

盛小平

国内外数字图书馆发展的比较研究

摘要 选用了14个国内外数字图书馆建设项目,从馆藏资源建设、检索特征、输出格式三方面进行了比较研究。研究表明,今后我国数字图书馆建设应在结构模式、数字资源建设和信息检索技术研究等方面下大功夫,以取得突破性进展。表3。参考文献11。

关键词 数字图书馆 比较研究 数字资源 信息检索

分类号 G250.76

ABSTRACT In this paper, the author compares fourteen digital library projects in China and in foreign countries in the aspects of library resource development, search characteristics and output format.

Then, the author proposes some suggestions for future development in China. 3 tabs. 11 refs.

KEY WORDS Digital library. Comparative study. Digital resources Information retrieval.

CLASS NUMBER G250.76

数字图书馆发展经过起源初创期(1993年前)和概念形成期(1994~1999),已经进入快速发展期(2000年以后)。加强国内外数字图书馆发展的比较研究,将有利于促进我国数字图书馆建设与发展。

1 14所数字图书馆情况简介

这里选用14个数字图书馆建设项目作为比较对象,它们的基本情况见表1^[1]。

表1 14个数字图书馆项目基本情况

项目名称	所属组织与国家	整体目标	始建年代	网址
计算机协会数字图书馆(ACM)	计算机协会(美国)	提供ACM期刊和会议论文全文的访问。	1996	www.acm.org/dl
美国记忆(AMMEM)	国会图书馆(美国)	主要提供反映美国历史和文化的数字化信息。	1996	lcweb2.loc.gov/ammem
卡内基—梅隆大学数字图书馆(CMDL)	卡内基—梅隆大学(美国)	开发一个集成声音、图像、语言识别以及电子书籍、艺术、音乐、电子期刊的数字视频图书馆。	1994	www.ul.cs.cmu.edu
密执安大学数字图书馆(UMDL)	密执安大学(美国)	通过分布式网络环境,提供环境研究和其它交叉学科领域包括生命、自然和社会科学的电子信息资源。	1994	www.lib.umich.edu/libhome/dig.html
伯明翰大学集成图书馆开发与电子资源(BUILDER)	伯明翰大学(英国)	开发混合图书馆的工作模型以便在教学与研究中通过万维网无缝集成存取大规模印刷型与电子型信息资源。	1998	Builder.bham.ac.uk
英国图书馆电子化贝奥伍夫项目(BL)	英国图书馆	利用图像和网络技术来提高对馆藏数字化文献的存取。	1993	www.bl.uk
昆士兰郡镇图像项目(DIGILIB)	昆士兰大学(澳大利亚)	建立一种大规模的有关昆士兰和布里斯班家庭、公共、矿业和农业建筑物的数字图像馆藏。	不详	www.architect.uq.edu.au/digilib/index.html
学术电子文本和图像服务(SETIS)	悉尼大学(澳大利亚)	促进存取机构内部的和远程文本与图像数据、技术程序和电子文本的创立与存贮。	1996	setis.library.usyd.edu.au

续表

项目名称	所属组织与国家	整体目标	始建年代	网址
加拿大国家图书馆	加拿大国家图书馆	NLC 的电子馆藏项目旨在通过万维网来利用加拿大 500 多个图书馆的联机书籍、期刊和目录。	1995	www.nlc-bnc.ca
网关电子媒体服务 (GEMS)	南洋技术大学(新加坡)	通过校园网为全体教职工和学生传输大量的各种媒介的信息资源。	1999	www.ntu.edu.sg/library/media/gems/gems.htm
新西兰数字图书馆 (NZDL)	怀卡托大学(新西兰)	开发基础技术来帮助其它图书馆建立和管理各自数字馆藏和为公众所利用。	1996	www.cs.waikato.ac.nz/~nzdl
中国国家图书馆 (NL-CN)	中国国家图书馆	研究与开发数字图书馆的体系结构、标准和规范、应用系统,领导和组织中国数字图书馆工程建设。	1996	www.nlc.gov.cn
上海数字图书馆 (SHDL)	上海图书馆	采取统一的界面、软件、管理,充分考虑满足当前需要、适应资源共享和可持续发展的目标,实现远程、快速、全面、有序、智能、特色六大服务。	1998	www.libnet.sh.cn
清华大学数字图书馆 (TDL)	清华大学	借助计算机技术完成馆藏资源数字化存贮和管理,通过网络技术向分布广泛的用户提供便利的服务,从总体上提升图书馆的各方面功能。	1995	www.lib.tsinghua.edu.cn

这 14 个数字图书馆建设项目代表了公共图书馆、高校图书馆、科研图书馆三大系统中数字图书馆建设的最新进展。下面从馆藏建设、检索特征、输出格式三方面对它们进行比较研究。

2 国内外数字图书馆发展的比较研究

2.1 馆藏建设

ACM 馆藏包括 39378 篇 ACM 期刊和会议论文的全文,以及自 1985 年以来 7000 多次 ACM 期刊论文的引文目次和近 35000 次 ACM 会议论文的引文目次。

AMMEM 拥有 100 万多条与美国历史和文化有关的馆藏记录,同时也包含那些记录美国历史的各种文件、电影、手稿、照片和语音记录。AMMEM 根据不同主题组、年、地点、原有格式、数字化格式、图书馆类目和用户格式来对馆藏进行归类。每种主题类目分成 13 个子目,馆藏按字母顺序排列。

BL 存有贝奥伍夫遗留在英国图书馆的古代英语诗、11 世纪盎格鲁——撒克逊人史诗的手稿、Cotton

Vitellius 十五世的肖像、十分珍贵的 18 世纪抄本、1815 年编辑的 19 世纪初手稿校勘副本、一个综合词汇集索引及其新版与副本。原稿图像被组织起来检索整个版本、特定行或特定页码。

BUILDER 藏有印刷型和电子信息资源,各种考试试卷和两种电子期刊——《法医语言学》(Forensic Linguistics)和《国内历史》(Midland History),正在开发混合图书馆的检索界面。文献根据系、标题、课程代码和试卷号来组织。

CMDL 是一个多媒体数字图书馆,能播放 1000 多小时的数字视频、音频、图像和文本信息,提供 300 多种电子日报、期刊和电子图书的访问。CMDL 馆藏分为艺术、图书、文集、期刊、多媒体、音乐和科研项目。每组下设有子组,子组下设有更细的小组,成树状结构。每组及子组都按字母顺序编排,在线图书是根据作者和标题来组织的。

DIGLIB 建立了昆士兰历史建筑物馆藏,它包括大量的家庭、公共、矿业与农业建筑物。它们中的许多以前没有用任何格式记录过,目前已经存贮了

1030多张图片。图像和照片根据市镇、类型、特征、结构、素材和上下文来组织。

GEMS能提供网络光盘数据库、中文光盘标题、联机检索服务、电子期刊、视听资源、OPACs和网站。馆藏包括310多种电子期刊、项目报告、学位论文、会议论文和职员与学生捐献给图书馆的出版物,并能提供其它信息资源,如院历、课程信息、注册细节、时间表、未清账单等。GEMS能实现对馆藏光盘、联机数据库OPAC、科研项目报告、数字化学位论文、会议和其它出版物的访问,可以浏览数据库和电子期刊标题字顺表,并根据72种标题把文献分成若干组。

加拿大国家图书馆(NLC)的电子馆藏是通过与加拿大在线图书与期刊出版机构的正式合作来进行的。电子馆藏名目上标有网址可供利用,现有1800万条书目记录、55万条作者记录和由加拿大500个图书馆包括国家图书馆提供的300万册数字化藏书。文献按照标题字母顺序排列,并用杜威十进分类法和全文本格式来组织。电子出版物的全文本格式包括美国国家信息交换标准代码(ASCII)、超文本标识语言(HTML)、文本、Word和WordPerfect。

NZDL提供13种馆藏的存取,主要涉及计算机科学,但同时也包括人机交互书目、常见问题解答(FAQs)等。最大的馆藏资料是计算机科学技术报告,它包括25000份来自世界300个地方的研究报告。馆藏中FAQs资料也很多,并提供了*Computists Communique*杂志的全文本索引。

SETIS能提供大量网络型和内部人文学全文数据库的存取。除文学、哲学和宗教文本外,SETIS致力于一些文本与图像的创建工程。大规模馆藏如美国诗全文数据库、1840年以来澳大利亚文学数据库、英语诗数据库、英语戏剧数据库、牛津英语词典和分布式数字化研究生学位论文数据都能全文浏览,并按关键词、著作标题、作者姓名、出版日期、出版地、出版者、作者姓氏和作者日期与文献时代来编排。

UMDL馆藏建设集中于期刊文献与参考资源,如McGraw-Hill科学与技术百科全书、美国百科全书、英国百科全书和200种核心与知名期刊,并可提供1100种Elsevier期刊的访问,密执安大学数字化期刊与报纸共计超过3000种。UMDL资源是按照标题、类目和服务资源字母顺序3种方式来组织的,它分为9个标题:即艺术与人文、商业与经济学、工程、一般参考资料、政府信息和法律、健康科学、新闻、科学与社会科学。

中国国家图书馆(NLCN)目前正在抓紧馆藏文献书目数据的制作,已完成1949年以来的中文书目数据近100万条,完成1992年以来的西文书目数据近30万条;现正进行馆藏民国时期中文图书、古籍、舆图和金石等文献书目数据的制作。同时,还在进行一批如“中国年鉴信息”等专题数据库的制作;其次,还抓紧馆藏印刷品文献的数字化和馆藏缩微制品数字化,以及馆藏珍贵文献数字化^[2]。

上海数字图书馆(SHDL)正在进行9个资源库的建设,拥有数据近200GB。其中“上海图典”拥有2万余幅图片;“上海图文”收录了114种上海年鉴和115种新上海地方文献及地方文献书目;“点曲台”收录了15个剧种的5000余份(种)音频资料;“古籍善本”已完成3233种古籍善本的数字化;“民国时期图书”已完成1000多种代表著作的数字化和全文网络浏览;“科技会议录”收录了1986年至今共约27万余篇会议论文;“中国报刊”目前每年收录哲学社会科学文章16万篇;“西文期刊目次”收录15000余种西文期刊;“科技百花园”收录了100集系列科普片《新科技3分钟》和41集《科学智慧8分钟》共计约700分钟的录像节目^[3]。

清华大学数字图书馆(TDL)建立了大规模的“本馆电子资源”和“学科网络资源”。其中“本馆电子资源”包括Ei、INSPEC、FirstSearch、CSA、PQDD、DIALOG、ABI、NTIS、EBSCO、UMI、CAPSXpert、Web of Science(SCI、SSCI、A&FCI)、JCR、Web of Science Proceedings(ISTP、ISSTP)、DII、BIOSIS Previews、Chemistry Server、Current Contents、Elsevier Science、Academic Press、IEEE/IEE、JSTOR、Springer Link、Kluwer Online、Wiley等25种外文电子资源数据库和科技期刊报导、高校学位论文查询、万方数据库、China InfoBank、联机光盘库、光盘网络新资源、中国期刊网等7种中文电子资源数据库;“学科网络资源”包括网络导航、国内主要网络站点、国内上网图书馆、国外上网图书馆、国内外主要大学列表、Internet搜索工具、Internet教室、Science Online、中文核心期刊表、专利、虚拟图书馆、科技报告12种资源。正在建设中的数字资源系统包括“清华大学建筑数字图书馆”、“清华大学网上图书馆”、“清华大学学位论文检索系统”、“清华周刊”、“馆藏文物珍品”和“数字图书馆研究相关信息”^[4]。

2.2 检索特征

比较14个数字图书馆的检索方法与特征,得出表2。

2.3 输出格式

这14个数字图书馆具有各自不同的输出格式,见表3。

表 2 14 个数字图书馆的检索特征对照

项目名称	浏览/索引	简单检索	布尔检索	多字段检索	截词检索	近似检索	自然语言检索	比较检索	主题词表检索	短语检索	语音检索	词根检索	分类输出
ACM	Y	Y	1, 2, 3	Y	Y	1	N	N	N	Y	Y	Y	Y
AMMEM	N	Y	4, 5	Y	Y	N	N	N	Y	Y	N	Y	Y
BL	Y	Y	1, 2, 3	Y	Y	N	N	N	N	Y	N	N	N
BUILDER	N	Y	1, 2, 3	N	Y	Y	N	N	N	Y	N	Y	Y
CMDL	Y	Y	1	Y	N	N	N	N	N	N	N	N	N
DIGILIB	N	Y	4	Y	Y	N	N	Y	N	Y	N	N	N
GEMS	N	Y	1, 2, 3	Y	Y	Y	N	N	Y	N	N	N	N
NLC	Y	Y	1, 2, 3	N	Y	Y	Y	Y	N	Y	N	N	N
NLCN	Y	Y	1, 2, 3	N	N	Y	N	N	Y	Y	N	N	N
NZDL	Y	Y	1, 2, 3	Y	Y	Y	Y	N	N	Y	N	Y	Y
SETIS	Y	Y	1, 2	Y	Y	Y	N	N	N	Y	N	N	N
SHDL	Y	Y	1, 2, 3	Y	Y	Y	N	N	Y	Y	N	Y	N
TDL	Y	Y	1, 2, 3	Y	N	N	N	N	Y	Y	N	N	N
UMDL	Y	Y	N	N	N	N	N	N	N	N	N	N	N

注: Y 为可利用的功能, N 为不可利用的功能(下同);“布尔检索”中 1 为“和”、2 为“或”、3 为“非”、4 为“隐式和”、5 为“隐式或”。

表 3 14 个数字图书馆的输出格式

项目名称	输出格式	排序工具
ACM	每屏能显示 24 个条目, 内容包括标题、作者、出版信息、查准率、不同单元的利用率。	N
AMMEM	最多能显示 5000 个条目, 附作者和标题字段。	查准率
BL	图像显示可以放大到 3 倍, 并可显示手稿页码和手稿馆藏编号。	N
BUILDER	每屏能显示 10 个条目, 内容包括标题与一部分摘要信息。	N
CMDL	没有限制每屏条目, 每种馆藏已分组并字母顺序显示。	N
DIGILIB	每屏能显示 12 个图像, 有单页、摘要表和详细摘要三种显示方式, 图像压缩成 JPEG 格式。	特征/ 条件/ 材料/ 利用 / 前后关系/ 位置对象
GEMS	每屏能显示 10 或 25 或 100 或 200 个条目。	N
NLC	每屏能显示 20 个条目, 内容包括字数、标题字母、文献超链与大小。	作者/ 标题/ 日期
NLCN	能显示图标、标题、摘要、链接网址等。	N
NZDL	能显示书目信息和摘要, 不同种类馆藏有不同的输出格式。	N
SETIS	每屏能显示 100 或 200 或 300 个条目, 内容包括标题和摘要的一部分。	分组匹配
SHDL	每屏能显示 10 或 20 个条目, 不同种类馆藏有不同的输出格式。	N
TDL	能显示记录总数目、资源库名、分类主题、出处等	N
UMDL	每屏显示条目与字段取决于个体数据库。	N

由表 3 可知, 不同数字图书馆有不同的显示字段。ACM、AMMEM、NLC、NLCN、SETIS、SHDL 只显示了诸如作者、标题、期刊名、日期等少量细目; ACM、NLC、NLCN、TDL 能够利用超链显示所检索文献的摘要或提要; CMDL、SHDL、TDL 能显示多媒体文献的多媒体标题、电子图书的书名和作者名、电子期刊的期刊名; DIGILIB、NLCN、SHDL 和 TDL 能显示、打印或下载、编辑照片和图像; SETIS 和

BUILDER 能显示所检索的书目信息及其少量说明; ACM、NZDL、SHDL、TDL 和 UMDL 的合法用户可以下载全文或摘要; 只有 AMMEM、DIGILIB、NLC 能排序输出, SETIS 尽管不能实现排序, 但可分组输出。

3 今后我国数字图书馆建设的 3 个关键问题

目前我们应在数字图书馆结构模式开发、数字资源建设和信息检索技术研究等方面下大功夫,取得突破性进展,以推动我国数字图书馆发展。

3.1 数字图书馆结构模式

“浏览器——Web 服务器——数据库服务器”是目前普遍接受的数字图书馆结构模式。Web 服务器主要接收读者客户端的查询请求、进行数据处理和处理结果的发送,管理 HTML 构成的信息空间,提供对数据库的存取接口;数据库服务器主要负责管理数字化馆藏,通过接收 Web 服务器请求,对数据进行处理,然后把处理结果传送给 Web 服务器;读者客户端通过各种网络实现与 Web 服务器的连接,通过浏览器访问 Web 服务器提供的各种功能和丰富的数字化馆藏^[5]。这就是常说的“三角形结构模型”“图书馆客户机——图书馆服务器——多媒体对象服务器”是它的另一种提法。辽宁省数字化图书馆、美国国家图书馆^[6]和斯坦福大学数字图书馆 InfoBus^[7]都是运用这种结构模型来构建的。

密执安大学数字图书馆(UMDL)项目小组开发了一种基于代理协作的数字图书馆结构模型^[8]。它由用户接口处理(UIAs)、中介代理(Mediators)和馆藏接口代理(CIAs)三部分组成。用户接口代理提供用户接口的通信封装,这种通信封装有两种功能:一是用适当方式为 UMDL 协议封装用户提问;二是为各种代理发布用户简表,以指导检索过程;中介代理提供中介服务:将查询从 UIA 指引到某个馆藏,监视查询的进展情况,传递处理结果,转换格式,记帐等;馆藏代理给馆藏信息提供通信封装,执行翻译任务,发布馆藏内容和功能等。UMDL 定义了 Conspectus 语言,用来描述代理对某个代理协作组可做的贡献及其局限。UMDL 还设计了注册代理,由它负责维护 UMDL 系统中所有代理的内容和能力。这种基于代理协作的数字图书馆结构模式具有很强的模块性、换算性和扩展性,有利于充分发挥数字图书馆的多种功能。

面对网络信息和用户各项需求的不断增长,人们着手构建一种开放体系分布式数字图书馆结构模式^[9]。它由数字对象、信息仓库、索引服务器、收集服务器、代理服务器和用户接口网关、Handle System 几部分组成。数字对象即代表各种媒体形式的信息资源(包括文本、图像、音频、视频等),数字对象以通用资源名 URN(Universal Resource Name)所标识;Handle System 服务器可以将一个 URN 解析

为由 URN 所标识的一个或多个位置的数字对象;信息仓库提供对数字对象存放(Deposit)、存储(Storage)、访问(Access)等服务;索引服务器(Index Server)提供对数字对象的发现机制;收集服务器扫描一系列索引服务器,阅读其元数据,按收集定义原则决定索引服务器索引的哪些对象是指定集合的元素;代理服务器和用户接口网关利用一个或多个收集服务器和索引服务器所提供的信息建立查询路径,以允许对这些集合中对象的搜索和访问。代理服务器和用户接口网关与 Handle System 交互以处理索引服务器返回的 URN,并将 URN 所代表的一组数字对象的 URL 返回给客户端。这种数字图书馆结构模式的主要优点是允许创建任意数量的联邦型数字图书馆实例(Instance)——按特定协议构建的服务器集合,以响应服务请求并返回结果,具有无限可扩展性。这种结构模式能够对分布式数字资源的收集、存储、发布、检索等服务给予可靠的管理机制,我们应该进行深入的研究、开发与应用。

3.2 数字资源建设

加强我国数字资源建设必须采取协同开发、合作建库、统一标准、分步实施、突出特色、避免重复建设的基本准则。首先,协同开发与合作建库是数字资源建设的关键。数字图书馆资源建设不可能仅仅依靠几个图书馆和情报部门,而是必须依靠社会上所有信息资源拥有者的大力合作,形成数字资源建设群体。各单位、各部门在建设本地数字资源的同时,将元数据提供给数字图书馆中心及分中心。美国国家基金会早在 1994 年的“数字图书馆倡议”中,就将多方协同,联合进行,作为项目中标的必备条件;其次,我国数字资源建设应统一标准规范,按照统一标准加工、标引数字信息,避免出现各自为政、互不兼容的现象,保证数字图书馆资源建设的科学、有序。数字资源建设标准包括文献分类标准、数据描述标准、数据压缩标准等。其中文献分类标准有 DDC、UDC、LCC、IPC、《中国图书馆分类法》、《中国科学院图书馆图书分类法》、《中国人民大学图书馆图书分类法》、《中国档案分类法》等;数据描述标准包括文献著录标准、MARC 标准、置标语言标准、元数据标准等;数据压缩标准包括 JBIG、JPEG、MPEG、P * 64 等^[10]。我们应该从这些国内标准中选择并确定我国数字图书馆建设的同一标准;再次,我国数字资源建设应分步实施、突出

资源特色。《一期规划》分为准备和实验阶段(2000年)、初步实用阶段(2001 - 2002年)和规模型成长阶段(2003 - 2005年),这符合我国数字图书馆发展现状与规律。最后,我国数字资源建设还应避免重复建设,应成立属于中国数字图书馆工程建设联席会议办公室领导下的资源小组,协调、安排全国各单位的数字资源建设,并强化管理,避免资源重复建设。

3.3 数字图书馆信息检索技术

由表2可知,各国数字图书馆信息检索技术已经取得了一定进展,但目前这种状态很难满足今后智能化、个性化服务需求,因而我们必须进一步开发数字图书馆信息检索技术。所谓基于内容的检索(CBR),是指根据媒体对象的语义和上下联系进行检索。它主要包括基于内容的图像检索、音频检索、视频检索。

基于内容的图像检索是根据图像所包含的颜色、纹理、形状以及对对象(图像中子图像)的空间关系等信息,建立图像的特征矢量作为其索引来进行检索的。它包括3种实现方法:基于颜色特征的图像检索法;基于纹理特征的图像检索法;基于形状特征的图像检索法。

基于内容的视频检索就是在大量的视频数据中找到所需要的视频片断,一般由视频数据库生成模块、视频查询检索模块两部分组成。前者主要完成视频源数据的生成、视频数据的预处理及视频特征库的生成;后者根据用户提问完成指定的查询和检索任务。在分析视频数据后,就可进行基于关键帧、运动的检索和浏览。

基于内容的音频检索是将输入的字符序列和音频数据库中的字符序列相匹配,它主要是针对频域信息或其它声学属性,以及声音的概念(主观)特性的查询。上海交通大学数字图书馆创建了一个音频数据库的旋律检索系统,能够使非音乐专业人

员可以方便地采用常规方法和基于音乐内容即旋律的检索方法在网上进行乐曲的检索;音乐专业人员可以用乐句进行全曲检索^[11]。

今后我国数字图书馆应该整合基于内容的文本检索、图像检索、音频检索、视频检索技术,为用户提供高智能化的信息检索服务。

参考文献

- 1 N. meyyappan etc. A review of the status of 20 digital libraries. Journal of information science, 2000, 26(5)
- 2 周和平. 关于建设中国数字图书馆工程的问题. 中国图书馆学报, 2000(5)
- 3 刘炜等著. 数字图书馆引论. 上海: 上海科学技术文献出版社, 2001
- 4 <http://www.lib.tsinghua.edu.cn/>
- 5 陈敏. Internet时代的数字图书馆建设. 情报学报, 1999(6)
- 6 肖珑. 美国国家数字图书馆项目的进展. 情报学报, 1998(3)
- 7 Andreas Paepcke. Building the InfoBus: a review of technical choices in the Stanford Digital Library Project. <http://www.diglib.stanford.edu/diglib/wp/public/doc217.pdf>
- 8 William P. Birmingham. An agent-based architecture for libraries. D-Lib, 1995(7). <http://www.dlib.org/dlib/July95/07birmingham.html>
- 9 张健等. 开放体系分布式数字图书馆原型设计. 计算机应用, 2000(6)
- 10 高文等著. 数字图书馆原理与技术实现. 北京: 清华大学出版社, 2000
- 11 薛锋等. 基于内容的音乐检索. 大学图书馆学报, 1999(4)

盛小平 湘潭工学院图书馆馆员、硕士。主要从事数字图书馆和信息资源研究。通讯地址: 湖南湘潭工学院图书馆。邮编 411201。

(来稿时间: 2001-06-27)

本刊 迁址

中国图书馆学报编辑部已迁至国家图书馆东 618 室办公。
通讯地址: 北京市海淀区中关村南大街 33 号中国图书馆学报编辑部。
邮编: 100081。电话: 010—88545141。传真: 010—68417815