论文献信息资源的数字化建设

刘 路 孟连生

【摘 要】 文献信息资源的数字化是促进资源共享、加快数字图书馆建设的重要基础。本文在分析数字化信息资源的类型、特点以及现阶段数字化建设中几个关键性问题的基础上，探讨了文献信息资源数字化建设的一些基本原则，以期探索出一条科学化、标准化的数字化建设道路。

【关键词】 数字化 信息资源建设 标准化 共建共享

无论未来图书馆如何发展，其最终的目的都是为了进行信息服务。而信息服务的前提则是信息资源建设。当前，图书馆发展的大趋势是建设数字化、网络化的数字图书馆。因此，作为数字图书馆形成和发展重要基础的文献信息资源数字化建设，就变得非常重要和迫切。它主要包括三个方面的内容：一是将传统图书馆馆藏资源数字化；二是采集、整理和存储网络信息资源；三是购买数字化产品。而将馆藏资源数字化，则是文献信息资源数字化建设中的重点和难点，而且其中还存在很多亟待解决的问题。

信息资源的数字化，就是将信息资源通过机器设备的处理，转化为计算机可识别的二进制代码，从而可以方便地对它们进行复制、检索、传递和存储。本文将以信息资源的数字化为中心，较为全面地探讨文献信息资源数字化建设的相关问题。

1 数字化信息资源的类型及特点

图书馆传统的馆藏主要是图书、期刊、报纸、技术报告、学位论文、专利文献、会议录等印刷型文献和缩微制品、音像制品等非印刷型资料。将这些馆藏的文字材料、声音视频信息进行数字化转换和处理，并以计算机可以识别和处理的字符编码形式或图像形式、多媒体形式存贮在各种光、磁性介质上，具备查阅、检索、复制等功能，就形成了数字化信息资源。

目前，数字化信息资源从组织形式上来看，主要有以下三种类型：

1.1 文件型

在文件型的数字化信息资源中，大部分是那些古典、经典文献信息的数字化产品，比如中国古典文学四大名著、四书五经、金庸小说等，它们都以印刷本的册或卷为单位，一个单位对应一个文件。其中主要的文件格式有：纯文本文件(以 txt 为文件扩展名)，可以在 Windows 环境下用记事或写字板等软件打开，在 DOS 下也可以阅读(加中文平台)；Word 文件(以 doc 为文件扩展名)，在 Windows 环境下用 Microsoft Word 字处理软件打开、阅读和编辑；PDF 文件(以 pdf 为文件扩展名)，可使用 Adobe 公司的 Acread 阅读器软件阅读。

除上述几种文件格式外，还有很多自行开发用来处理数字化信息的软件，需要相应的阅读器软件来阅读和编辑，在这里就不逐一列举了。

另外，图形、图像、图表、音频和视频等非结构化信息，也都可以用一定的文件格式存储下

来，主要情况如下：

	图形、图像	图表、表格	音频	视频
应用软件	Adobe photoshop；Microsoft Painter；ACDSee	Microsoft Excel	Microsoft Windows Mediaplayer；Winamp；	Microsoft Windows MediaPlayer；超级解霸 RealPlayer
文件扩展名	.psd .bmp .jpg .gif	.xls	.wav .midi .mp3	.mpg .avi .rm

单一文件型的数字化信息资源，在存储介质上放于一个或多个目录中，用户通过类似Windows资源管理器的目录树来查找和利用。Internet上的FTP服务以及大部分光盘资料，都是以单一文件的方式提供数字化信息资源的。其主要特点是：管理和利用简单方便。计算机已经有一整套成熟的文件管理的技术与方法，在组织数字化信息资源时可以借用。用户在使用时能够直接访问，简单方便，非常直观；信息独立性强，便于有选择性地复制和传递。由于文件型数字化信息资源均以单一文件的形式存放，彼此之间相互独立，既没有链接关系，也不受数据库结构的限制，因而方便了用户的随意选择和复制；缺点：信息在不断地增多，单一文件的组织形式无法组织管理海量信息。除此之外，文件系统只能反应信息之间简单的位置关系，对于结构化信息就无能为力了。

所以，文件型组织方式虽然在一定范围内应用较多，但是由于其固有的局限性，使之不可能成为数字化信息资源类型的主流。

1.2 超媒体型

超媒体型数字化信息资源是超文本与多媒体技术结合的产物。它以超文本的方式，将文字、表格、图像、音频和视频等多媒体信息组织起来，使用户可以通过高度链接的网络结构在各种信息库中随意浏览，找到需要的信息资源。

Internet上的数字化信息资源，主要就是以超媒体的形式存在，它们多以htm或html、asp、phtml、dhtml等为文件扩展名。主要的特点是：伸缩性，在超媒体型信息中，互相链接的文件可多可少，在很大程度上还能够改变其结构，其中任何一个文件，都只是整个网状结构中的一个节点，可以随时删除、添加或更新一个或多个文件；非线性，超媒体型信息以相互链接的网状形式存在，用户不必考虑文件存于何处，只要有超链接关系，就可以访问和使用目标。这种非线性的形式符合人们思维的关联和跳跃；缺点，网状的组织方式，使用户在通过浏览方式进行信息搜索时，很容易出现“迷路”的情况，超媒体型信息对用户而言，很难加以归类，不利于相关主题信息资源的搜集。

1.3 数据库型

数据库是对大量的规范化数据进行组织管理的技术。目前建设的数据库主要有书目数据库、题录数据库、文摘数据库和全文数据库等几种类型。数据库技术利用严谨的数据模型对信息进行规范化处理，利用成熟的关系代数理论进行信息查询的优化，大大提高了信息资源管理的效率。其主要特点是：提高了对大量的结构化数据的处理效率，弥补了文件型和超媒体型组织方式的缺点；大型的数据库，是对海量信息资源进行优化和组织管理的强有力工具；检索利用数据库内的信息资源十分方便，现代数据库技术也提供了许多有用的检索工具和检索手段，

使得检索结果在检全率和检准率方面都有极大提高；数据库的不足之处大致有三个方面：一是数据库技术限于严格的数据模型规范，不能提供数据信息之间的关联；二是界面缺乏直观性和良好的人机交互性，这是由关系数据库系统的检索结果以记录集合形式出现造成的；三是对非结构化信息的处理比较困难，面对日益增多的多媒体信息以及表格、程序、大文本等非结构化信息，数据库技术还显得有些力不从心。

总的来说，尽管以上三种类型的数字化信息资源各有优劣之处，但它们仍然是当前组织数字化信息资源的主要方式。不过，无论数字化信息资源采用何种类型进行组织，其根本的基础还是数字化信息资源的建设。

2 文献信息资源数字化建设的几个关键问题

文献信息资源数字化建设的基础是信息资源的数字化，简言之就是将印刷型及其它非数字化信息资源进行数字转换。信息资源的数字化，是一个重建信息资源体系的过程。在这个过程中，我们无法回避现实中存在的三个关键问题：数字化技术、数字化对象范围和数字化信息资源的共建共享。

2.1 数字化技术

信息资源数字化技术在目前还并不成熟，而且成本较高。现阶段的两种主要方式是键盘录入和扫描输入。

键盘录入是一种手工转换，常用的汉字输入法有五笔字型、自然码、拼音码、智能码等，其缺点在于速度慢、成本高、录入不全、效率低。笔者曾参与《燕京大学学报(哲社版)》回溯性全文数据库的原文文献数字化工作，由于文献中有大量的古文字、生僻字、象形符号无法用输入法录入，只能通过造字和编码进行数字化转换，因此大部分工作人员都不得不忙于查找、记录这些文字或符号。由此可见，键盘录入技术实非有效的办法。

而使用扫描仪扫描输入，计算机获取信息又只能以图像形式存储，不能进行检索和文字编辑。虽然可以通过光学字符识别(OCR)技术结合图像分析加以转换，变成计算机可读的字符编码，但由于图像质量不高、字符分类、源文献的背景干扰、扫描速度、单位成本较高等因素，使得这一技术的实用性也不强。例如，美国国会图书馆曾计划在 2000 年 200 周年馆庆时完成对 500 万幅历史馆藏的数字化转换，但在预计的时间内完成的数字化图像却只有 1.4 万幅。

不难看出，提高信息数字化技术和降低相关成本，是进行文献信息资源数字化建设的过程中必须解决的一个关键问题。

2.2 数字化的对象范围

图书馆馆藏信息资源非常丰富，如中国国家图书馆截止 1999 年底，馆藏文献已近 2 260 万册(件)；中国科学院文献情报中心也已收藏各类文献 560 余万册(件)；高校中的北京大学图书馆馆藏已达 461 万册(件)。如果要把这些馆藏都转化为数字化信息，不但任务繁重，耗费大量的人力、物力和财力，而且也毫无必要。文献信息资源数字化建设并不是盲目地全盘数字化，必须建立一定的标准，有选择地决定哪些信息资源应该数字化、哪些信息资源不需要数字化以及哪些信息资源应该优先数字化。

根据图书馆处于社会文献信息服务机构的这一地位，我们应该首先选择利用率最高的那部分文献信息资源优先实施数字化。文献信息资源数字化建设同传统的信息资源建设一样，应该服从于图书馆信息服务的宗旨，最大限度地满足绝大多数读者的主要需求。“在适当的时候

提供给适当的读者以适当的图书。”我想，阮冈纳赞这句至理名言在数字化、网络化的信息时代，对于图书馆工作仍然非常适用，只不过这句话中的“图书”二字要改为“信息资源”或者“数字化信息资源”了。所以，在确定数字化的对象范围时，应首先考虑的就是读者利用率最高的那部分信息资源。

接下来，就应该选择本馆的珍藏和特藏文献信息资源实施数字化。中国国家图书馆的藏书可上溯到700多年前的南宋皇家缉熙殿藏书，最早的典藏可以远溯到3000多年前的殷墟甲骨。国家图书馆的珍品特藏包括善本古籍、金石拓片、古代舆图、敦煌遗书、少数民族图籍、名人手稿、革命历史文献、家谱、地方志和普通古籍等260多万册(件)。外文善本中最早的版本为1473～1477年间印刷的欧洲“摇篮本”。这部分藏品极为珍贵，闻名遐迩，世界瞩目。其他各馆也都有自己的珍藏，一般都是孤本或残本。将它们进行数字化转换，一是为了制作“复本”更保险地将它们保管和留传下去；二是使它们可以重新展示在世人面前，并为更多的研究者提供研究素材。因此，将这些珍藏纳入数字化的对象范围，具有现实和长远的意义。

图书馆各自的特藏，体现了每个图书馆不同于其他图书馆的特色和价值，是图书馆在合作与竞争并存的信息时代更好地生存与发展的重要保障。馆内特藏也是数字化的重要对象，例如美国国会图书馆于1989年启动的“美利坚记忆工程”(American Memory Project，1989～1995)，就是以特藏为标准。这项工程首先选择了反映美国历史、文化和立法方面的文字手稿、图书、图片、地图、照片、音乐、乐谱、电影等作为数字化对象，将它们转换为电子格式，使得人们能够共享国会图书馆的珍贵特藏。

2.3 数字化信息资源的共建共享

第66届国际图联理事会全体大会于2000年8月13～18日在中东的耶路撒冷召开，大会的主题是：“用于合作的信息：构建起未来的全球图书馆。”其关键词就是“合作”。对于数字化建设来说，合作就是数字化信息资源的共建共享。

我国的图书馆体系存在不同的系统，其中有公共图书馆系统、高校图书馆系统、科学院图书馆系统等等；在地域上，也存在条块分割、各自为政的情况。这不但导致了各个图书馆在传统的信息资源建设中，形成“大而全”、“小而全”的信息资源体系，造成书刊大量重复，整体上文献信息资源种类匮乏。而且，这种固有的状况，还对文献信息资源的数字化建设产生了极为不良的影响。主要表现在如下几个方面：数字化产品的重复建设，造成人力、物力和财力的严重浪费；不利于知识产权的保护，在文献信息资源数字化建设中非法数字化、非法拷贝和传递等侵害版权的情况屡见不鲜；由于数字化信息资源成本高，在一定程度上具有独占性，因而价格高、服务差，大大降低了数字化信息资源的流通率和利用率；文献信息资源数字化建设这种无序、无规的现状，加上来自图书馆以外的竞争，使得图书馆在信息时代的地位和竞争力有所下降。

因此，在文献信息资源的数字化建设过程中，同样需要开辟一条共建共享的路子。

3 文献信息资源数字化建设的原则

进行文献信息资源数字化建设，必须有应该遵循的原则，以便在建设中能够有章可循、有法可依。就目前的各种情况和问题而言，主要有以下几个方面值得注意：时效性原则。即是在数字化技术不断进步的前提下，提高数字化建设的速度和效率，不能拖延数字化建设的进程，更不能将数字化建设的计划或工程无限期地推迟甚至不了了之；经济性原则。这主要是指在建设中节约资金，避免重复建设。其次，做到专款专用，集中力量把数字化建设搞好。最后是不能

盲目追求全面数字化，需要有层次、有目的地选择；质量原则。数字化建设中，最忌讳的就是只重数量，不重质量。在确立的数字化对象范围之内进行高质量的数字化建设。要树立精品意识，生产出来的数字化产品不但要忠于原文文献，还要便于检索、利用和合法的传播；标准化原则。在数字化建设过程中，不论是建设数据库型数字化信息资源，还是单纯地形成文件型或是多媒体型数字化产品，都应该遵照统一或者相互兼容的标准和协议，这是数字化信息资源共建共享的一个前提。

现阶段，国内主要是由文献工作标准化技术委员会制定相关标准，另外，还应该遵循其他技术委员会制定的有关信息管理、数据转换、信息检索、文献著录等方面的规范。当然，最好是都能够统一按照国际标准来开展工作。以下是三大国际标准化团体：ISO(国际标准化组织)、IEC(国际电工委员会)、CCITT(国际电报电话咨询委员会)制定的部分标准、协议：

类别	标准/协议名称	适用范围
电子出版物	ISO8879 标准通用置标语言，即 SGML	可用于数字化信息、电子出版物的内容格式描述
	ISO/IEC10744 多媒体基于时间的结构化语言	用于描述依赖于时间的数字化信息
	ISO/IEC DIS10719 文体语义与规范语言，即 DSSSL	描述文体通用树结构的转换和语义构造
	ISO/IEC DIS10180 标准页面描述语言，即 SPDL	用于文体页面的描述
	ISO/IEC9541-1，9541-2，9541-3，9541-4	字型信息交换系列标准
网络服务与协议的标准	ISO10162,ISO10163，Z39.50(美国国家标准化组织)	信息资源检索协议标准
馆际互借(ILL)	ISO10160	ILL 服务模型及服务定义
	ISO10161	ILL 协议说明

4 结语

总之，在文献信息资源的数字化建设过程中，还有很多需要继续深入研究的问题。但是，通过对以上各个方面问题的探讨，我们大致可以清楚在文献信息资源的数字化建设中，至少必须兼顾这几个方面：避免文献信息数字化的重复劳动，降低数字化建设成本；严格保护数字化产品的版权和生产者的权益；方便用户的检索与查询，节省时间并注重提高检全与检准率；提高数字化及非数字化信息的利用率和流通率，即提高信息资源的共享程度；要致力于增强图书馆和文献信息机构在信息时代的竞争力。

参考文献

1 孟广均，徐引篪．国外图书馆学情报学研究进展．北京：北京图书馆出版社，1999

2 孟连生．简评 90 年代中国文献数据库建设．情报科学，1999(3)

(下转第 66 页)

企业高素质的信息人才结合起来，创造出一流的信息产品。另一方面，现代社会是个竞争异常激烈，同时依附性又很强的社会，虽然图书馆的公益性使其没有直接的竞争压力，但社会的信息化、一体化趋势，要求图书馆的信息服务必须是开放型的，必须社会化。图书馆应积极主动向读者宣传“社会事业社会办”、“图书馆是我们自己的家”，恳请社会各界人士参与图书馆的服务与管理，同时，也可采取图企联姻、军民共建、社区共建、横向联合等多种途径，来促进图书馆事业的发展，提高图书馆满足社会需求的能力。

二是强化全民信息意识。强化全民信息意识，激发全社会对信息服务的需求，是图书馆信息服务社会化的重要条件。目前，我国国民的信息意识不强，现有信息基础设施及信息资源尚未得到充分利用。一般说来，当潜在的信息需求没有转变为现实需要时，是不能刺激信息利用的。因此，要强化全民信息意识教育，增强公众对信息、对图书馆的需求，以促进图书馆信息服务的快速发展。

三是注重培养信息服务人才。第一，要加强对信息服务人员的信息技术、信息处理和信息检索知识的培训。第二，通过各类培训班或高等教育，加紧培养，使他们成为复合型人才，使他们既懂得信息学和信息加工整理技术，又要精通电脑和通信技术，还要懂得市场营销，把信息产品直接推销给用户。第三，完善各类管理和考核制度，做到“人尽其才，各尽所能”。

四是组织好网上信息资源。创建与完善有自己特色的图书馆网站，并通过它来加强网上信息资源的引导服务。一方面可组织专业人员对网上丰富的信息资源进行分类，做成文摘或索引，增加信息服务的选择性和针对性，另一方面，还可以建设或购买一些有针对性的数据库放在网上，或者通过选择链接一些商用数据库、免费信息资源等，为用户提供全方位直接的信息服务。

参考文献

1 周敬治．网络环境下信息服务的发展趋势及对策．中国图书馆学报，2000(4)

2 周向荣．网络环境下图书馆的信息服务．图书馆，2000(3)

3 孙达武．高校图书馆网络建设的社会化趋势．大学图书馆学报，2000(1)

4 李纲．论信息服务业产权重组与机制转换．中国图书馆学报，2000(2)

5 王蔽雯．对图书馆发展信息产业的认识与思考．图书馆建设，2000(2)

林 清 福建建筑专科学校图书馆

（上接第 41 页）

3 陈子玲．知识经济时代图书馆的生存与发展．天津：天津人民出版社，1999

4 陈春．网络环境下的图书馆文献资源建设与共享．中国科学院第十一次图书馆学情报学科学讨论会文集．北京：中国科学院文献情报中心，2000

5 卢子博．跨世纪的思考——中国图书馆事业高层论坛．北京：北京图书馆出版社，1999

6 刘路，沈英．信息时代图书馆的地位、功能、模式探讨．图书情报工作，2000(5)

7 张智雄，宋小冬．联合西文期刊篇名目次系统的设计与实现．现代图书情报技术，2000(1)

8 http://elsevier.lib.tsinghua.edu.cn

9 http://www.ifla.org/iv/ifla66/66intro.htm

刘 路 中国科学院文献情报中心硕士研究生

孟连生 中国科学院文献情报中心博士生导师