

浅析古籍数字化建设的组织模式 ——以哈佛燕京图书馆中文古籍数字化项目为例

□ 吴茗 / 国家图书馆 北京 100081

摘要: 古籍数字化规范的建设是一个复杂的过程, 同时也是一个不断变革的领域。针对目前古籍数字化存在的问题, 文章以哈佛燕京图书馆中文古籍数字化项目为研究实例, 通过其数字化加工方式、数字化过程、参数选择等方面的具体实践, 重点对图像和元数据的数字化建设和组织进行了分析和探讨, 为古籍数字化标准规范的建设提供借鉴和参考。

关键词: 古籍数字化, 数字化实践, 组织模式, 哈佛燕京图书馆

DOI: 10.3772/j.issn.1673-2286.2012.03.010

1 前言

古籍数字化就是从利用和保护古籍的目的出发, 采用计算机技术, 将常见的语言文字或图形符号转化为能被计算机识别的数字符号, 从而制成古籍文献书目数据库和古籍全文数据库, 用以揭示古籍文献信息资源的一项系统工作^[1]。通过数字化的工作可以保留古籍的原样, 与古籍原件具有“异曲同工”的效果, 作为古籍原件的替代品完全可以投入使用, 同时, 使用它们可以在更大范围、更便捷、更有效地使古籍文献为人们所利用。

古籍文献卷帙浩繁, 情况错综复杂且收藏分散, 目前的古籍数字化工作参与的机构有很多, 古籍数字化成果斐然, 但参与者的素质参差不齐, 又缺乏统筹兼顾、统一管

理, 绝大多数开发者以独立开发为主, 格式不统一, 互不兼容, 重复投入, 低水准重复建设, 既造成资源的重复开发, 又存在许多信息孤岛。虽然一些文献古籍已经数字化, 但图像格式、元数据格式、存储结构等各不相同, 不能在有效的范围内进行整体检索, 更不能实现资源共享。

因此, 如何在新的信息技术高速发展的背景之下, 保护好、利用好古籍文献并做到保护性的利用, 使之尽可能以其完备的形态呈现于读者的面前, 是目前面临的亟待解决的难题。有鉴于此, 本文将结合哈佛燕京图书馆中文古籍数字化项目的实践, 对其建设方式和组织模式做一些探索性的研究, 以期对古籍数字化的标准规范的建立提供有益的借鉴。

2 哈佛燕京图书馆中文古籍数字化项目介绍

哈佛大学哈佛燕京图书馆的馆藏中文善本古籍特藏以其质量高、数量大著称。该项目对哈佛燕京图书馆藏的所有中文善本古籍和齐如山专藏合计4210种51889卷进行数字化拍照。这是中国国家图书馆在国际合作史上规模最大的文献数字化项目。

中国国家图书馆负责提供资金、技术支持和数据质量控制。哈佛大学图书馆提供技术设备, 承担中文善本古籍数字化、元数据制作和数据传递工作, 并在双方共有的基础上使用数字化成果。项目成果将由双方各自保留一份完整的数字化图像文档, 分别用于各自网站免费发布和其他学术研究及文献保存之应用。

为保证项目数据得以顺利接收和发布,项目组成员在项目实施前后,投入大量精力对项目细节进行研究和探讨,进行技术攻关。按照工作计划,经过质量检查和元数据标引后,通过国家图书馆网站发布。

3 图像的组织

3.1 图像数字化的方式

古籍全文数字化就是以古籍的内容为揭示对象,对古籍资源的全文进行数字化处理以便为读者提供全文阅读、全文检索或智能分析服务。根据全文数字化的技术处理情况,可以将其分为完全图像方式、全文本方式、图文对照方式^[2]。

(1) **完全图像方式**:是一种最简单的数字化方法,就是将古籍书页扫描形成图像存储在光盘或硬盘等介质上,为读者提供文字信息服务。图像版数字化古籍的优势是技术实现容易,运行成本低廉,可以保存古籍原貌;其缺陷是占据空间大,不能检索。

(2) **全文本方式**:是指将古籍书页转换成文本字符的形式,存储在光盘或硬盘等介质上,并附加全文检索和链接系统等功能,为读者提供全文阅读或全文检索服务。文字版数字化古籍的优势是储存空间小,具备全文检索功能,可以支持学术研究。全文检索版数字化古籍的缺陷是文字输入难度较大,开发成本高,不能保持古籍原貌。

(3) **图文对照方式**:该方式是以上两种方式的综合体。它不但将古籍以图像和文本方式存储,还对古籍中的每一个汉字与它在原书

中的图像建立链接关系,用户在进行全文检索的同时,又可以方便地查看图像资料。显然,图文对照方式数字化古籍是最理想的数字化古籍开发模式。

数字化方式的选择应根据项目的具体情况,要综合考虑时间、成本、技术成熟度、适用范围等各种因素,不必追求一次到位。哈佛燕京古籍数字化项目旨在方便海内外学人便捷地利用这些资料进行研究,同时以数字化形式保存这些中华古籍精品,促进海外古籍的数字化回归。在这种特定前提下,又综合考虑了工期、成本等因素,选择以完全图像方式进行数字化加工,同时确定合适的可进行二次开发的参数,以便将来进行增值性开发。

3.2 图像数字化过程

古籍的数字化加工过程与其他文献的数字化加工过程基本一致,大体可分为三个阶段,即图像扫描或拍照、图像处理和图像的发布。哈佛燕京图书馆的古籍数字化项目同样遵循这样的加工过程,首先在图像数字化的第一阶段,根据资源原件的自身特征,参照相关的数字化资源建设标准,确定较为合理的、具有可操作性的技术指标。因为在本项目中拟数字化的书籍为线装古籍,为了在保护古籍的基础上再现古籍的原貌,采用拍照的方式,同时在保证不影响品质的前提下,为了降低数字化成本,采取一张数字图像包括两页文本的方式,并有针对性地确定各项技术指标,对图像的格式、分辨率、清晰度、尺寸、色彩等方面的标准进行了深入的研究,充分体现了古籍数

字化的“保真原则”,并具有一定的前瞻性。第二阶段是图像处理,该项目组成员进行了多次专题研究,在图像画面的选取、图像发布的格式、水印的嵌入、图像文件压缩与优化、图像文件传输与存贮、数字资源长期保存等方面,进行了反复的测试、比较,从而选取最佳的技术指标。第三阶段是图像的发布,本项目采用PDF格式打包发布,并根据浏览效果,对图像作进一步的调整。

3.3 图像资源的主要参数

对图像资源进行数字化描述需要的参数很多,主要包括成像格式、分辨率、图像色彩等。

3.3.1 成像格式

在古籍文献数字化过程中,根据使用数字图像的不同目的和被成像资源的不同类型及其特征,合理选择恰当的成像格式是非常重要的。不恰当的数字化格式选择将导致生成的数字化复制品损失原始文件的信息,或使原始文件从外观、视觉感受上蒙受巨大损失。

就古籍文献来说,选用的图像格式根据不同的应用级别,可以分为长期保存级和发布服务级两种,目前长期保存级一般都采用不压缩或无损压缩TIFF5/TIFF6、JPEG2000等格式,发布服务级一般采用JPG或封装到PDF。

(1) **TIFF (Tagged Image File Format)**,即标记图像文件格式,是一种通用的图像格式,在图书馆界应用广泛,目前大多数图书馆数字化文件生成的图像都采用TIFF文件格式,其优点是存

储的图像质量高,独立于软硬件平台,支持多种色彩模式和多通道等。TIFF格式具备一定的应用基础,但由于其本身的局限性,经其处理的数字化图像只适用于长期保存,而不便于利用,因而通用性受到一定限制^[3]。而TIFF提供给用户使用的时候,通常是转化成其他的文件格式。

(2) JPEG2000则是档案图像文件格式中的新贵,在国内图书馆界的应用较少,相比TIFF而言, JPEG2000由于采用无损压缩的方式,文件的体积能够成倍减少,适用于长久保存,还能够采用有损压缩成为体积更小、更灵活的图像文件提供利用。JPEG2000因其所占空间小、灵活压缩等特点,作为一种新兴的数字图像文件格式,有着TIFF无法比拟的优势,大有取代TIFF的趋势。

哈佛燕京采用无损压缩的JPEG2000格式来保存图像。MIME类型是image/jp2。发布时采用PDF封装。

3.3.2 分辨率的确定

分辨率越高,图像的质量越高,但对数字资源的存储与传输也是一个极大的挑战。由于分辨率的选择直接关系到数字化资源的品质与其占用空间和传输速度等,所以在分辨率的选择上,不仅要充分考虑资源的自身特点,如原书尺寸、版本、年代等,还要充分考虑经济成本和时间成本、存储成本、转换效率等综合因素。分辨率选择还要有一定的前瞻性,在考虑当前需求的同时,尽量考虑未来的需求,以便提高数字化产品在将来使用的可能性。这意味着我们在着手

数字化工程实践时,需要综合考虑这些因素;在保证原始文献最小字符或最有意义信息清晰可读的前提下,选择适宜的分辨率格式水平来进行数字化转换。一般来说,用于长期保存级的图像一般选用300-600dpi。

该项目在综合考虑各方面因素,尤其是考虑使用目的后,认为追求过高的分辨率在本项目中并无太大的意义,最终将图像输出分辨率定为300dpi。

3.3.3 色彩管理

在文献数字化过程中应该根据古籍特点选择适当的图像色彩,对于仅用于文字浏览或OCR的图像,可以采用灰度或黑白二值图像,既可以节约存储空间,又能提高数字化效率。而由于古籍的特殊性,为了能够充分体现古籍数字化的保真原则,再现古籍原貌,同时出于对古籍保护原则的考虑,一般采用24位真彩图像。在哈佛燕京的项目中,为了保证图像输入设备、输出设备的色彩匹配,实现不同显示器的色彩显示的一致性,本项目对图像的色彩管理有明确的要求。考虑到数码影像在各种处理系统之间的匹配问题,采用sRGB色彩空间,消除了不同显示系统在色彩还原上的原有差异,提高了它们与打印输出设备间的色彩匹配,保证了色彩的一致性。除此之外,还在每册的封面图像左侧添加有24个色块的DC色卡,用以鉴别或鉴定色彩的还原度和偏色度。

4 元数据的组织

哈佛燕京图书馆中文古籍数

字化项目采用美国国会图书馆推荐的METS元数据编码和传输规范作为元数据的封装标准。

METS (Metadata Encoding and Transmission Standard) 是美国数字图书馆联盟DLF (Digital Library Federation) 开发的、用来将一个数字图书馆中的数字对象相关的描述性元数据、管理性元数据和结构性元数据进行编码的一个标准,采用W3C (World Wide Web Consortium) 的XML Schema语言表达。具体来说就是将有关数字化资源的元数据进行打包,包括所有描述性的、管理性的、结构化的、权限及其他可用于数字化资源检索、保存和服务的元数据。它可用于在不同的数据库中的数字对象交换数据,也可以用作对数字对象本地化的输入、检索和显示的封装机制^[4]。

随着数字文献的长期保存问题逐渐成为数字图书馆研究领域中的热点问题, METS规范也逐渐被越来越多的信息服务机构采用,开始应用它作为编码和传输元数据的标准,并且在相关的系统中开始应用。

目前,国外已经有很多知名大学图书馆和机构开始在实际工作中采用METS作为数字资源保存系统中的一个重要实现,如美国国会图书馆的Audio-Visual Prototyping Project、Virginia大学和Cornell大学合作开发的Fedora Project、RLG (Research Library Group) 的Cultural Materials Service项目等。

哈佛燕京图书馆古籍数字化元数据包括三个部分:描述元数据、结构元数据、技术元数据。

描述元数据:为每一种书提

供一条描述性书目记录。它记录了METS所有子对象的描述性元数据,该元数据可以是指向到METS文件外部的描述性元数据或者包含内部嵌入的描述性元数据。

结构元数据:采用METS规范对每册书的结构进行描述,展现文本的物理排列。它是METS文件的中心部分,概括出了数字化对象的层次结构,链接了内容文件的结构元素和属于每个元素的元数据。

技术元数据:至少应包括MIME类型、MD5校验和、图像宽度、图像高度、采样频率、位深、文件大小等信息。这部分提供的是文件如何创建和存储、知识产权、图书馆数字化对象原始来源的数据、图书馆数字化对象文件出处的信息。同样,管理性元数据可在METS文件内部和外部编码。

5 数据封装方式

哈佛燕京项目使用bagIt规范对元数据和对象数据进行封装。

bagIt是一种分层级的支持数字内容的硬盘存储和网络传输的文件打包格式。一个“bag”由“有效载荷数据”和“标签”组成,即是用来保证“bag”存储和传输的元数据。

一个“bag”有一个基本目录,包括一系列顶级标签文件“tags”和一个名为“data”的子目录,内涵有效负载数据,“data”包含任意的文件层级,一般将图像数据封装在data的文件目录下。

具体来说,一个“bag”的目录结构如下:

```
<base directory>/
| manifest-<algorithm>.txt
| bagit.txt
```

| [optional additional tag files]

\--- data/

| [optional file hierarchy]

6 结语

古籍数字化工作是数字图书馆建设的重要组成部分,其组织、加工和发布应依循数字图书馆的理想发展模式,应遵循与现代图书统一的标准规范,并在统一标准规范的框架下进行细化,并使古籍数字化具备开放性、层次性和结构化的特征^[5]。然而,随着数字化技术的发展与完善,这些都不是一成不变的,只有在实践中不断地摸索和总结,才可能得到较为理想的数字化成果。

参考文献

- [1] 毛建军. 古籍数字化的概念与内涵[J]. 图书馆理论与实践, 2007(4): 82-84.
- [2] 何贤英, 李秀娟. 中文古籍数字化浅析[J]. 科技情报开发与经济, 2008(18): 13-14.
- [3] 徐曼. 馆藏资源数字化过程中图像格式的选择[J]. 图书情报知识, 2005(3).
- [4] 董丽, 吴开华, 姜爱蓉, 等. METS元数据编码规范及其应用研究[J]. 现代图书情报技术, 2004(5).
- [5] 陈力. 中文古籍数字化的再思考[J]. 国家图书馆学报, 2006(2): 42-49.

作者简介

吴茗 (1975-), 女, 国家图书馆数字资源部。E-mail: wmsxy@sina.com.cn

Analysis on the Organization Patterns of the Digitization of Ancient Books – Taking "The Harvard-Yenching Library Digitization Project" for Example

Wu Ming / National Library of China, Beijing, 100081

Abstract: The construction of the standard about ancient books digitization is a complex and changing field. Taking the "Harvard-Yenching Library Digitization Project" for example, this paper analyzed how to build and organize the images and metadata through the practice of the digital processing, methods, parameter selection, etc. This can be used for reference to the construction of digital standard about ancient books.

Keywords: Digitization of ancient books, Digitization practice, Organization pattern, Harvard-Yenching Library

(收稿日期: 2011-08-25)