

古籍数字化基础工作述略及思考

张伟丽

近二十年来，古籍数字化一直受到公共图书馆、科研院所、古籍整理研究者的高度关注，涉及地域、范围日趋广泛，相关从业人员也越来越多。经过多年的发展及学者们的共同努力，中国古籍数字化工作取得了突破性的进展，廓清了一些基本概念，完成了一些重要古籍数字化项目。目前较为全面的古籍数字化的概念为“从利用和保护古籍的目的出发，采用计算机技术，将古籍内容复制或转移到其他载体，以达到利用和长期保护的目的”¹。国家图书馆作为国内较早开展古籍数字化的公共图书馆之一，不仅较早提出了“古籍数字化”概念，并取得了丰硕成果，完成了“甲骨影像”、“数字方志”、“敦煌写卷”等多个数字化项目的工作。

2011 年初，为了有针对性的完成古籍数字化工作，国家图书馆古籍馆设立了数字化工作室，以项目为中心，对馆藏古籍进行数字化处理。五年来，数字化工作室从无到有，从草创初期到形成了一套规章制度，从最基础的数字化工作做起，不断完善工作流程，使之规范化、制度化。截至 2013 年 8 月底，古籍馆数字化工作室已经完成馆内外 150 余个项目。藉由国图宏富馆藏的天然优势，工作室的数字化对象级别高、类别多，包括善本古籍（善本、手稿、地方志）、法帖、舆图、外文善本、民文古籍、新善本、老照片等。我们在完成数字化项目的同时也积累了一些实务经验，攒聚成文，以求正于方家并与诸同仁分享。

作为数字化对象，古籍有其特殊性，古籍数字化工作更是数字化工作中比较独特的一种，相关学者纷纷指出从业人员既要有古籍文献知识，也要掌握一定的数字化技术和标准，这样复合型人才比较缺乏，培养也需要一个过程，就基层工作实践经验来看，目前“最有力的解决办法就是古籍整理专业人员与计算机技术人员的通力合作”²，故工作室的日常工作主要由这两类人员组成。古籍馆在馆内抽调人员加入经典文化推广组的数字化工作室并开展数字化知识培训，成员需要掌握格式转换、拼接和裁切等图像处理方法。但扫描、复制工作、调试机器、

¹ 王立清《古籍数字化研究》，北京：中华书局，2013 年，70 页。

² 朱锁玲《我国古籍数字化进展与研究述评》，《图书馆理论与实践》2009 年第 9 期，18 页。

色彩管理等则需采购外包服务，由两类人员共同合作完成数字化项目。特别需要指出的是，这两类人员都要对藏品保有高度的责任心和安全意识。

整体来看，古籍馆数字化工作主要包括三个步骤：前期数据采集，数据管理与交付，后期数据（库）加工。目前国家图书馆古籍馆数字化工作室主要负责在前期数据采集和数据管理上。古籍馆洽谈好项目后，经由馆报批和签署协议等流程，将确定的目录提供给工作室，由典阅组负责按照项目书单提出古籍交与数字化工作室加工，下面就具体情况详述之。

一、古籍数字化工作设备及流程

一个完整的古籍数字化工作室包括如下设备：数字化设备、衬纸工作台、数据质检工作站、图像后期处理工作站、古籍存放保险柜、书车、光盘和硬盘柜。目前工作室古籍数字化主要有两种途径：拍照和扫描。工作室数字化设备主要采用法国的 I2S、意大利 METIS 和德国赛数三种大型专业书刊扫描设备。拍照设备若干和翻拍架等辅助设备，但只在最开始的一段时间使用拍照，目前主要采用扫描方式。数字化工作流程如下：

首先，经过库房人员按照项目要求筛选、提取古籍，即前整理。前整理完成后，数字化工作室将藏品提出，对照工作单核对书目。这一过程看似简单，但关系重大，扫描数据量和扫描种类累积，数据重复的风险增大，若不核对，很可能把已经加工过的藏品流入下一环节。



图 1 不同规格的衬纸



图 2 衬纸前后对比（左侧衬纸，右侧未衬纸）

其次，检查书页是否透字，如果扫描点数较高，则需要衬纸，衬纸的大小不一，根据古籍的大小要分配相应的衬纸。衬纸规格有 28×9.5cm、32×13cm、

35×15.5cm、60×24cm 不等（图 1），常用的为 32×13cm。衬纸后扫描图像效果会更清晰（图 2）。

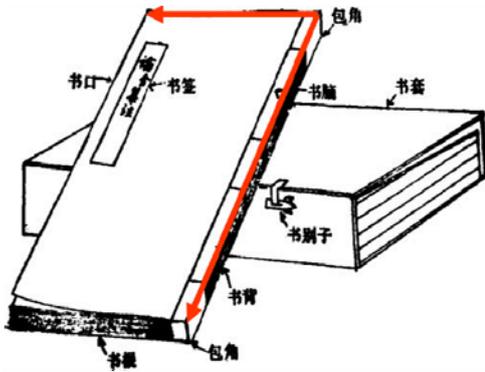


图 3 书籍大小测量方法



图 4 中缝夹字

第三，量好书籍大小尺寸（图 3），记录数据。检查古籍是否有夹字问题（图 4），在藏品安全的前提下，经藏品所在科组的组长认可后，可会请修复组人员拆书线，扫描完成后再缝合。



图 5 古籍色卡放置标准

第四，上述工作全部完成后才能进入图像扫描阶段。扫描时需加放色标和灰度卡（图 5）。加放色标后成像有色彩对比依据，能够更好地还原古籍原始面貌。

第五，扫描完成以后，填写扫描记录单，内容包括书号、书名、项目名称、古籍尺寸、扫描日期、扫描人等。再转入质检环节，将数字化原件与扫描数据

核对，这一环节至关重要，容后详谈。质检完成后，质检员填写质检表，将合格的图像存入移动硬盘中，存储格式为 TIFF。最后将数字化完成的藏品归库。

此为数据采集过程，全部完成后进入到数据管理阶段。每个工作日都有专人登记扫描记录单和质检记录单，录入数据总表中，总表中添加“存储地”一项，即标明该种古籍数字化图像存在哪块硬盘中，便于后期数据交接和查找。由专人定期整理数据总表，并负责数据交接。数据交接时要填写《数据交付单》，根据项目性质和合同要求，提供相应分辨率的 JPG 或者 TIFF 格式图像。数据交付后，

定期接受反馈，如数据出现问题，则需要核对原件或原件缩微胶片。工作室扫描数据也可供查看、研究之用，但拷贝数据则需要填写《数据流出单》，内容包括使用人、用途、数据量大小等内容，获领导批示方可流出。

经过五年的积累，数字化工作室的数据管理已经初步规范化、制度化，很好地保证了藏品和数据的安全。

目前数字化工作主要以项目为中心，以项目为中心的优点是具有针对性，数据集中，便于管理，不易间断，缺点是缺乏系统性。将来如何有计划有步骤地将馆藏数字化，如何利用这些数据资源，这些都是管理者、从业者需要思考的问题。

二、古籍数字化质检工作详解

在保证藏品、数据安全的前提下，质检是数字化工作中比较重要的一步，前承扫描环节、下接数据存储管理。在数字化工作室成立之初，整理数字化图像时，发现扫描（拍照）中难免出现漏扫、颜色异常等情况，为保证工作效率，避免重复提书，便强化了质检工作，要求当日加工数据当日质检完成，并细化质检内容，从而大大提高了图像采集的准确性。目前质检主要采用 ACDsee 软件，需要质检员一叶一叶地翻检古籍原件，核对扫描好的数据，一旦发现问题，交还扫描人员重扫，合格后将补扫后的图像替代原图像。笔者从事质检工作近两年，发现图像采集常见问题主要有如下几个：

- 1.未放色标。色标对于扫描色彩对比度较强的扫描对象非常重要，如法帖，地图等。这样能够集中显示扫描的色差及色差度。法帖对于扫描要求比较高，特别是乌金拓，扫描时极易反光，整个字体呈白色。还出现过白色字体中空的现象，一方面需要调白平衡，另一方面亦需检查机器的玻璃盖，如果尘土太多也会导致字体中空。

- 2.图像发虚或变形。这是扫描时速度太快或者书籍移动导致。

- 3.颜色异常。即图像发红的现象，需要调白平衡。

- 4.漏叶。分两种情况。一种是扫描时漏扫某些书页。这种情况，如果没有特殊要求，漏扫只需补扫再存入相应文件夹即可。另一种是书籍原有缺叶。如果质检登记未注明漏叶原因，就无法确定是古籍原缺还是扫描漏扫，只能核对原书或缩微胶片，费时费力。有两个办法解决，一是质检时登记原书是否有缺叶信息，

另一种是借鉴其他图书馆的古籍数字化相关经验：“对于有缺页的古籍，按照所缺页数，插入相应的空白页，以便到其他收藏单位找到相同版本，采集所缺的页数补上。”¹在原缺位置插一张空白页以做标识，数据交付的时候对方亦会一目了然，不会再怀疑是漏扫。

5.特殊要求之命名。有些项目要求特殊命名，如“中华字书项目”要求图像封页、序文命名成“ZHYS+乙方提供编号+000×+A00000×”，正文命名成“ZHYS+乙方提供编号+000×+P00000×”，除此之外的内容要命名成“ZHYS+乙方提供编号+000×+Z00000×”面对这样复杂的情况，需要质检人员以高度的责任心和耐心应对。

6.异形藏品的处理。对于异形藏品，特别是超大藏品，如舆图，需要单拍扫完后后期拼图。

7.浮签的处理。有的藏品叶上贴有浮签，多者至每半叶五至六张浮签，且位置不一，只能一一掀开扫描，浮签透字者还需衬纸。遇到这样的情况，质检人员要倍加小心，不致漏扫浮签，更要保证浮签不掉、不错位。

最后重点说一下补扫数据存储（回插）问题。图像回插工作需要极大地耐心和细心，如果回插图像数量过多，一个一个核对非常费时。在做“思溪藏项目”时，我们就遇到了这种问题。如何提高大量回插图像的工作效率且保证不出问题？我们的诀窍是将不合格图像的叶码标注在夹条上，请扫描人员扫完即按照夹条上的标注改名，再统一拷入相应文件夹，替代原图，这样质检的工作效率大大提高了。如果是漏扫回插，可以命名为“××.1”或者“××—1”，“××.1”表示在图××之前，“××页—1”表示图××之后，补扫后的图像依此命名，质检人员一目了然，依次回插入对应文件夹后，再统一重命名即可。

完成质检之后要填写质检表，内容主要包括书号、书名、拍数、是否合格、问题叶码、质检人和质检日期。鉴于数字化对象情况较复杂，应进一步细化质检表，如增加“是否原缺”、“原缺页码”等内容。

三、古籍数字化工作相关问题思考

无论哪一单位或部门的古籍数字化工作，首先面临的的就是数据安全和存储问

¹ 孙建春《古籍图像采集经验谈》，《科技情报开发与经济》2012年第14期，25页。

题，大量的数据存储在手机硬盘中，如何保证最大限度地保存图像，延长硬盘保存图像时间，是否需要适时备份数据，这些都是亟待解决的问题。

其次，优化相关从业人员知识结构。目前数字化工作室的重点主要放在图像采集上，将采集好的图像信息交给数字资源部门编目和整理。前期工作人员并不了解后续的工序，是否需要加以培训，使之对数字化工作有宏观的了解，特别是后期数据库的建设，能够使前期图像采集工作更好地配合后期加工，更好地实现最终目标。特色数据库的建设，更需要从业人员有复合知识结构，比如中国台湾地区基于 GIS 的古籍数字化地图（Chinese Civilization in Time and Space）“它以谭其骧先生《中国历史地图册》为主……辅以各类其他历史地图和遥感影像等基础图像资源……用户可以古代地图为入口了解到当时的地名设置、区域大小、重要道路以及其他重要文化信息。”¹这种特色数据库就需要前期图像采集与后期制作紧密配合才能实现最终意图。目前古籍数字化数据库常见的主要有文本数据库和信息数据库，由于我们是全图像采集，是否能在文本数据库中突出图像优势，突出版本特征，也是从业者需要考虑的问题。

第三，更好地为读者服务。古籍具有不可再生性和无可替代性，但是它最大的价值不是束之高阁，而是能更好地为读者服务，但是藏用冲突一直是不可避免的矛盾，“古籍数字化是缓解日益严重的古籍使用与保存这一特定矛盾的有效之道”²。全图像扫描比起缩微胶卷有明显的优势，能够最大限度地体现藏品细节，对版本研究多有裨益。因此，适时针对数字化的古籍编目和编制索引，特别是影像的索引，可以给读者提供更为专业的服务。毕竟“图书馆在数字化与古籍整理的过程中扮演着枢纽的地位，不仅因为图书馆是保存者，还肩负着规划、执行和协调的角色”³。

第四，数字化产品版权问题。学者们关心这个问题由来已久，多次讨论并提出建议。2013年8月28日复旦大学举办的“第十二届国际小说戏曲暨文献数字化研讨会”上专家再次提出数字化版权问题，数字化产品一旦形成商品之后，相关版权问题以及涉及到的政策法规是否能够进一步明确，也是古籍数字化研究者应考虑的问题。

¹ 范佳《“数字人文”内涵与古籍数字化的深度开发》，《图书馆研究》2013年第3期，19页。

² 张雪梅《古籍数字化与文献信息资源共享》，《天津工业大学学报》第21卷第3期，23页。

³ 庄舒蓉《数字化图书馆建设中的古籍数字化探讨》《武汉职业技术学院》第11卷第5期，89页。

结 语

就目前的研究成果来看,与古籍数字化宏观理论发展相比,数字化基础工作研究相对滞后。基层工作人员要正确认识数字化基础工作,并能从中总结出经验,提高工作效率,为古籍数字化后期数据管理和提升产品发展空间提供有力的支持。同时能够在工作中更好地运用既有理论,在理论指导下工作,为形成新的理论贡献力量,实现理论与实践的良性循环,从而为建设数字化图书馆贡献力量。

参考文献

- [1]王立清 《古籍数字化研究》中华书局 2013 年。
- [2]朱锁玲 《我国古籍数字化进展与研究述评》《图书馆理论与实践》2009 年第 9 期。
- [3]孙建春 《古籍图像采集经验谈》《科技情报开发与经济》2012 年 第 22 卷第 14 期。
- [4]范佳 《“数字人文”内涵与古籍数字化的深度开发》《图书馆研究》2013 年第 3 期。
- [5]张雪梅 《古籍数字化与文献信息资源共享》《天津工业大学学报》第 21 卷第 3 期。
- [6]庄舒蓉 《数字化图书馆建设中的古籍数字化探讨》《武汉职业技术学院》第 11 卷第 5 期。