174391(传真) 66126156(门市部)

E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com

经销 新华书店

印刷 北京华正印刷有限公司

开本 787×960 毫米 1/16

印张 19.875

版次 2007 年 7 月第 1 版 2007 年 7 月第 1 次印刷

字数 270(千字)

书号 ISBN 978-7-5013-3443-8/G·692

定价 38.00 元

前　言

当把这本文集呈现在读者面前时，我感到一丝欣慰，但更多地感到不安。

所谓欣慰，是这些年作为数字图书馆技术领域的学习者、实践者，参与了许多数字图书馆技术的研究，尤其是有机会参加了国家科学数字图书馆的设计、组织和建设，也参与了其他重要数字图书馆系统的研究与论证，对数字图书馆技术与系统的一些问题有所研究、有所理解、有所应用，见证了数字图书馆技术促进和提升文献信息服务的能力，见证了用户通过数字图书馆系统迅速提高了检索和利用信息的能力。数字图书馆的根本意义，或者说，研究数字图书馆的根本意义，就在于它可以大幅度地提高用户的信息能力，而只有当我们的研究真正转化成了用户的信息能力，我们的研究才有价值（也许才能“感到欣慰”）。从这个意义上讲，我更愿意作为数字图书馆的一个实践者，根据用户需求及其变化，根据数字环境变化，来不断地探索数字图书馆技术和应用的发展，这大概也是这个文集中的研究文章内容覆盖面比较广的原因之一。

所谓不安，是现在作为文献信息服务的组织者、实践者，面临迅速发展变化的用户需求和信息环境，深感挑战的严峻和迫切：（1）数字信息资源已经成为科研、教育、医疗、政务等高信息密集度、高知识依赖度和高协同工作度领域的主流信息资源，而且这些资源的每一个比特都可以标识、解析、关联、分析和重组，这为信息的分析处理，为知识的发现与组织，提供了与传统图书馆环境截然不同的信息机制。（2）我们正面临数字科研、数字教育、数字出版、数字媒体、数字文化等的汇聚，当用户习惯于这样的汇

聚空间及其丰富的功能和服务时，当用户通过集成机制去个性化、集成化地发现、交流、组织、利用这些空间中的信息内容和服务时，当这种汇聚和相应的集成应用成为用户的主流信息模式时，以文献收集、组织、检索和传递为主的图书馆模式如何继续维系和提高自己的贡献就成为严重的问题。(3)数字环境下，尤其是多重数字空间聚合的环境下，用户的信息需要和信息行为已经不再是简单的检索、获取和阅读文献(数字或者印本)，而是根据问题，根据研究和解决问题的过程，把多源、多样化、多层次的信息本身作为工作内容、分析对象、研究机制和研究目标，例如 bio-informatics、chemo-informatics、geo-informatics、legal-informatics 等等，将信息组织、信息分析、信息产生、知识验证等整合起来，将信息利用和信息服务推向了一个更高、更深入、更贴近核心需求和更能作出核心贡献的阶段，传统的文献检索与传递服务在此面前很容易被边缘化。(4)以 Web 2.0 为代表的 Interactive、User-driven、Community-oriented、Dynamic-structured 的网络信息交流与服务环境，也对我们所熟悉的以正式机构为基础的资源中心、学习中心、服务中心模式提出了挑战。我们太习惯于一种要把用户吸引到自己系统或机构的“中心”化服务模式，但它是否适应未来的用户？它如何去适应未来的用户？(5)我们更忘不了在信息内容和信息服务领域不断创新、不断扩展的各类信息服务商，例如 Google、Yahoo、Microsoft、百度等。它们的用户覆盖面、信息内容覆盖面、信息服务覆盖或汇接面、市场吸引力、服务创新力等，逐步使得它们成为用户(尤其是新一代用户)的主流信息服务平台，而且它们不遗余力地推动各种数字化计划，不断连接各种图书馆和数字图书馆服务，我们如何应对这种不可避免的趋势？当然，这些并不完全是技术问题，但确实又是研究数字图书馆技术与系统必须面临的问题。数字图书馆技术及其发展，从根本上讲，不是为图书馆服务的，而是为用户服务的；不是解决现有图书馆的当前问题，而是解决用户的不断变化的需求。从这个意义上说，数字图书馆的发展，功夫可能在图书馆之外。其实，对于整个图书馆来说，又何尝不是如此。我的不安，在于我个人、我们的领域，也许对新的挑战还远没有做好准备。因此，这个前言，似乎更

应该是一个新的学习计划，为了未来。

最后要指出，我的许多研究工作是与我的同事、学生合作进行的，这里收集的文章中相当多的是合作撰写的。和他们的合作给了我灵感和力量，在此也特别向他们表示感谢。

《当代中国图书馆学研究文库》编委会

一尊还酹江月

——《当代中国图书馆学研究文库》(第一辑)总序

30年前结束“文革”后的中国,和其他所有事业一样,图书馆学也获得了历史性的转折。经过恢复、改革、批判、建设,摆脱“意识形态化”的影响,挣破经验主义的束缚,走向科学化的发展道路,中国图书馆学进入了一个欣欣向荣的新时期。

在这一伟大进程中,一大批中青年学者成为建设新时期图书馆学的主力军。他们较少旧传统的束缚,勇于提出新的见解,发表了许多名篇佳作,推动我国图书馆学研究不断向前发展。其中涌现了不少杰出人才,成为新时期的弄潮儿。特别是在世纪之交的时候,现代信息技术的发展给图书馆学研究带来了许多新的问题,“数字图书馆”等新的研究内容层出不穷,中青年学者在这方面充分发挥了知识结构的优势,他们的研究成果具有鲜明的时代特征。

展现在我们面前的这一束文集,虽只是其丰硕成果中的一小部分,但已可看出他们的学术成就与青春活力。从毅然提出“转变图书馆学研究的方向”(张晓林),到以生命的代价“追问图书馆的本质”(黄纯元);积极“探索新图书馆学发展轨迹”(范并思),深入进行“中国图书馆学本土化的思考”(刘兹恒);热烈“呼唤新世纪”的到来(吴建中),实现“从传统向现代化的转型”(富平);潜心钻研“数字技术应用”(朱强),时刻关注“图书馆精神”(程焕文);或驰骋于“期刊领域研究”(叶继元),或在“知识产权研究”沃土上耕耘(陈传夫)。阅读这些文稿,不由产生丰收的喜悦。

他们是幸运的,成长于新时代,受到良好的高等教育与科学方法训练,有不同程度的跨学科与国外留学经历。这是他们的客

观条件。

他们也是努力的，具有为事业献身的精神，崇尚理论与实践的紧密结合，辛勤劳动，刻苦钻研，善于与他人合作。这是他们的主观因素。

长江后浪推前浪，一代新人换旧人。

进入21世纪以来，我国经济持续平稳增长，科学、教育、文化事业都有了长足的发展。图书馆事业建设也有了更好的条件，无论事业的规模，或是服务的内涵，都将有极大的变化。图书馆学研究也要面对新的课题，提出新的研究成果。

中国图书馆学正在走向世界，需要有一批学术上的领军人物，相信这些中青年学者有能力担负起这一历史重任。

这束文稿虽然不能反映新时期图书馆学成果的全貌，也不是每篇均为精品，但绝大多数都很有价值，是图书馆学发展新时期的历史记录。相信还有更多更好的佳作未曾齐集，需要我们继续发掘，以扩大收获。

东坡词云："多情应笑我，早生华发。人生如梦，一尊还酹江月。"

吴慰慈　陈源蒸

2006年2月28日

目 录

数字图书馆机制的范式演变及其挑战

1 前言

迅速发展的信息网络和数字信息资源体系正在造就一个全新的信息服务环境，其中信息资源、信息组织工具、信息传递工具日益聚合在同一数字空间，信息资源系统、信息服务系统和用户信息系统（例如电子邮件信箱、个人网页、课题网站、机构信息系统、业务信息管理系统等）日益连接在同一网络空间，各种基于网络、基于知识、基于协作的信息处理机制也日益成熟，它们之间的链接、交换、互操作、协作和集成也日益成为可能。

数字图书馆（以及所有信息服务系统）的根本目标是通过一系列服务机制有效支持用户利用信息来学习和创造知识。当信息资源、信息服务和用户（信息活动）都聚合在同一数字空间时，就有可能从新的技术基础出发，从用户信息利用全过程及其复杂信息活动的角度来重新审视信息服务系统的功能与结构，构建全面和直接支持用户信息活动的信息服务机制。因此，随着数字信息资源、信息服务系统和用户信息环境的不断发展，数字图书馆机制也从基于数字信息资源的系统形态逐步过渡到基于集成信息服务的系统形态，并开始向基于用户信息活动环境的系统形态过渡。本文将对这一范式演变过程以及它为数字图书馆建设和图书情报理论与实践带来的挑战进行初步分析。

2 数字图书馆的范式演变

我们可以根据数字图书馆建设的基点、体系形式和所解决的

关键任务等，将数字图书馆分为不断递进和深化的三代范式[1-3]：

2.1 第一代——基于数字化资源的数字图书馆（Resource - based digital library）

第一代数字图书馆主要在特定文献资源数字化的基础上建立数字信息资源系统，它们往往作为独立系统，嵌入到传统图书馆系统或上层机构信息系统中，将跨时空检索和传递特定数字化资源作为其主要任务，称为基于数字化资源的数字图书馆。这类数字图书馆的范例包括美国 LC 的 American Memory 系统（http://memory. loc. gov/）、密歇根大学的 JSTOR（http://www. jstor. org/）、拉斯阿拉莫斯国家实验室的 Physics E - Print（http://www. arXiv. org/）、美国 NASA 的 Astrophysics Data System（http://ads. harvard. edu/）以及我国实验数字图书馆项目等。目前，许多图书馆、档案馆乃至博物馆等的数字化馆藏系统基本属于这一范围。我们可用图 1 来表示这一类数字图书馆的基本逻辑结构，这时资源库管理与检索系统往往是与数字资源库直接捆绑（Hard-binding），其功能形式、技术方法和操作管理机制往往取决于资源库的内容格式、元数据格式、知识组织体系和特定软硬件平台，可能利用专用甚至私密的代理模块来实现，最后通过 Web 平台供用户直接检索浏览。

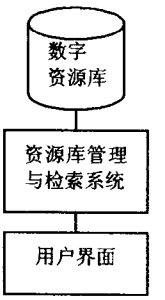

图 1

这类数字图书馆的任务范畴一般包括：数字化对象的选择（例如珍贵文献、手稿、档案、地方特色文献或经过授权的出版物等），文献数字化方法（Digitization），数字文献格式标准体系（例如数字文本标记格式、数字图像扫描格式等的标准化），描述和管理具体数字文献的元数据，数字资源库组织（包括标识符机制、内部标识与检索机制、存储管理系统等），检索与呈现方法（包括并行检索、基于内容检索和简单数字对象的呈现等），初步的知识产权管理（包括用户使用控制和数字水印保护等），数字化工作流程等。

2.2 第二代——基于集成信息服务的数字图书馆（Service-

based digital library)

为有效利用数字信息环境中分布、多样化、往往是异构的数字信息资源,第二代数字图书馆致力于支持分布的数字信息系统间的互操作(interoperability),支持这些系统间无缝交换和共享信息资源与服务,并由此构造一个逻辑的集成信息服务机制,形成基于集成信息服务的数字图书馆。这一代数字图书馆不再以文献数字化和具体数字资源库建设为核心,而主要是面向分布和多样化数字信息资源(包括由出版商、学术机构、各类机构等拥有的"正式"和"非正式"资源),通过服务集成(包括虚拟资源体系建设、跨系统多系统检索、分布式使用管理、分布式权益管理、分布式数字参考咨询服务、长期保护协调等)构造统一的信息服务系统,将形成与传统图书馆不同的新系统形态和组织形态,是目前数字图书馆技术研究、应用试验和开发的主要趋势,其范例包括加州大学的 California Digital Library(http://www.cdlib.edu/)、计算机科学领域的 Networked Computer Science Technical Reference Library(http://www.ncstrl.org/)、学位论文领域的 Networked Digital Library of These and Dissertations(http://www.ndltd.org/)、美国 OhioLink(http://www.ohiolink.edu/)、英国 National Electronic Site License Initiative(http://www.nesli.ac.uk/)、英国 Distributed National Electronic Resources(DNER)[4] 以及我国的 CALIS 等。

图2是 MOA2 项目提出的第二代数字图书馆体系架构[5],将第一代数字图书馆作为具体数字资源系统,强调通过一系列搜寻、转换、整合工具来集成这些分布的系统,支持集成服务。

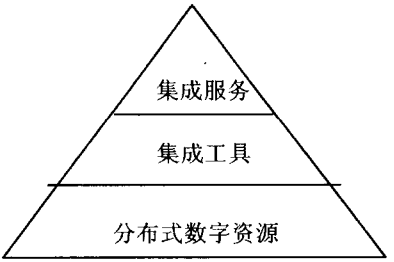

图2

这一代数字图书馆的任务范畴将更集中于:分布式系统结构,系统互操作(例如基于分布式对象、基于代理与协调技术、基于搜寻或整合协议的互操作),数字对象与数字对象唯一标识符,元数据互操作,数字资源开放链接,分布式开放式使用管理,分布式开放式数字权益管理,网络化资源建设和资源组织,以及基于内容和集成的检索技术(基于内容检索、基于知识体系和语义的检索、跨系统检索、跨语言检索等)。

2.3 基于用户信息活动的数字图书馆(Work-based digital library)

随着聚合数字信息空间的逐步形成,人们第一次有可能摆脱传统图书情报系统(甚至传统数字图书馆系统)单纯基于信息资源的服务形态和将信息系统与用户信息利用过程相对隔绝的局限,以支持用户灵活地处理信息、提炼知识和交流协作为核心,围绕用户信息活动和用户信息系统来组织、集成、嵌入数字信息资源和信息服务,从而更直接、深入、有效地支持用户检索、处理、利用信息以解决问题的全过程。在这种理念推动下,数字图书馆的前沿研究者们已开始探讨以用户信息活动为基础的数字图书馆机制。

例如,美国国家科学基金会的"国家科学、数学、工程和技术教育数字图书馆项目"(NSDL)[6-7]明确提出建设成围绕着用户协作化学习过程的分布式资源网络和学习机制,个人或集体可充分和动态调用各种数字化资源和工具(包括合作学习系统、远程实验室、虚拟实验室等)来个性化和协作地检索、集成、处理信息并以此支持合作学习。将于今年九月举行的欧洲数字图书馆会议(ECDL)已将研究开发嵌入到用户信息空间和用户合作过程的数字图书馆系统作为其三大主题之一[8],马里兰大学 MiND 项目提出新一代数字图书馆应直接支持用户在其信息利用过程中灵活处理信息对象[9],奥地利 Maurer 提出数字图书馆应成为用户交流媒介来支持用户对数字信息的注解、交流和协作处理[10],美国数字图书馆研究著名专家 Ed Fox 也提到数字图书馆的下一步发展将走向虚拟个性化数字图书馆和嵌入到用户工作环境中的数字图书馆[11]。实际上,一些试验或者应用系统已出现,例如

NCSU 图书馆的 MyLibrary 系统(http://my.lib.ncsu.edu/)支持建立个性化门户,Questia 数字图书馆(http://www.questia.com/)在提供全文图书检索浏览的同时支持用户对图书内容的析取、注解和交流,SOSIG 主题信息网关(http://www.sosig.ac.uk/)在提供网络精选资源导航的同时支持用户发布信息、围绕特定主题资源构造用户社区和进行协作。

图 3 给出了第三代数字图书馆的可能模式。在分布式数字资源系统(分布资源层)和集成信息服务体系(集成服务层)基础上,通过一定的个性化定制机制形成适应用户或用户群组需要的可能是动态过滤、析取和组合的资源、工具和服务集合(个性化定制层),这些集合被有机地嵌入到用户信息系统或用户信息利用环境(用户系统层)中,直接支持用户的信息利用活动。

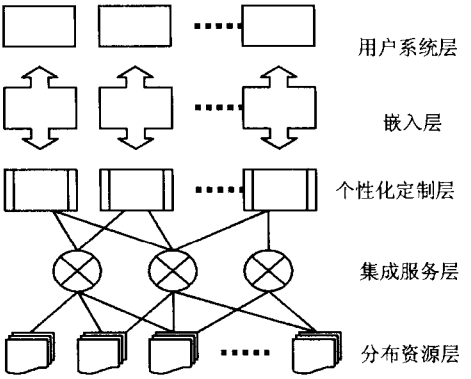

图 3

这一代数字图书馆的任务范畴可能涉及:数字对象与分布式对象代理技术,智能代理技术,个性化机制,动态文献和动态文献集技术,知识组织系统技术,信息协作处理机制,基于 XML 技术体系的信息处理技术和信息系统定义、构建、集成技术,用户信息系统和信息处理流程中数字化信息资源与服务的嵌入与定制机制等。

3 基于集成信息服务的数字图书馆模式分析

面对异构、多样的分布式资源系统和分布、移动的用户,集成

信息服务体系将：

a. 支持分布和多样化资源系统的方便接入，同时支持各资源系统的自主性和本地服务；

b. 支持对这些资源系统的逻辑集成，支持以标准形式跨资源系统进行搜寻、检索、转换和整合；

c. 支持基于整个分布资源体系的集成服务和服务管理机制，支持对分布式第三方工具或服务系统进行动态和无缝调用，形成逻辑整体服务系统；

d. 支持对资源体系进行逻辑重组，以构建适应不同用户群的虚拟信息资源和服务系统；

e. 支持整个机制的开放性、可伸缩性和可扩展性，能方便接入和动态组合任意数量或类型的资源与服务系统（包括新的资源形式、系统形式和服务工具）。

集成信息服务体系的核心问题是分布、异构系统的互操作，可能的实现形式包括：

a. 基于中间协调与转换代理的联邦式系统（Federated Systems），例如 NCSTRL +[12]、基于 Z39.50/ILL 协议的集成检索系统、英国 MIA 结构[13]及 DNER[4]系统，可通过全面和复杂的代理机制支持强有力的集成、转换服务；

b. 基于标准搜寻协议的开放资源体系（Open Harvesting Systems），如 OAI 协议机制[14]，资源系统（作为数据提供者）可通过简单开放机制，提供基础性元数据和读取功能，从而支持服务集成者搜寻和提取元数据，建立元数据库，读取数字对象及提供其他第三方服务；

c. 基于整合检索协议的跨系统搜集整合机制（Gathering/Integration Systems），在诸如 LDAP、WHOIS + +、SDLIP 等协议支持下，这些机制可直接利用各数字资源系统的可公开获取信息进行整合检索，例如 CrossRoads[15]、Isaac Network[16]、Imesh[17]、LFDL[18]等。

图 4 给出了一个集成信息服务体系结构，通过各种协议和机制来支持对分布式数字资源的检索和获取，通过一系列转换、整合、服务调用、工具调用和管理机制来提供逻辑集成和服务管理，