异构分布式环境下的数字图书馆互操作技术

张付志 著

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

异构分布式环境下的数字图书馆互操作技术

异构分布式环境下的数字图书馆互操作技术

张付志 著

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

内容简介

本书系统地论述了异构分布式环境下的数字图书馆互操作技术。全书分为6章：第1章介绍了数字图书馆的概念、分类、特点和发展趋势以及数字图书馆互操作的概念、研究数字图书馆互操作的目的、意义和面临的困难；第2章对数字图书馆互操作的研究现状及关键技术进行了综述与讨论；第3章介绍了基于元级搜索服务的数字图书馆互操作解决方案；第4章介绍了数字图书馆包装层的生成技术；第5章介绍了数字图书馆元搜索引擎的实现及评价；第6章介绍了基于移动Agent的数字图书馆互操作方案。本书内容丰富，语言精炼，论述深入浅出，通俗易懂。在撰写过程中力求理论与实践相结合，突出实用性。

本书适合计算机、图书情报、网络文化教育等领域的研究人员、工程技术人员、研究生和本科生阅读，也可作为数字图书馆技术研发人员的技术参考书。

图书在版编目（CIP）数据

异构分布式环境下的数字图书馆互操作技术 / 张付志著.—北京：电子工业出版社，2007.12

ISBN 978-7-121-05560-7

Ⅰ. 异… Ⅱ. 张… Ⅲ. 数字图书馆－研究 Ⅳ. G250.76

中国版本图书馆 CIP 数据核字（2007）第 191384 号

责任编辑：董亚峰
印　　刷：北京天宇星印刷厂
装　　订：涿州市桃园装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：850×1168　1/32　　印张：8.25　字数：304 千字
印　　次：2007 年 12 月第 1 次印刷
定　　价：25.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

数字图书馆为分布式数字信息资源的管理提供了一种有效手段，它将从根本上改变目前 Internet 上信息资源分散、不便使用的现状。虽然 Web 上已有许多数字图书馆，但都属于自治的信息系统，它们具有各自的查询界面、体系结构、通信协议和管理策略。由于数字图书馆自身所具有的异构性和分布性以及数字图书馆之间所缺乏的互操作性，使得 Web 上各数字图书馆之间难以共享信息资源和服务，成为困扰用户充分使用数字图书馆中信息资源的一大障碍。用户为了查找所需要的资料，往往需要访问几个数字图书馆，同一查询不得不重复提交给每个数字图书馆。由于 Web 上大多数数字图书馆是基于数据库驱动的，现有流行的 Internet 搜索引擎不能对其内容建立索引。因此，对于那些需要跨越多个数字图书馆查找资料的用户来说，现有的 Internet 搜索引擎不能满足这种需求。如果能够为 Web 上的数字图书馆提供一种统一的访问界面，并将用户提交的查询实时地映射到不同数字图书馆的查询服务，这样用户看到的将不再是一些异构、分布的数字图书馆，而是一个能够共享资源和服务的联合数字信息资源库。要实现这一目的，必须解决数字图书馆的异构性和分布性问题，实现互操作。

本书系统论述了异构分布式环境下的数字图书馆互操作技术。全书分为 6 章，各章内容分述如下。

第 1 章　绪论。本章对数字图书馆的概念、分类、特点和发展趋势进行了介绍，给出了数字图书馆互操作的概念，介绍了研究数字图书馆互操作的目的和意义，分析了实现数字图书馆的互操作所面临的困难。

第 2 章　数字图书馆互操作的研究现状及关键技术。本章对数字图书馆互操作的研究现状及关键技术进行了综述与讨论，内容主要包括：数字图书馆互操作的层次结构、数字图书馆互操作协议、数字图书馆中间件、数字图书馆中的信息发现、语义互操

作、元数据（metadata）互操作、数字图书馆互操作的实现方法及评价标准。

第 3 章　基于元级搜索服务的数字图书馆互操作方案。本章介绍了 Internet 搜索引擎的概念及分类、元搜索引擎的概念、结构模型及工作原理，分析了数字图书馆元搜索引擎的设计要求，设计了一种基于元搜索引擎的数字图书馆互操作框架，给出了数字图书馆元搜索引擎的工作流程，并对基于 Web 服务的数字图书馆互操作方案进行了探讨。

第 4 章　数字图书馆包装层生成技术。本章介绍了包装层的概念、作用及建立方法，并从数字图书馆包装层 Agent 的结构模型、查询映射、数字图书馆查询服务的调用以及结果提取等方面讨论了建立数字图书馆包装层的关键技术，介绍了基于 XML 和 Java 的数字图书馆包装层半自动生成方法以及实现数字图书馆包装层生成器的关键技术。

第 5 章　数字图书馆元搜索引擎的实现与评价。本章详细介绍了实现数字图书馆元搜索引擎的关键技术，内容包括：数字图书馆元搜索引擎的体系结构，用户接口 Agent、中介层 Agent 和包装层 Agent 的实现技术，中介层与包装层之间通信的实现技术。并对数字图书馆元搜索引擎的性能进行了模拟测试。

第 6 章　基于移动 Agent 的数字图书馆互操作方案。本章针对网络方面的因素（如网络低带宽、高延迟、负载重和不可靠网络连接等）对数字图书馆互操作服务质量的影响，探讨了基于移动 Agent 的数字图书馆互操作解决方案。指出了基于移动 Agent 的数字图书馆互操作的实现途径，设计了一种基于移动 Agent 的数字图书馆互操作框架，并介绍了原型系统的实现技术。

由于作者水平有限，书中难免有不足之处，敬请读者批评指正。

作者

2007 年 10 月

目 录

第1章 绪 论

数字图书馆是为了有效利用 Internet 的信息资源而产生的，其研究与开发得到了世界各国的重视，已成为国际高科技竞争中新的制高点和评价一个国家信息基础设施水平的重要标志之一。本章介绍了数字图书馆产生的背景及国内外研究概况，给出了数字图书馆的有关定义、特点、功能及分类，指出了数字图书馆的发展趋势，介绍了数字图书馆互操作的概念，分析了研究数字图书馆互操作的目的、意义及面临的困难。

1.1 引 言

20 世纪 90 年代以来，随着以计算机技术、通信技术、网络技术、高密度存储技术和多媒体技术的高速发展和有机结合，特别是 Internet 在全世界的迅速普及与应用，引发了世界范围内信息环境的改变。作为信息拥有者和提供者的图书馆也在经历着这场信息革命浪潮的冲击，出现了“电子图书馆（Electronic Libraries）”、“数字图书馆（Digital Libraries，DLs）”和“虚拟图书馆（Virtual Libraries）”等概念。这些概念也随着数字技术和网络技术的发展，最终归结为数字图书馆的建立和发展。

数字图书馆主要是为了解决 Internet 上信息资源的有效利用而产生的，它将从根本上改变目前 Internet 上信息资源分散、不便使用的现状。数字图书馆要解决的是目前 Internet 上存在的主要问题，即用户查找信息困难、异构信息仓储（repository）之间的互操作（interoperability）和缺乏对大规模分布式数据的操作机制。数字图书馆的关键技术是研究数字化信息的有效组织结构，解决

各信息仓储之间的互操作问题，形成数字图书馆的基础体系结构，以便有效地操作大规模的、分布的数字化信息，实现跨越异构信息仓储的统一检索服务，为用户提供一个虚拟的、统一的信息网络[1]。

数字图书馆是世界各国研究与开发的热点领域。在数字图书馆的研发方面，西方发达国家开展得最早，投入的资金最多，取得的成果也很显著。美国将数字图书馆的建设作为国家信息基础设施的主体结构，数字图书馆的研究和建设被认为是“美国面临的挑战”。1994 年 9 月，美国国家自然科学基金会（NSF）、美国国防部高级研究计划署（DARPA）和美国国家宇航局（NASA）联合出资 2400 万美元，资助 6 个为期 4 年（1994.9～1998.8）的第一期数字图书馆先导研究项目（DLI-1）。DLI-1 的目标是提升用来收集、存储、组织和使用那些包含各类广泛分布的知识资源以及各种以电子形式存储的内容[2]。该计划的 6 个项目分别由美国的斯坦福大学等 6 所著名大学承担，每个项目都要开发一个数字图书馆的测试平台，实现相关研究。1999 年 9 月，美国联邦政府又投入大约 4400 万美元，资助了 24 个为期 5 年（1999.9～2004.8）的第二期数字图书馆先导研究项目（DLI-2）。DLI-2 计划是在 DLI-1 计划所取得的研究成果的基础上，开展以人为中心的研究、基于内容及内容收集的研究和以系统为中心的研究。该计划是一个多机构首创计划，力求在下一代数字图书馆发展的基本研究中起到领导作用，促进全球分布的、网络化信息资源的利用，鼓励现有和新加入的项目将重点放在创新性的应用领域[2]。

美国在数字图书馆方面的研究和取得的成就对世界各国产生了深远的影响。1995 年 2 月，由欧盟主持的西方七国“信息高速公路”部长级会议上，制定了一个发展全球信息高速公路的 11 项计划，其中一项就是建立全球数字图书馆计划。英国以大英图书馆为龙头，联合各大学图书馆和信息机构，致力于名为“电子图书馆”的研究，并为此根据信息技术的发展，提出了分阶段的发展计划和具体研究项目。日本也在数字图书馆的研究方面投入巨

资，建立了相应的原型试验系统，耗资 4 亿美元的日本“关西图书馆工程”项目计划建成日本最大的数字图书馆和亚洲地区的文献中心。

在一些发达国家大力加强数字图书馆的研究和建设的影响下，我国于 1997 年 7 月正式开始了自己的数字图书馆研究计划——“中国试验型数字式图书馆项目”。国内一些大型图书馆机构和高校也积极开展了数字图书馆技术的相关研究，并且有计划、有步骤地加紧本系统内部信息的数字化建设。目前正在实施的“中国数字图书馆工程”项目把数字图书馆作为知识经济的重要载体来建设，其总体目标是实现中国数字图书馆工程的总体架构，建设超大规模的优质中文信息资源库群，并通过国家高速宽带网向全国及全球提供服务，最终形成世界上最大、最全面的网上中文信息基地和服务中心。

数字图书馆代表着 21 世纪信息资源共享的方向，既是知识的网络，又是知识的中心。数字图书馆作为知识网络中组织知识内容和利用知识内容的核心模式，其开发与建设对传播知识具有十分重要的作用。特别是对于教育领域，数字图书馆将成为非常重要的教育设施。与传统图书馆不同的是，数字图书馆是一种新的基础设施和知识环境，数字图书馆中的资源不再是孤立地分散在世界各地的图书馆中，而是通过集成和利用最新的计算技术、通信技术及数字内容，建成超大规模的、可扩展的、可互操作的知识库群，成为人类共享的知识财富。数字图书馆可以向用户提供更方便、更快捷、更先进的服务，改变以往人们获取信息、组织信息和利用信息的方式。

目前，数字图书馆已成为国际高科技竞争中新的制高点和评价一个国家信息基础设施水平的重要标志之一。数字图书馆的建设、普及与应用，必将大大影响教育的质量和生活的质量。

1.2 数字图书馆的定义及特点

1.2.1 数字图书馆的定义

目前，数字图书馆的研究正日益广泛和深入，数字图书馆的建设也逐步从实验阶段向实用阶段转变。但是，对于什么是数字图书馆，至今还没有一个科学、严谨、普遍认可的定义。不过对数字图书馆这一概念已基本上形成了一致的理解，即数字图书馆的内涵要比传统的图书馆广泛得多。下面给出几种有关数字图书馆的定义。

定义 1.1 数字图书馆可非正式地定义为有组织的信息馆藏及相关服务，信息以数字化形式保存，并通过网络进行访问[3]。

该定义强调了数字图书馆中的信息是有组织的，并且通过网络加以利用，同时还应提供选择信息、组织信息、存储信息和发布信息的相关服务程序。

定义 1.2 数字图书馆是一个大系统，它具有分布式的、大规模的和有组织的数据库和知识库。用户或用户团体可对系统内的数据库和知识库进行一致性的访问，获得自己所需要的最终信息[4]。

该定义强调数字图书馆已不再是传统意义上的图书馆，而是图书馆自动化发展的更高级阶段。

定义 1.3 数字图书馆是采用现代高新技术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、不便于使用的现状。通俗地说，数字图书馆是没有时空限制的、便于使用的、超大规模的知识中心[5]。

这是中国国家图书馆关于数字图书馆的定义，其核心在于：数字图书馆是一个数字信息资源系统，是通过 Internet 向用户提供信息服务的超大规模知识中心。它体现了数字图书馆既是知识的网络，又是知识的中心这一内涵。

定义 1.4 数字图书馆是一个全球性的、分布式的大型知识

库，即以分布式海量数据库群为支撑，基于智能技术的大型、开放、分布式信息库[6]。

该定义的实质是把数字图书馆等同于可共享的大规模分布式知识系统，是通过互联网提供智能信息检索服务的知识中心。

上述 4 种定义虽然算不上是对数字图书馆的规范、权威的定义，但是从这几种定义可以看出，数字图书馆不是传统图书馆的简单数字化，而是包含着复杂的技术和更深的内涵。数字图书馆的定义将会随着数字图书馆研究和建设的深入而不断地得到完善。

1.2.2　数字图书馆的特点

（1）信息资源的数字化

信息资源的数字化是构成数字图书馆的馆藏（collections）资源的物质基础，因为数字图书馆的其他特点都建立在信息资源数字化的基础上，这也是数字图书馆与传统图书馆的区别所在。如果没有数字化的馆藏资源，数字图书馆就成了无源之水，无本之木。

（2）信息访问的网络化

高速的数字通信网络是数字图书馆存在的基础，通过互联网数字图书馆可以向用户提供更方便、更快捷、更先进和跨时空的信息服务。

（3）信息资源的分布式管理

数字图书馆中的信息资源库可以分布在不同的地理位置，将这些分布式的资源库进行无缝集成，通过 Internet 实现跨库检索服务，达到高度的资源共享。

（4）高新技术的集成化

数字图书馆是一个十分复杂和庞大的分布式信息系统，涉及的技术领域很多，它是数据仓库技术、互联网技术、网络存储管理技术、信息检索技术、多媒体技术和数据挖掘技术等多种高新技术的综合应用。

1.3 数字图书馆的功能与分类

1.3.1 数字图书馆的功能

目前，数字图书馆主要提供以下功能。

（1）数字化处理功能

数字馆藏是数字图书馆的基础，数字图书馆的一切活动都是围绕着它进行的。因此，存储在物理介质上的声、文、图、像资料要成为数字图书馆的馆藏资源，必须进行数字化处理。

（2）信息的存储与管理

目前的数字图书馆在信息的存储与管理上大多数采用三层B/S模式，客户机、Web服务器和数据库服务器构成了信息传递的核心结构。其中，Web服务器负责接受来自客户端Web浏览器的查询请求，进行信息处理和结果发送，管理HTML构成的信息空间，并提供对数据库服务器的访问接口。数据库服务器负责管理数字图书馆的馆藏资源（包括元数据库和数字对象数据库），响应Web服务器的服务请求，进行信息处理，并将处理结果传送给Web服务器。最后，由Web服务器将查询结果交给用户，实现了数字图书馆中对象数据的传送。

（3）信息的有效检索与查询

数字图书馆不仅应当提供基于文本内容的检索与查询服务，而且还应当提供基于多媒体内容的检索与查询服务。因此，建立良好的馆藏元数据结构，是实现有效检索与查询的关键。

（4）信息的发布与传播

数字图书馆中数字信息的发布是基于面向对象机制的，信息资源的拥有者可以选择多种方式来发布信息。对于多媒体数字信息的发布，还需要依靠高速宽带网络技术的支持。

（5）信息安全与权限管理

数字图书馆运行于开放环境（如Internet）之中，对安全性提

出了特殊的要求。它不仅要提供一般计算机网络系统的管理功能，而且要对各类用户的访问权限进行管理，以保护信息拥有者的利益，并采用适当的技术（如数字水印技术），以确保版权人的数字信息资源不被滥用。

1.3.2　数字图书馆的分类

目前，关于数字图书馆还没有一种普遍接受的分类方法。通常将其分为以下三种类型[7]：独立数字图书馆（Stand-alone Digital Library，SDL）、联邦数字图书馆（Federated Digital Library，FDL）和虚拟数字图书馆（Virtual Digital Library，VDL）。

（1）独立数字图书馆（SDL）

独立数字图书馆相当于传统图书馆的完全计算机化，其馆藏资源是经过数字化处理后的数字文档，这些数字文档集中保存在本地的仓储（repository）中，通过网络进行访问。目前许多数字图书馆都属于这种类型，如国内自主开发的超星数字图书馆。

（2）联邦数字图书馆（FDL）

联邦数字图书馆是指正式或非正式合作的操纵 SDLs 的一些组织，它们统一支持一组共同的服务和标准（如元数据的标准），以便在联盟成员之间共享 SDLs 的资源和服务，实现互操作。FDL 实际上是由多个自治的 SDLs 构成的网络图书馆，它提供一个能够透明访问各联盟成员 SDLs 的统一用户界面。由于不同的 SDLs 之间往往存在着异构性（因为不同的仓储通常采用不同的元数据格式和标准），所以互操作问题是建立 FDL 面临的主要挑战。网上计算机科学技术报告图书馆 NSCTRL[8] 和网上学位论文数字图书馆 NDLTD[9]就属于联邦数字图书馆。

（3）虚拟数字图书馆（VDL）

虚拟数字图书馆的馆藏中仅包含有元数据，具体的数字文档