第 58 卷 第 6 期
2005 年 11 月
武汉大学学报(哲学社会科学版)
Wuhan University Journal (Philosophy & Social Sciences)
Vol. 58. No. 6
Nov. 2005. 880 ~ 884

■情报学

馆藏文献数字化调研报告

黄 萃
(武汉大学 信息管理学院, 湖北 武汉 430072)

[作者简介] 黄 萃(1978-),女,江西新余人,武汉大学信息管理学院博士生,主要从事数字信息资源管理研究。

[摘 要] 2003 年 12 月至 2004 年 3 月间,作者参与对文化部档案处、国家文物局、中共中央对外联络部档案处、外交部档案处、教育部档案处、国家图书馆、北京市档案局、清华大学档案馆、电影机械研究院、京东方软件股份有限公司等机构的馆藏文献数字化工作进行了调研。此报告在这些调研基础上写成。

[关 键 词] 馆藏文献;数字化;调研报告
[中图分类号] G202 [文献标识码] A [文章编号] 1672-7320(2005)06-0880-05

馆藏文献数字化是通过采用摄影摄像、图形图像转换等数字信息技术手段,将传统介质的图书、档案文书等文献资料转化成计算机可以识别和处理的数字信息,经过整理和组织后,存储在计算机存储设备里,目的是为了快速检索、提高档案信息的利用率和提供远程信息服务。其目标在于建立起具有多媒体、多视角、多维立体文献信息体系,实现文献目录信息、文献全文信息、多媒体图形图像信息的一体化集成管理,便于提供利用和信息资源共享,从而更好地弘扬中华文化,共享人类遗产。馆藏文献数字化是电子政务建设、数字图书馆建设、数字档案馆建设、数字博物馆建设的前提与基础。国家自然科学基金课题组"国家文化资源数字化战略"的 4 人调研组,在 2003 年 12 月至 2004 年 3 月间,对文化部档案处、国家文物局、外交部档案处、教育部档案处、国家图书馆、北京市档案局等机构的馆藏文献数字化工作进行了调研。下文就是根据被调研机构中文献数字化开展状况、技术模式、管理模式等方面进行的总结与分析。

(一)各机构馆藏文献数字化开展状况

各调研机构的馆藏文献数字化工作开展状况如表 1 所示。

各被调研机构在馆藏文献数字化工作上,大致都根据馆藏基础,分析用户需求,研究馆藏体系,分三步走,首先实现馆藏文献目录数字化,编制电子目录检索工具,实现目录级的管理与检索利用;其次,有条件的单位机构有步骤地实现馆藏纸质文献全文数字化,在充分调研的基础上选择最优的数字化方案;然后,有实力的单位机构开始着手实现多媒体图形图像信息的数字化处理。

从理论上说,全部馆藏文献都可作为数字化的对象,例如卫生部档案处就将 1948 年至 2002 年的所有文书档案都进行了数字化处理。但其余单位机构一般都是从实际情况出发,根据自身的财力物力以及数字信息需求等基础要素进行思考,有选择有计划地进行馆藏文献数字化。(1)根据馆藏文献的保存价值进行数字化的,将保存价值高的馆藏文献优先数字化,例如文化部档案处,将 1 万卷保管期限为永久的档案划入了数字化行列;(2)根据馆藏文献的利用价值进行数字化,选择数字化的内容主题与社会信息利用需求相结合,把利用率较高的馆藏文献进行数字化,例如,国家图书馆的中国研究资源库;(3)根据馆藏文献的差异价值,确定本机构的特色文献,并选择这些馆藏文献优先数字化,提供独具优势的服务,例如国家图书馆的数字方志资源库、石刻拓片资源库、甲骨文献资源库、博士论文资源库、民国时期中文期刊资源库。从某种意义上说,特色就是优势,特色馆藏信息和特色服务往往是赢得用户和创造效益与价值的关键所在;(4)根据用户需求进行数字化,一些专业信息利用者要求文献信息机构提供综合性强、全方位的信息服务,因此,要以用户的需

收稿日期:2005-03-12
基金项目:教育部人文社会科学基金项目(01JA870010)

求为导向，做到按“需”数字化。例如，中国国际广播电台将所有的照片档案进行了数字化，积水潭肿瘤医院将所有的医疗档案进行了数字化。

表 1　各调研机构馆藏文献数字化状况表

调研机构	馆藏文献数字化程度	文献数字化规划范围
卫生部档案处	83 万多页馆藏文献进行了数字化	将非典期间 300 多盒文献资料进行数字化；1979 年以前所有档案文献进行数字换；1948 年至 2002 年所有文书档案进行数字化（工作状态：已完成）
国家文物局	一千多卷馆藏文献进行了数字化	所有的历史档案（工作状态：已完成）
文化部档案处	处于起步阶段，有 539 卷进行了数字化，	1 万卷保管期限为永久的历史档案（工作状态：进行中）
教育部档案处	尚未全面开展，进行了一些数字化前期准备工作，	1965 年前的几万卷历史档案（工作状态：尚未开始）
外交部档案处	已有 250 万页馆藏文献进行了数字化	馆藏文献 30 多万卷历史档案（工作状态：进行中）
北京档案局	600 多页纸质档案进行了数字化，1000 多个小时声像档案进行了数字化	2005 年，将 5-10％馆藏进行数字化；2008 年，将 20％馆藏文献进行数字化，大约 2400 多万页文献资料（工作状态：规划中）
国家图书馆	已完成 6380 万页图书、近 2000 部影片、22 万首音乐作品、4000 页馆藏西夏文资料、8000 幅金石拓片、180 万拍民国期刊、近 8 万篇博士论文等数字化工作	数字方志资源库；石刻拓片资源库；甲骨文献资源库；博士论文资源库；民国时期中文期刊资源库；音频视频资料资源库；网络信息资源库；馆藏各类文献书目数据库；中国研究资源库；《永乐大典》资源库（工作状态：进行中）

馆藏文献数字化应在充分调研的基础制定优先次序方案。目前，世界上发达国家在进行馆藏资源数字化时的优先选择对象，一般锁定三类文献：特色文献（如美国国会馆的“美利坚记忆”）、珍贵文献（如日本国会馆优先对珍善古籍进行数字化处理）和利用率高的文献。我国现阶段在选择馆藏文献优先数字化对象时应当借鉴国外经验，同时从我国国情出发，综合考虑各方面要素的制约条件，然后再去审慎抉择。最终要实现馆藏文献数字化工作的经济效益和社会效益最大化。

（二）各被调研机构文献数字化的工作模式

各被调研机构文献数字化工作模式的选择大致如表 2 所示：

表 2　各被调研机构文献数字化工作模式说明表

工作模式	业务外包	自主开展	数字化业务承包方
被调研机构	文化部档案处、国家文物局、外交部档案处、教育部档案处	国家图书馆、北京市档案局	电影机械研究院、东方基业

各组织机构在进行文献数字化工作模式的选择时，主要考虑了以下一些主客观因素：（1）管理制度方面，有没有进行数字化建设的组织保证。即是否建立起有关数字化建设的科学管理体制、规章制度、整体规划，有没有明确数字化的范围，领导及其各工作人员的职责是否明确分工等；（2）硬件方面，有没有进行数字化建设的物质保证。即是否有进行数字化工作的硬件条件，包括计算机、服务器、高速扫描仪、打印机、信息存储设备等的配置，以及信息系统与网络环境的建设等；（3）技术方面，有没有进行数字化建设的技术保证。文献数字化需要数据库技术、数据压缩技术、高速扫描技术、数据存储技术、智能检索技术、光学字符识别技术、视音频捕捉技术等技术支持，因此各机构要根据自己的实际情况和技术条件，在进行数字化建设之前进行客观地分析；（4）人力资源方面，即有没有进行数字化建设的人力资源保证。

（三）各机构馆藏文献数字化业务流程模式

文献数字化是一个系统性业务工作，其工作业务流可以简要概括为文献信息获取、信息处理、信息存储、信息发布与利用等 4 个核心业务过程：（1）文献信息获取有多种渠道，但主要是三种文献信息获取形式，即纸质文献扫描加工、模拟

文献(如磁带、磁盘文献)的数字化转换以及采取数码照相技术将文献实现数字化。(2)信息数字化处理是数字化解决方案的核心功能,主要包括对文献信息的编目、标引、图像文件处理、图像识别处理以及将图像与文献目录信息进行一致性关联等内容。该过程的每一个功能模块都需要借助于软件开发平台建立相应的用户操作环境。(3)信息存储是整个系统得以有效运行的支撑保障。这个环节中首先应根据系统存储量的需求、安全管理的基本要求以及应用访问的速度等因素选择存储设备,如磁盘阵列、光盘塔、网络存储设备等,其次是选择各类电子文献信息的存储和访问方式,如采取文件存储还是采取数据库存储方式等。(4)信息发布与利用是文献数字化的主要目的之一,该环节中除了需要建立文献信息的查询与利用平台以外,更多的是需要考虑哪些信息可以在网上公开发布,哪些信息是采用权限控制的方式在网上进行查询利用,这些问题的解决必须服从各行业领域的实际业务管理条例。

1.纸质文献数字化加工流程。纸质文献数字化,就是将纸质文献转化为基于原文影像及标引信息(或全文信息)的数字文献信息的过程。其工作流程主要包括文献整理、扫描、OCR 识别(如果需要实现全文检索,可采用 OCR 技术)、图文编辑、图文质检、重新装订、备份等多道工序。该流程要求,支持工序回馈,形成一个质量控制系统。

文献整理是数字化加工流程的预备工序,主要是将文献资料按归档要求进行分类、组卷、排列、修补、编写案卷号与页号,并根据文献的内容编制目录。

扫描加工是通过中高速扫描仪和专用扫描软件将整理和分检好的文献资料批量转化成图像文件,并自动实现图像的压缩存储。扫描过程不严格要求页号顺序,但是必须保证图像质量清晰,与原件基本一致,一样清晰。

OCR(Optical Character Recognition)识别是通过 OCR 软件将扫描生成的光栅图像文件自动辨识成文本字符的过程。根据需要可对文献的部分内容(如标引信息)或全文进行识别。鉴于字体、纸张状况以及识别算法等诸多因素,OCR 的识别率不可能达到 100%,因此在自动 OCR 处理之后,还需要进行人工校对和补录(简称 OCR 后处理)。OCR 后处理功能可在图文编辑工序中具体实现 OCR 后处理功能。需指出的是,OCR 识别属于可选工序,一般仅适用于较清晰的印刷体文本或较规范的表格类资料,对于手写体文献,OCR 识别率不很高。

图文编辑是建立数字化文献的核心工序,主要实现:图像处理,页号排序,建立文献标引/全文信息(人工录入或 OCR 后处理),案卷与图像挂接,目录与图像挂接,密级设置等功能。

图文质检是一个模拟查询调阅的过程,用来综合检查档案扫描和图文编辑工序的加工质量,主要包括文字内容校对、原文图像质检、图像挂接检查与密级校核等过程,以综合检查文献扫描和图文编辑工序的加工质量。

重新装订即根据被拆开的文献原件上的页号排列顺序,并且根据装订要求重新装订。装订完的文献经过质检员检查后才可以归还文献库房。

加条形码是在文献整理工序中加贴表示不同意义的条码,可以实现案卷号、文献分类等关键标引信息的自动识别。文档数字化加工的全过程,通过采用条码可以实现下述自动化处理:文献移交过程中,可利用条码自动进行文献的逐卷核对;文献盘点时,通过条码扫描枪或无线数据采集器进行条码扫描,可实现库房文献数量的精确统计,同时还可以实现库房实物文献与计算机中存储的文献信息的核对。在文献的借阅管理工作中,可应用条码进行自动化的出入库管理。

数据备份是文献数字化加工完成后必须进行的安全管理的一环。系统维护人员使用备份软件定期将加工好的电子文献(原文影像及文字信息)从服务器中转储到光盘或磁带上,以作长期备份。

2.多媒体音像文献数字化加工流程。多媒体音像文献数字化,就是将录音、录像等各种形式的多媒体原文资料通过音频、视频转换设备进行转换、识别,压缩,生成标准格式的电子文件,并编目以及建立标引信息的过程。其处理流程就是:(1)通过数码相机、数码摄像机、录音设备形成媒体原始资料;(2)通过转换设备进行转换与识别,生成标准格式文件;(3)进行编目与标准;(4)进行数据备份。

(四)文献数字化技术参数指标

文献数字化过程涉及数字化扫描技术、图文编辑、图像格式、图像存储等技术方案的采用和参数的选择。功能包括扫描加工、质量检查、去污处理以及加工后的图像文件与文献标引信息的关联等。

1.扫描技术参数的选择。扫描过程中尽量采用标准的 TWAIN、ISIS 编程接口,编写应用程序直接控制各类扫描仪,自动实现图像压缩存储。一般要求支持连续和平板两种扫描方式,支持 A3、A4 等多种幅面,支持黑白二值、灰度和彩色等多种图像格式,有盖章、照片的页面采用灰度或彩色图像处理。

扫描过程中分辨率的选择是需要根据实际业务的需要进行灵活设置,一般情况下,为了满足网络化查询利用,黑白图像采用 200dpi 就足以满足要求,彩色图像的扫描分辨率还可以低一些,具体参数可根据扫描清晰度和质量因素进行综合选择。对于特殊的利用可以采用较高的扫描分辨率来进行。

2.图像文件格式的选择。文献图像数字化后,常用图像文件格式采用以下两种组合方式:

（1）TIFF/JPEG 格式。TIFF 是一种支持多页存储的图像文件格式，它支持多种压缩算法（如 CCITT、LZW、JPEG 等），但 TIFF 本身并不是一种压缩算法。而 JPEG 既是一种单页存储的文件格式，同时又是一种标准的压缩算法。TIFF 的重要特点是支持多页存储、多种压缩方法，而且扩展性强，因此在专业图像应用领域得到了广泛的应用；JPEG 格式一般用于压缩、存储单页图片的灰度或彩色图像，不支持多页存储。在数字化档案的应用中，其主要技术指标如下：每卷档案作为一个图像文件，采用 TIFF 多页存储格式，能将任意多页的黑白二值、灰度、彩色、各种不同幅面图像压缩到一个图像文件中；黑白图像压缩采用 CCITT_GROUP4 压缩算法；灰度、彩色图像压缩采用 JPEG（YUV 4：4：4）算法；200DPI，A4 幅面，黑白二值图像压缩效果：TIFF CCITT_GROUP4，平均每页大小为 20K 左右[1]（第 3 页）。

（2）JPEG2000/ JBIG 格式。JPEG2000 作为 JPEG 升级版，是新一代灰度/彩色图像压缩国际标准，其压缩率比 JPEG 约高 30％左右，同时它还支持无损压缩、渐进传输和感兴趣区域等先进特性。JBIG 是同一标准化小组 WG1 制定的新一代的二值（黑白）图像压缩国际标准，其压缩率比 CCITT_GROUP3 高 20％～80％。JPEG2000/JBIG 格式应用有如下问题：目前大多数通用的图像应用软件还不支持或者不充分支持这些新的格式；由于该格式采用了复杂的小波变换算法，在普通 PC 机上的图像处理速度明显慢于 JPEG 和 TIFF CCITT_GROUP4，经过实测：JPEG2000 图像首次放缩有 3 秒左右的延迟时间，而相应 JPEG 图像的首次放缩几乎没有视觉延迟。

3.OCR 识别技术。OCR 技术可用于文献标引信息识别和全文信息识别。在理想的测试条件下，其主要技术指标如下：（1）识别字体：识别宋体、仿宋、楷体、黑体、魏碑、隶书、圆体、行楷、行书等近百种字体。（2）识别功能：支持印刷文稿、纯英文、中英文混排、较工整的手写文稿等多种类型。对印刷材料的识别率达 98％以上。（3）识别速度：在普通配置的计算机上印刷体汉字达 120 字/秒以上。（4）要求图像分辨率：一般不低于 300DPI。需要指出的是，OCR 在实际应用中自动识别的准确度和稳定性会有较大的折扣，这使得馆藏文献数字化工作不得不面临艰苦而繁琐的人工校对和补录工作。

4.图像文件密级定义技术。为了在较细粒度上实现对电子文献图像信息的安全访问与控制，可进行通过采用逐页定级或者页内区域进行定级，即指定某一页或页内某一区域的保密级别。页内区域定级的技术实现方式是：可采用“区域定位”技术实现页区域级别的设定。即在一页文献中的某个区域可以由这个区域左上角相对于该页的坐标、区域的长和宽来惟一定位。因此，可以把区域的位置信息和区域的级别拼成字符串来表示和存储页区域级别信息。保密级别可分为公开级、国内级、内部级、秘密级、机密级、绝密级等。定级之后，通过用户角色授权，便可实现精细的数据访问控制和权限管理[2]（P.26）。

页密级与页内区域密级授权访问方式为，若某个页内区域没有显式设定密级，则自动继承所在页面的密级；若页内区域设定了密级，而且该级别与所在页面的级别不同且高于其所在页面的保密级别，则覆盖所在页面的密级，也就是说以页内区域的密级为准。

5.媒体音像文献数字化技术参数。多媒体音像文献数字化，就是将录音、录像等各种形式的多媒体原文资料通过音频、视频转换设备进行转换、识别，压缩，生成标准格式的电子文件，并编目以及建立标引信息的过程。

资料类型	转换工具	文件存储格式	技术参数
照片	数码相机/扫描仪	JPEG	清晰度、分辨率、黑白/灰度/彩色等图像的选择
录音（磁带、文件）	数字音频压缩卡	MP3	信息失真度、压缩比选择
录像（磁带、文件）	数字视频压缩卡	MPEG	信息失真度、压缩比选择

（五）馆藏文献数字化的后期存储备份

1．数字文献的存储方式。数字化后的数字文献信息包括文献目录信息和图像原生信息两大类，为了实现网络化利用，文献的目录数据库必须采用支撑网络化系统运行的数据库，如 SQL Server2000、Oracle、Sybase、Informix 等商业化的关系型数据库管理系统[3]（第 17 页）。而数字化图像文件的存储则可以选择文件存储方式或数据库存储方式任何一种存储方式。如果选用数据库存储，则要求数据库服务器的存储容量足够大；如果选用文件存储，则应考虑存储在文件服务器上文件的存储规则和命名规则，以方便实现图像文件与目录数据库的检索。

（1）文件存储方式：在文件存储方式中，将数字文献影像以文件形式存储于文件服务器上，相关联的标引信息存入数据库中。该种存储方式降低了数据库的庞大性，提高了数据的更新效率；有利于数字文献数据的交换和标准化管理。但此种方式需通过软件实现文件和数据库的一致性备份，增加了程序编写的复杂性，同时备份也比较复杂。（2）数据库存储方式：在数据库存储方式中，将数字文献影像被直接存储到数据库的 Blob 字段中。数据库这种存储方式简便了数据的备份，安全性相对较强，但由于数据库容量较为庞大，增加了数据库的管理与维护难度，对数据更新效率有一定影响。

2．数字文献的存储设备。（1）数据存储设备：推荐采用磁盘阵列（RAID5）存储电子文献。磁盘阵列具有存取速度快、数据冗余校验、故障恢复、支持热插拔等多种先进特性。如有条件也可采用更为先进的 NAS（网络附加存储）或 SAN

(存储区域网络)存储体系结构。(2)数据备份设备:可采用可擦写光盘(MO、DVD—R、CD—R)、光盘库、磁带、磁带库等多种存储介质,推荐采用光盘/光盘库作备。

(六)馆藏文献数字化建设中应注意的问题

1.重先期规划。组织机构要先了解自己的经费、馆藏规模、信息需求、文献特色、信息政策等,根据实际情况进行科学合理地调研评估,制定出馆藏文献数字化的优先次序,明确本机构数字化建设目标。

2.注意知识产权问题。馆藏文献数字化的最终目的是实现数字化文献信息的网络化,实现资源共享。网络传输作为一种新的传播方式,已经打破了原有传播格局,信息的传播、利用超越了时空的限制,但也为信息侵权带来广阔的空间,数字化产品侵权行为时有发生,因此在数字信息环境下,更应注意加强知识产权的保护力度。根据《世界知识产权组织版权条约》、《世界知识产权表演与唱片条约》、《WTO 版权条约》以及新修订的《著作权法》等条约规定,作品的网络传输权利是属于著作权人的专有权。所以,在实际操作中,一定要注意分析馆藏文献的所有权和著作权,如果在网上传播受版权保护的文献信息及其编研成品,必须取得著作权人的授予权,未经著作权人的同意不能随便把数字化的档案信息上网传播,否则非法上载是侵权行为,要承担相应的法律责任。

3.数字信息的安全性。数字化信息与传统馆藏文献相比,具有明显的不稳定性,数字化信息的内容和位置易发生变化,它们是"动态"的、"积极"的,信息的易逝性、易变性和可操作性极大地威胁着信息安全。因此,在数字化过程中,要做到三个确保:第一,通过录入或扫描方式得到数字化信息的过程中,要确保文献原件的安全;第二,在处理和存贮数字化信息时,要确保数字化信息的内容与文献原件相吻合;第三,确保有密级文献信息内容不泄密。确保数字化信息的安全应与馆藏文献数字化同时并进。

4.推行信息标准化。由于各行业领域中馆藏文献信息种类多、数量大,各种不同数据格式的识别以及不同的信息传输方式带来了许多兼容性问题,所以必须加强信息的组织与传递的标准化、规范化,不断提高共享信息的可利用率。考虑数字信息的存储格式、压缩算法、检索方法等,提倡按照统一的技术参数指标、统一的文本格式和统一的工作流程模式进行数字化。

5.经济成本与效益评估。在馆藏文献数字化方面,建设一定规模的馆藏文献全文数据库,以现有的技术条件而论,经济成本是相当高的,如果还要维持一个带有全文数据库的镜像站、网站之类,其管理、更新维护的成本就更加高昂。目前,这两类成本主要是依靠国家财政拨款的方式予以承担,但这种状况不可能长期维持。因此,必须要遵守效益性原则,讲究数字化工作的效益。

[参 考 文 献]

[1] 陈 冠,韩为民. Adobe 电子文档管理解决方案的设计和技术特性[J]. 中国图像图形学报,2000,(5).

[2] Salton, G. Automatic Text Processing[M]. Addison-Wesley Publishing Company, 1998.

[3] 王 伟. 基于内容的图像信息检索信息的研究[J]. 中科院计算所博士论文,1999,(1)

(责任编辑 涂文迁)

Report on the Collections Digitalization

HUANG Cui

(School of Information Management, Wuhan University, Wuhan 430072, Hubei, China)

Biography: HUANG Cui(1979-), female, Doctoral candidate, School of Information Management, Wuhan University, majoring in digital information management.

Abstract: The author describes current conditions of collections digitalization in some institutions, then probes into the work mode for the collections digitalization in terms of the quality of collections, and expounded the technical project for collection digitization and its corresponding terms and conditions.

Key words: collections; digitalization; work mode