【资源·共享】

图书馆馆藏文献全文数字化建设探究

●陈洁薇 陈锦波 (广东药学院 广州 510006)

[摘 要] 文章阐述图书馆馆藏文献全文数字化建设的意义、选择标准;探索馆藏文献全文数字化的制作方法、保存模式和数字化资源的使用方式。参考文献5。

[关键词] 图书馆 文献 数字化建设
[中图法分类号] G250.7 [文献标识码] A [文章编号] 1003-7845(2012)01-0065-03

1 图书馆馆藏文献全文数字化建设的意义

1.1 数字化资源占用空间小,存储量大

大量的数字化信息存贮在磁盘或光盘上,经全文数字化的馆藏资源,与原有的文献载体相比,其占用的空间小,存储量大。

1.2 数字化信息能为用户提供便捷的信息服务

馆藏文献经过全文数字化处理,用户即使不到图书馆,只要具备互联网接口和计算机终端,就能通过网络登录图书馆网站,进入馆藏全文信息数据库界面,直接查询、浏览和下载图书馆馆藏数字化文献。

1.3 提高图书馆的服务质量

图书馆通过对馆藏文献全文数字化建设,优化图书馆网站界面和提高网络传播速度,提高用户对图书馆信息服务的满意度,全面提高图书馆的服务质量和服务水平[1]。

1.4 保护和延续馆藏文化遗产

图书馆珍贵的文化遗产,经过全文数字化处理,实现原件的全文复制,用户不须再通过翻阅原来的书籍就能够获取到这些珍贵的文献,图书馆把珍贵的文献原件作为善本资源,保存在更适宜的环境中。

2 全文数字化馆藏文献内容的选择[2]

2.1 选择图书馆拥有版权的文献或已失去版权保护的馆藏文献

馆藏文献的版权与文献出版的国家与地区有关,每个国家、地区都有知识产权保护法规,依法赋予文献作者对其文献拥有知识版权。馆藏资源的知识版权有三种形式:一是图书馆拥有版权的文献或馆藏文献已超出知识版权的保护年限,这类文献是图书馆全文数字化的主要选择对象,由这些文献建成的馆藏全文数据库,图书馆拥有其知识版权,可自由对外发布和对外提供服务;二是馆藏文献的版权归出版单位所有,这类资源在进行数字化之前必须征求文献出版单位的意见,最好是让文献出版单位开具该文献数字化授权书或许可证明书,防止以后出现不必要的知识产权纠纷;三是文献的版权归个人作者所有,这类文献在数字化之前一定要征得作者的授权或许可,否则不能作为馆藏文献数字化的选择内容。

2.2 馆藏文献的价值

评价文献的价值主要有:文献资源的唯一性、重要性;文献资源的领域所覆盖的广度和深度、实用性和准确度;文献资源特定主题领域中记录的内容具有强化历史价值;文献数字化产品潜在的经济价值等[3]。

2.3 文献载体的特性

不同载体材质的文献,对其进行数字化处理时,需要采用不同的数字化技术。如文献载体是普通纸张,则采用数字扫描技术;文献载体是光盘,则采用数字转换技术;文献载体是以视频的形式储存在磁盘,则采用视频-文字转换技术。

2.4 文献全文数字化技术条件

数字化技术条件是决定文献全文数字化是否成功的关键因素。具体包括以下内容:利用数字技术生成的数字信息是否与原始文献所包含的信息相符,数字信息的质量及其在用户终端机上的显示是否完整、清晰,图书馆现行的网络支持平台和网络环境是否与数字信息的存取要求相匹配,数字信息在图书馆的局域网和广域网的传播速度是否正常,现有的搜索引擎是否能运用于数字信息的搜索等。

2.5 文献全文数字化的制作成本

馆藏文献全文数字化制作成本组成因素很多,

例如,全文数字化文献越多,所需要的成本就越大;图像分辨率越高,采用高新技术设备,所需要的成本也越大;采用的彩色图像越多,需要的成本就越大。此外,便于全文检索且占用存储空间少的纯文本文件数字化,采用光学字符识别技术(OCR)和人工校对,需要的成本较低;带有标记的文本文件,带有各种分析数据功能和数据库管理功能的文本文件,需要的成本较高;文献单页扫描比装订在一起扫描所需要的成本低;扫描保存状态良好的文献比扫描保存状态差的文献所需要的成本要低;规模越大的项目单件数字化需要的成本越低。

3 馆藏文献全文数字化的制作与保存[4]

3.1 文档型数字资源的制作与保存

(1)字符编码。计算机中的信息包括数据信息和控制信息,数据信息又分为数值信息和非数值信息。非数值信息和控制信息包括了字母、各种控制符号、图形符号等,它们都以二进制编码方式存入计算机并得以处理,这种对字母和符号进行编码的二进制代码称为字符编码。常用的字符编码有ASCII码(美国标准信息交换码)和EBCDIC码(扩展的BCD交换码)。字符编码是将字符表示成数字形式的一种算法,它通过将字符系列转换为8位数字系列来实现,比如EBCDIC码使用8位,表示出2的8次方个字符,即256个字符。字符编码要求数字资源在运用字符编码时要标明文档所用的编码方式以解释文档的代码,因此,馆藏文献全文数字化按要求须标明数字文档所使用的编码,比如,可扩展标记语言(XML)文档就应该将其编码方式记录在可扩展标记语言的标签中;在可扩展超文本置标语言(XH-TML)文档,必须在HTTP-EQUIV属性和META元素中记录编码的方式。

(2)文档格式。文本文档是以TXT后缀名的文件,馆藏文献数字文档的创建与管理一般采用结构化格式,方便数字文档转换为可扩展超文本置标语言(Extensible Hypertext Markup Language,简写为XHTML)和可扩展标记语言(Extensible Markup Language,简写为XML)。很多情况下,将文本型数字文档保存为标准通用标记语言(Standard Generalized Markup Language,简写为SGML)或符合已经公开发布的标记符的语法规则(DTD)或可扩展标记语言架构(XML SCHEMA)的可扩展标记语言(XML)格式是最好的选择,但对相应的标记符的语法规则(DTD)或架构(SCHEMA)也是有效的。文本型数字资源一般用文件方式存储,也可以存储在数据库里。项目应该清晰解释采用标准格式对文本进行编码的目的,并以这种格式进行数据存储。馆藏文献数字化将文本型数字资源以最新的超文本标记语言(Hypertext Markup Language,简写为HTML)或可扩展超文本置标语言(Extensible Hypertext Markup Language,简写为XHTML)版本格式存储。在特殊情况下,馆藏文献数字化也可采用可移植文档格式(Portable Document Format,简写为PDF)保存文本型数字资源。但可移植文档格式(PDF)是一种专用格式,需要采用Adobe Acrobat Reader浏览软件来浏览内容。就像采用其他所有专用格式一样,这种解决办法是有风险的,应该评估采用这种方式的潜在成本,并为其探索数据迁移策略,为将来转换为开放标准格式做准备。

3.2 馆藏文献数字化中静态数字图像的制作与保存

静态数字图像分为两类,光栅图像和矢量图像。光栅图像也叫做位图、点阵图、像素图,简单的说,就是最小单位由像素构成的图,只有点的信息,缩放时会失真。光栅图像采用栅格或矩阵的形式,矩阵里的每个图像元素都有一个唯一的定位和一个可以被编辑的独立的颜色值;矢量图像是由画图程序根据一组数学算法来完成的,其记录的是点、线、面的位置和颜色信息的描述,矢量图没有直接记录点、线、面、基本图形等的信息,重现时利用看图软件就能把这些描述重绘出来。

光栅图像通常是在馆藏文献数字化的过程中产生的,没被其他应用程序进行后续处理,并以非压缩形式保存。光栅图像采用标签图像文件格式(Tagged Image File Format,简写为TIFF)、联合图像专家组(Joint Photographic Experts Group,简写为JPEG)、图像互换格式(Graphics Interchange Format,简写为GIF)、流式网络图形格式(Portable Network Graphic Format,简写为PNG)保存。使用光栅图像需要考虑两个参数:空间分辨率和颜色分辨率。空间分辨率是指图像中每英寸的像素数量,颜色分辨率是指表示颜色信息的位数,比如,用8位表示颜色,就可以表示256种颜色。此外,光栅图像质量参数的设置取决于馆藏文献的原始尺寸、内容数量和用途三个因素。光栅图像制作时应设置最高空间分辨率和颜色分辨率,设置图像的最低质量要求。在一般情况下,低廉的数码相机所产生的数字图像比较适合以JPEG/SPIFF格式保存,采用这种格式保存的图像幅度小、质量低,比较适合于Web站点小

型图像的展示。矢量图像是对多维实物性资源数字化的结果，其创建和保存应该采用开放的格式，采用可扩展标记语言（XML）语言来描述图形，也可采用Macromedia Flash专有格式，但必须考虑数据格式的迁移策略，Macromedia Flash格式要避免使用文本，以便将来开发多语种版本[5]。

3.3　馆藏文献中视频资源的制作与保存

馆藏文献视频资源的数字化制作，应该考虑视频资源的用途，设置合适的空间分辨率、颜色分辨率和帧速率，设置每个视频资源的最低质量要求。视频资源的保存格式是非压缩格式，不需要编码，不需要进行任何后续处理，直接从录像设备中获取。视频资源也可保存为动态图像专家组（Moving Pictures Experts Group，简写为MPEG）格式、图元文件（Windows Metafile，简写为WMF）格式、高级串流格式（Advanced Streaming Format，简写为ASF）或音频视频播放器格式，比如QuickTime格式。

3.4　馆藏文献中音频数字资源的制作与保存

馆藏音频资源数字化制作时，不需要经程序进行任何后续处理，直接从录像设备中获取相关信息。音频数字资源一般保存为非压缩格式，也可以使用压缩格式，如动态影像专家压缩标准音频层面3（Moving Picture Experts Group Audio Layer III，简写为MP3）、微软音频格式（Windows Media Audio，简写为WMA）和即时播音系统（Realaudio）等。

4　馆藏文献全文数字化信息资源的使用方式

（1）单机光盘版，是指馆藏文献全文数字化资源存放在光盘上，或存放到硬上，可以直接使用光盘或硬盘上的数据库，但只能单机使用。

（2）网络光盘版，是指文献数字化资源存放在光盘存储设备，或存放到硬盘存储设备，可以直接使用光盘存储设备或硬盘存储设备上的数据库，只能在网络的终端机上使用。

（3）镜像站点，是指对另一个站点内容的拷贝。镜像用于为相同信息内容提供不同的源，为下载量大的时候提供一种可靠的网络连接。镜像站点通过主服务器增加转移存储地址来实现信息的异地备份。通常一个镜像会定期访问主网站，以更新其内容。镜像站点服务器及数据存储设备由设站单位提供，将软件系统和数据库托管到服务器和磁盘阵列上。当设站单位交纳会员费后，单位内部的终端可免费使用数字化信息资源。镜像站点除了对内服务外，还可对外提供服务，为不具备建立镜像站点的用户提供服务。

馆藏文献全文数字化建设是节省图书馆的馆藏空间，保护珍贵的文化遗产，拓展图书馆的服务模式和提升图书馆服务质量的有效途径；在当前图书馆人力资源和经费紧张的条件下，馆藏文献全文数字化内容的选择、制作和保存等内容对数字化项目的具体操作具有积极的指导意义。

参　考　文　献

[1] 陈金刻. 图书馆数字化信息资源建设对高校科研的作用[J]. 福建图书馆理论与实践，2009(1)：7－8.

[2] 臧国全. 图书馆信息资源数字化内容选择原则研究[J]. 图书情报知识，2006(1)：20－24.

[3] 王　军. 图书馆信息资源数字化项目实施原则解析[J]. 图书馆理论与实践，2006(6)：3－6.

[4] 肖希明. 图书馆学研究进展[M]. 武汉：武汉大学出版社，2007：547－612.

[5] 臧国全，庞桂娟，姜　燕. 图书馆信息资源数字化项目实施标准框架解析[J]. 图书馆理论与实践，2006(4)：5－10.

[作者简介] 陈洁薇，副研究馆员；陈锦波，馆员，现在广东药学院图书馆工作。

[收稿日期] 2011－09－01　　（宋小华　编发）

The Contribution of Full－text Digitalization of Library Collections

Chen Jiewei　Chen Jinbo

(Guangdong Pharmaceutical College, GuangZhou, Guangdong 510006, China)

Abstract　This paper states the significance of the full－text digitalization of library collections and its selection criteria and then explores its making methods, preservation models and usage modes of digital resources. 5 refs.

Keywords　Library. Literature. Digital Construction.