中国图书馆学报(双月刊)2004年第5期
ZHONGGUO TUSHUGUANXUEBAO

●吴建中

图书馆VS机构库

——图书馆战略发展的再思考

摘 要 从20世纪90年代开始,"DC热"、"语义热"及"机构库热"等新技术和管理方法,不仅对图书馆业务,而且对其生存发展提出了严峻挑战。图书馆必须重新审视其技术发展路线,积极应对范式转变,确立其知识交流中心的核心作用。参考文献14。

关键词 元数据 语义网 机构库 图书馆管理
分类号 G250

ABSTRACT Since the 1990s, DC, Semantic Web, Institutional Repository and other hot topics have presented great challenges to library operations and even the survival and development of libraries. The author thinks that libraries should review their own technological development paths, actively face the paradigm shifts, and establish their core roles in knowledge communication. 14 refs.

KEY WORDS Metadata. Semantic web. Institutional repository. Library management.
CLASS NUMBER G250

0 引言

在较长的一段时期里,图书馆利用现代信息技术在MARC和Z39.50等资源组织方式上进行了稳步而艰难的开发,取得了明显的成果。今天,读者之所以能够通过电脑迅速而准确地查询到图书馆的馆藏资源以及连接数个图书馆的联合目录数据库资源,正是得益于这些图书馆现代技术。但是从20世纪90年代开始,现代信息与通讯技术的飞速发展以及一系列接踵而来的技术性突破,如将结构引入非结构性数据,将结构化的描述方法引入管理非结构化的互联网信息内容,将语义描述机制引入结构化的信息内容管理以及开放源代码运动等,不仅从技术上对图书馆的业务活动提出挑战,而且将深刻地影响并改变图书馆的生存和发展模式。本文通过上述发生在图书馆外围的技术演变过程,探讨这些技术对图书馆产生的影响和启示,同时对图书馆的应对措施以及未来发展战略提出建议。

1 "DC热"对图书馆资源编目方式的冲击

美国加州大学伯克利分校研究人员发现,在1999年到2002年这3年间,世界范围内信息生产量以平均每年30%左右的速度递增。2002年,全球由纸张、胶片以及磁、光存储介质所记录的信息生产总量达到5万亿兆字节,约等于1999年全球信息产量的两倍,而新产生的信息中有92%记录在硬盘等磁存储介质上[1]。

在茫茫大海般的信息环境中,如何完整、准确并及时地查找到所需信息是摆在人们面前的一大难题。今天,已经很难想象单靠几个搜索引擎就能够管理并检索互联网上的所有资源。而且现在的搜索引擎技术基本上是基于字面匹配的,由于网络信息的多样性和不规则性,这种字面匹配的检索方式无法解决查准率低、误检率高的问题。为了有效地解决网络资源管理以及检索所存在的问题,20世纪80年代末、90年代初掀起了一股元数据研究热潮。很多元数据,如GEM、TEI、EAD、DC等相继出台。

实际上,图书馆界早就有人对MARC格式的繁杂性提出批评[2],他们看到了MARC的局限性,也感觉到传统的数据描述方式已经跟不上形势发展的要求,因此希望改变它、简化它,盼望着有一种既能解决数据的结构化问题、又能克服数据过于烦琐和复杂的资源描述形式能够早日出台。但是,由于当时DC(都柏林核心元数据)本身还处于研发阶段,尽管通过几次DC年会,形成了一定的技术规范,但其成熟程度仍与已实践多年的MARC(机读目录)体系无法相比,加上围绕元数据发展的技术和网络环境尚未成熟,图书馆界的争论焦点更多地集中在DC与MARC之间孰优孰劣上。

在相当长的一段时期里,MARC和AACR(英美

编目条例)一直是书目数据描述领域的主流工具。从世界范围来看,绝大部分的书目记录都是依据上述方式编制的。无论是从数据描述的丰富性,还是从数据检索的查准率来看,MARC/AACR都是名列前茅的,现在还没有哪一种元数据格式可以在这两个方面超过它们。但如果说图书馆把信息资源的组织和整理仅仅局限于馆藏资源的话,那么现在MARC和AACR就足以应付了。但是进入数字时代,原有的数据描述手段就明显地跟不上形势发展的要求了。因此,图书馆需要思考的不仅仅是MARC对现代网络环境是否适应的技术性问题了。

2 "语义热"对图书馆资源组织方式的冲击

20世纪末,为了应付互联网的爆炸性发展,改变网上信息的无序化状态,人们开始探索将结构化的概念引进互联网,也就是说,将新的、基于人类认知和语义的协议引入网络,让机器不仅做到"可读"信息,而且能够"读懂"信息,于是语义网(Semantic Web)的概念便应运而生了。所谓语义网,按照发明者Tim Berners-Lee的表述,就是:"当前互联网的一个延伸,在语义网中信息被赋予清晰的定义与含义,从而使人机可以更好地合作。"[3]

将结构化的描述方法引入管理非结构化的互联网信息内容,将语义描述机制引入结构化的信息内容管理,是互联网向智能化发展的飞跃,它所引起的不是量的变化,而是质的变化。

现在让我们来回顾一下这一技术演进的过程。首先,人们为了让互联网无需通过服务器地址就能够直接连接数字资源,发明了URN(统一资源名)系统,为了让HTML页面不仅能显示和链接,而且能蕴含更多的"结构"和"关系",发明了XML(可扩展置标语言)[4]。XML的核心在于以一种标准化的方式来建立数据表示的结构,而将具体标记的定义留给了用户。其可扩展性使XML可以满足各种不同领域数据描述的需要,并可以对计算机之间交换的任何数据进行编码。国际码(Unicode)、统一资源标识(URI)、XML及其相关技术如命名域(Namespace)和XML模式(XML Schema)构成语义网的原始物理基础,奠定了数据的格式与语法。可是随之而来的问题是,两个用XML表示的消息或数据,如何才能实现交换?为了让机器能够"读懂"XML中的特定内容,并解决不同元数据的互操作性和兼容性问题,便发明了RDF(资源描述框架)。RDF及其语言(RDF Schema Language)通过定义概念之间的关系,使数据能够自我描述,建立了描述层的格式与语法环境。

但是这些还不够,还需要赋予数据以语义,即数据的含义以及数据之间的关系。要让信息系统能够处理并读懂这些数据以及数据之间的相互关系,于是便开发了知识本体(Ontology)。知识本体语言(OWL)是语义网发展过程中的一个重要里程碑,它在信息系统和数据库的设计中起着规范数据含义的作用,从而为语义网的实现提供一个必不可少的逻辑基础[5]。

3 "机构库热"对图书馆整体管理模式的冲击

网格计算和开放源代码是近两年来对互联网发展产生重大影响的最新进展,而且都与图书馆发展有关。网格这一名词来源于输电网。它试图提供这样一种思路:人们把自己的个人电脑插入网格,就像插入墙上的电器插座一样,就可以使用网格上的各种计算资源和知识资源。网格计算源于元计算(Metacomputing),其初衷是将分布的多台超级计算机连接成为一个可远程控制和访问的元计算系统,网格体系根据高速和共享两大特征,正不断地向"基于更大的网、共享更多的资源、构建更大的计算能力"的目标挺进[6]。数字图书馆追求的是海量信息的有序组织以及分布式跨库检索,网格技术将有助于解决数字图书馆大并发量访问、海量信息检索以及馆际资源共享等问题。开放源代码的意义在于把大量人类宝贵的创造力从"重复创造"中解放出来。近年来,Linux采取开放源代码这种新的理念,通过互联网集中了全球软件工作者的智慧,现已具备与Windows相抗衡的能力。Linux等开放源代码技术对未来几年互联网发展的影响,除了将改变世界软件业的格局以外,也将扩展开放源代码这一软件发展的新模式[7]。今天这一应用在数字图书馆领域已经相当普遍。美国的OCLC,英国的UKOLN,以及新西兰的DL项目都是图情领域推动开放源代码应用的代表。

当人类刚刚步入21世纪的时候,一个得益于上述几乎全部研发成果的新生事物正在形成,那就是机构库(Institutional Repository)。2001年,俄亥俄州立大学的几位高级行政官员找到该馆馆长布兰宁(Joseph J. Branin),探讨如何开发一个远程教育体系。在他们的策划下,一个机构库的雏形——俄亥俄州立大学知识库(Ohio State University Knowledge Bank)便由此诞

生了。后来,一些大学纷纷建立了以发展大学学术数据库为目的的机构库,有杜克大学乐谱库、约克大学考古库以及弗吉尼亚理工大学的影象库等。其中影响最大的是麻省理工学院(MIT)建立的 DSpace 系统。该系统建立于 2002 年,由该大学图书馆与惠普公司(Hewlett-Packard Co.)共同开发。该系统建立的初衷,是为了让每年由该校研究人员生成的 1 万多件电子版学术内容实现网上共享。DSpace 是一个开放源代码的软件平台,主要代码均为 Java 编写,可以运行于所有 UNIX 系统。该系统使用 DC、OpenURL 与 OAI-PMH 等一系列开放标准,人们可以根据自己的需要来修改和扩展它的功能[8]。该系统发展迅速,并很快形成了一个由剑桥大学、哥伦比亚大学等七家著名大学直接参与的联合机构库。在成立后不到一年的时间里,已经有 3500 家来自全球的机构下载了 DSpace 的开放源代码[9]。MIT 的史密斯女士(M.Smith)说,"在很短的时间里,我们已经看到了开放源代码的成果,很多机构在帮助我们侦错和改进系统","如果每一个机构库都发展各自代码的话,我们就不会有今天的进步"[10]。

由此可见,机构库是指收集并保存单个或数个大学共同体知识资源的知识库,在学术交流体系改革的诸要素中扮演着关键的角色,即扩大对研究资源的存取能力,重申学术机构对学术的控制力,增强竞争力,减少杂志的垄断性,提高经济自救力和与各类机构及图书馆之间的关联性等。同时,在为提高大学质量的具体指标方面,在提高研究活动的科学、社会以及经济的关注度方面,以及在增强研究机构的知名度、地位及公共价值等方面创造了必要的条件[11]。

4 重新审视图书馆技术发展路线

图书馆为什么不能够眼光向外,从网络界最近的发展得到一点启发呢?众所周知,元数据即结构化数据的概念早在数百年前就运用于图书馆了,但是网络工作者抓住了这一灵感,在十年左右的时间里网络世界围绕元数据发生了一系列技术革命。然而,图书馆依然循着自己的一条技术路线在艰难地跋涉着,MARC 编目体系走了 20 多年没有出现质的飞跃,反而派生出各种各样的区域格式,同时 Z39.50 不仅变得越来越庞大,而且越来越复杂。根据 OCLC 最近的一个调研报告表明,Z39.50 下一步发展可能以适应互联网(web)环境为目标,即 SRW(Search and retrieve On the Web)和 SRU(Search and retrieve URL)[12]。也就是说,Z39.50 今后发展的趋势将是向适应互联网环境的方向发展。

可以预测,在未来的几年里,元数据和内容标准将继续朝着以 XML 框架及与之相关的技术环境向前发展。这些发展将使得图书馆应用技术更少专门化和领域特定性(Domain Specific)。其实,图书馆可以运用的现成工具很多,而且不少都是开放性的。我们已经看到 MARC 等已经在向适应 XML 的方向发展,如美国国会图书馆发布的 MARCXML Schema 等,但是图书馆需要应用和开发的项目还有很多,为了降低成本,更重要的是共享资源,图书馆应该采取更加开放的态度,积极与网络工程师合作,利用适应互联网的先进技术及现有成果,更好地为知识资源管理与服务提供技术及环境支撑。

5 积极应对图书馆范式转变

数字图书馆所追求的目标,不仅仅是对资源的发掘,还应加上对资源所含内容的发掘。美国肯特州立大学曾蕾博士指出,我们不仅需要对一个资源作结构上的分解(以便发掘其结构部分),而且需要对其作语义上的分解(以便发掘其有用的内容成分)[13]。今天图书馆信息和知识组织的重心已经从对资源外形的描述,深入到对资源内容的开发上。由"结构"和"语义"所引发的一系列创新技术不仅在改变互联网,而且也在改变图书馆。图书馆的管理模式也将随着范式的转变而转变。

过去,图书馆管理的对象主要是书,从采访、编目、流通到保存,整个业务流程都是以书为中心展开的。后来增加了其他载体的资料,如音/视频资料、缩微资料和电子资料等,图书馆的管理方式依然照旧。从 20 世纪 60 年代开始,图书馆界开始强调以人为本,重视参考咨询服务,但原来那种以书为本的管理模式并没有发生多大的变化。现在,图书馆管理的对象转向知识,从信息的收集、描述、储存到服务,整个业务流程都围绕着知识展开,并且更加强调人与知识之间的关系。原来那种一成不变的管理方式已经适应不了以动态和开放为特征的新型管理方式。业务流程中环节之间以及资源之间的界限变得更加模糊,而图书馆与用户之间的供需关系变得更加密切,一切与知识相关的要素都被有机地整合和调动起来,形成一种面向需求、适应变化的知识管理与服务机制。

6 重新确立图书馆在知识交流中的核心作用

尽管机构库是在图书馆的影响下发展起来的一种知识库，但大多数都是由大学及研究机构管理或开发的。机构库是知识基础设施中有关知识存储价值链的一个数字空间，如DSpace。它把图书馆放在一个机构库体系中，并要求图书馆重新思考自己的生态体系，使其更加能够适应各类用户对知识的需求。他们认为图书馆一直在从事很多重复的劳动，而且现在的图书馆数字管理体系实际上仍然是以纸张为基础的，只不过是借用电子手段而已。机构库面向所有的人，但重点是知识群体，它不仅存储成形的信息，而且存储未成形的信息，如片段的数据或信息等。它的起源据说是MIT的一个教授想把自己做成数字式的成果捐献给图书馆，但图书馆无法收藏它们，所以他们认为现在的图书馆不能适应新的需求，遂发起建立一个跨越学校各院系和图书馆的虚拟空间，后来便形成了现在的数字空间。

作为一种知识库，机构库应该成为图书馆的一个组成部分，但目前它们之间似乎只是一种伙伴关系，一方面大学教研人员对图书馆的传统观念与服务手段一直持批评态度，另一方面由于图书馆管理体制本身的缺陷，图书馆员不思变革的惰性相当严重。在有关MARC与DC的争论中，我们可以感受到一些图书馆员的抵触情绪，使得争论的焦点一直停留在两者的孰优孰劣上，而很少冷静地去思考DC元数据的发展对图书馆的业务活动具有什么样的影响或启示。当一批元数据体系逐步走向成熟，并积极与其他网络技术结合，走出一条资源发现之路的时候，图书馆仍抱着观望的态度，沿着自己的一条技术路线往前走。今天，当机构库充分利用一切可运用的技术手段，在短时间内迅速崛起并形成规模的时候，图书馆不得不思考今后发展的定位了。

7 结论

在《战略思考：图书馆发展十大热门话题》中，我曾经说过这样一段话："如果认为图书馆没有必要去关注互联网资源组织和整序的话，或者如果认为图书馆没有必要在互联网上占有一席之地的话，可以说，靠着千百年来积累的经验和知识，图书馆这一专业似乎还可以维持一段时间，但是我们已经可以看到尽头了。"[14]今天，我们面临着更加严峻的挑战，机构库是在图书馆的外围发展起来的虚拟知识库，今后它有望形成一个全球共享的知识库。作为以"知识服务"为己任的现代图书馆，应该充分意识到这一挑战对自己意味着什么。未来的图书馆不仅要整合信息资源，而且要整合各类知识资源，如大学、研究机构以及实验室形成的知识资源，因此，机构库对图书馆发展具有深远的影响和意义。图书馆只有采取积极的姿态，迎接挑战，与时俱进，才能真正实现自己的价值，完成自己的使命。

参考文献

1 毛磊.美国研究揭示全球"信息爆炸"现状.http://www.fsonline.com.cn/IT/news/200310310048.htm(查询于2004年3月11日)
2 莫少强.网络环境下目录学发展的新课题——数字图书馆元数据和资源共享的研究与实践.http://www.chinalibs.net/book/wlml.doc(查询于2004年3月11日)
3 Tim Berners-Lee, et al. [Online]. Available: http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2 (accessed March 11,2004)
4 刘炜,张亮.数字图书馆的体系结构与元数据方案.http://www.1ibrary.sh.cn/libnet/sztsg/fulltext/reports/2002/metadatadesigning.htm#_ftnl(查询于2004年2月12日)
5 秦健.实用分类系统与语义网：发展现状与研究课题.现代图书情报技术,2004(1)
6 包冉.国家网格在路上——2003网格技术与应用研讨会侧记.http://www2.ccw.com.cn/04/0402/b/0402b17_1.asp(查询于2003年3月11日)
7 Linux等开放源码技术－软件业革命的"导火索".http://tech.sina.com.cn/other/2003-12-28/1433274441.shtml(查询于2003年3月1日)
8 Dspace研究.http://libweb.zju.edu.cn/04/dspace/index.htm(查询于2003年3月1日)
9 Tom Storey. University Repositories: An Extension of the Library Cooperative. OCLC Newsletter,2003(7):7~11.
10 Ibid.,10.
11 Raym Crow. The Case for Institutional Repositories: A SPARC Position Paper[Online]. Available: http://www.arl.org/sparc/IR/ir.html(accessed March 11,2004)
12 The 2003 OCLC Environmental Scan: Pattern Recognition. Dublin,Ohio:OCLC,2004,94~95
13 曾蕾.元数据与专业置标语言在数字图书馆中知识表述方面的功能.图书情报工作,2002(10)
14 吴建中.战略思考——图书馆发展十大热门话题.上海：科技文献出版社,2002

吴建中 上海图书馆馆长，研究馆员。通讯地址：上海市淮海中路1555号。邮编200031。

(来稿时间：2004-04-26)