珍贵古籍数字化概说

——以国家图书馆古籍馆经典文化推广组数字化实践为中心

赵大莹

国家图书馆古籍馆经典文化推广组（以下简称推广组）于 2011 年 3 月成立。有员工 15 人，年龄段是 1979 至 1989 年，可以说是全员 80 后。与古籍馆其他以藏品立组不同，经典文化推广组是以功能立组。目的是在古籍馆多年来以古籍特藏为中心的大众文化服务基础上，通过专门的业务科组来继续拓展渠道，提供高水平高质量的文化信息和相关文化产品，从而使古籍馆成为传统文化保存、发布、推广的重要文化单位。因此推广组的工作宗旨是提供经典文化普及的高端服务，以文化出版物、学术讲座、专题展览、古籍在线资源等为实现形式。在工作中，建设熟悉掌握经典文化精髓，具备经典文化整理、研究、再阐释能力、富有文化推广创新意识的高水平人才队伍。其基本业务工作可以分为两类：文献数字化和文化推广活动。

国内外对于古籍文献数字化的研究成果颇多，从古籍数字化的国际合作管理、古籍字库、元数据、数据库结构、知识挖掘等技术方面，皆有较为深入的讨论；对古籍文献扫描中的色彩管理等问题，也开始引起工作者的注意[1]。随着全国图书馆界古籍数字化项目的开展，古籍数字化的图像采集作业规范化操作等实践中产生的各种问题，受到越来越多的承担数字化工作的单位的重视。对专门从事扫描的工作人员、对藏品管理人员、修复人员等直接参与古籍数字化工作的人来说，这些实践中的细节，能够帮助他们较为快速的上手操作，少走弯路。因此，本文即以推广组成立五年以来的工作实践为主，简要概述工作中需要注意的相关内容。

[1] 兹举数例：关于中华古籍的数字化国际合作研究，如龙伟、朱云《中华古籍数字化国际合作及实践探讨》，《图书馆工作与研究》，32—35 页。李荣艳、李云龙、梁蕙玮《国际中华古籍数字资源整合研究及思考》，《图书馆学研究》2014 年第 6 期，50—53，34 页。张文亮、党梦娇《古籍数字化国际合作问题探析》，《图书馆学刊》2015 年第 3 期，1—4 页。对国内古籍数字化研究的进展分析，有常继红、魏晓峰《国内古籍数字化研究进展与启示》，《河北科技图苑》2014 年第 3 期，82—85 页。古籍数字化标准问题的研究，如葛怀东《论古籍数字化标准体系建设》，《图书馆学刊》2013 年第 1 期，47—49 页。王海花、王睿《西北地区古籍数字化现状及标准化建设研究》《农业图书情报学刊》2014 年第 1 期，33—36 页。古籍知识挖掘方面，如史睿《古籍文献索引与知识发现》，《2005 年中国索引学会年会暨学术研讨会论文集》，2005 年，2—9 页。色彩管理方面，如肖禹、王昭《论色彩管理在古籍数字化中的应用》，《图书馆学刊》2013 年第 9 期，20—22 页。

一、古籍数字化的主要内容

现存中文古籍数量巨大，保存分散，读者利用困难，因此通过现代技术手段将古籍的内容转移到其他载体，可达到对古籍长期保护与有效利用目的。这被称为古籍的再生性保护，也是目前我们文献数字化工作的主要内容。数字化工作涉及加工对象、工具、著录标准和操作者，以及协调管理的机构。古籍的数字化加工过程具体可分成两个步骤：

第一步，古籍影像数据采集。指利用现代信息技术，将以抄本、刻本、活字本、套印本等方式呈现的古代文献，转化为影像数据的形式。根据数据格式的不同，还可以分成图像、文本以及图像加文本三类；根据内容的完整程度，可以分成部分数据和全文数据两类。

第二步，古籍书目和影像的数字化加工。即对古籍文献进行加工、处理， 制成古籍文献书目数据库和古籍全文数据库，用以揭示古籍文献的内容信息，实现便捷准确的检索，满足知识发现的需要。

图 1 国家图书馆特色资源数据库

基于古籍馆所保存的资源类型，珍贵古籍数字化可以制作多种类型的数据库。

国家图书馆已经建成的特色数据库，包括：甲骨、敦煌遗书、少数民族文字古籍（西夏文文献）、老照片、汉文古籍（含方志）、金石拓片、年画、舆图等（参见图 1）。

其中，“甲骨世界”数据库内，既包括甲骨实物和甲骨拓片的图像，也有对应释文。该库收录甲骨目录 2964 条，影像 5932 幅；甲骨拓片目录 2975 条，影像 3177 幅。数据库资源的著录包括出土地点、时期、原骨属性、原骨尺寸、来源、释文情况、著录情况、旧藏编号、卜辞内容类别等，用户可依据著录设置的检索途径进行全文检索。该数据库还具有工具库链接功能，如《甲骨文合集》来源表及释文部分、《甲骨文字典》、《金文字典》等，以便读者参考[1]。

可以说，想要把如此丰富、好用的数据资源提供给读者，需要大量后台工作。包括数据的组织与管理，如书目数据库的规范，需要统一的机读目录格式，国家图书馆用的是 CNMARC；统一的古籍著录原则，国家图书馆现在使用的是国家标准 1987 年国家标准的《古籍著录规则》，并参考 2008 年的修订版；统一的古籍分类法标准，国家图书馆目前正在整合善本和普通古籍书目数据，其中的一个问题，就是善本的目录著录分类采用四部分类法，而普通古籍则采用《中国古籍总目》中的经史子集丛的五类分法；统一的主题标引依据，国家图书馆主要使用《中国分类主题词表》；统一的字库标准，例如在多种异体字存在下，选用哪个作为正字，可以对应那些异体字，等等。这些工作，非一组之力能够完成，通常需要几个科组乃至与馆外机构合作，才能继续开展。

不仅如此，数字化工作还包括藏品资源衍生品，包括各种整理出版的成果和研究专著、论文等的转化。例如“中华再造善本工程”中影印出版的古籍，古籍馆学人论著（包括《文津学志》、《文津流觞》）、文献整理成果（如《西夏文献中的汉文文献释录》）、文献保护会议论文集、善本书目等。经授权，这些资源可以整合在相关的数据库中，对文献著录、索引编制、知识拓展都是极为重要的资源。这方面，国际敦煌项目（IDP）已经有所实践，其数据库不仅是敦煌遗书的全部影像，对某号敦煌遗书的著录目录（如王重民《敦煌劫余录》）、研究成果皆做关联，因此在浏览某一文书图像时，可以同时了解其研究成果与进展。

二、古籍数字化的规范作业

[1] 对该数据库的设计和介绍，可以参见贾双喜《甲骨及甲骨拓片影像数据库的设计和实验》，《文津流觞》第 8 期。

无论是文献出版、高仿复制、专题展览、数据库建设，其来源都是数字化的藏品图像。目前，推广组主要负责古籍馆的珍贵古籍数字化的采集加工管理，以及部分专题数据库的策划。尤其是一线采集，经过五年多的具体实践，也逐步摸索出适应工作需要的一些业务规范。对于古籍影像的采集，值得一提的是傅斯年图书馆制作的详细的数字化流程，包括作业步骤、内容、具体规范等。这为我们制定相关工作规范提供了重要参考。

结合古籍馆的实际情况和数字化对象涉及的范围，推广组的规范作业要求人人了解库房管理制度，同时对数字化作业区进行安全管理，并在工作中细化和完善相关作业规范。

1. 协作范围

以古籍影印出版为例，从项目成立之日开始，数字化工作一般涉及五个科组，包括部门办公室、推广组、典阅组、修复组和立项科组（藏品所在科组）。部门办公室为总协调，一般负责立项与合同管理，提交成品，并协调藏品流动。立项科组负责提供藏品目录、配合修复组，给出藏品修复意见等。典阅组、修复组和推广组主要在加工流程中分工合作。其中典阅组负责藏品数字化前整理，包括书籍状况稽核、统计拍数、藏品出入库管理；修复组负责透字文献的衬纸或残损藏品的初步修整；余下的工作由推广组完成。

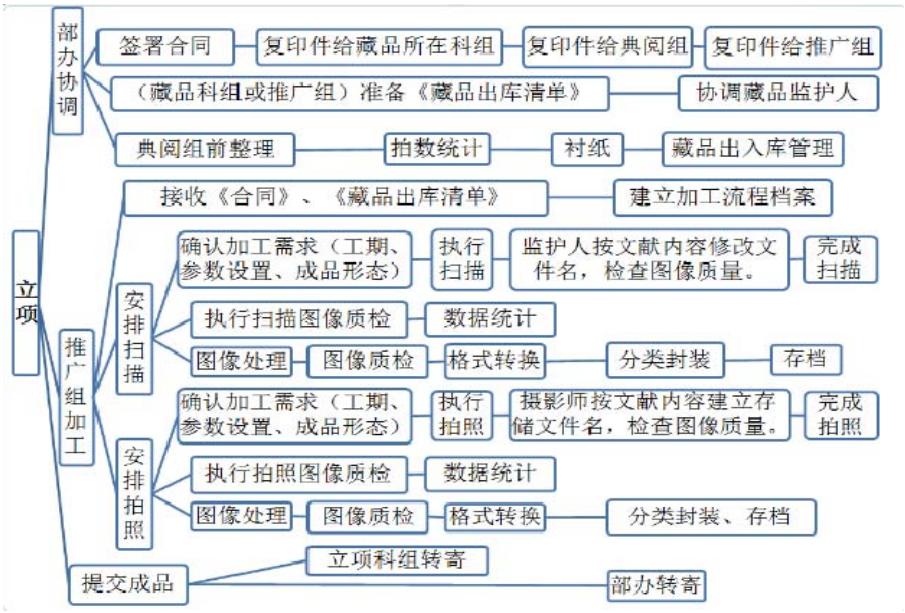

图 2 古籍数字化作业一般流程

从图 2 可以看到，接到藏品立项合同（附项目用书清单）、藏品出库清单后，

推广组负责提归藏品、安排扫描或拍照。这时要确认加工的需求，包括工期、数据参数、成品形态，在执行扫描和拍照过程中，要监护藏品安全，在完成扫描或拍照后，要进行图像质量检查，再根据要求转换格式、制作目录，刻盘或存盘，最后提交成品数据。这些流程内的关键环节，如合同所附藏品清单的索书号、版本情况的核对，藏品出库清单整理，扫描、质检等分类登记表制作，文献提交清单等，都要有相关文本存档，以备查询。

2. 扫描作业

2011 年以来，推广组数字化加工的主要形式是古籍扫描。使用的扫描仪有四种品牌：

（1）意大利 Metis 扫描仪：有两种型号，其一为 DRS 5070，适合扫描幅面比较大的文献，如地图等。但是超过双 A0 幅面的文献，则需要拼图。另一为 DRS 750，为书刊扫描仪，适合 A3 以内文献。

（2）台湾虹光书刊扫描仪：Avision fb6080E，适合扫描 A3 幅面以内的文献。

（3）法国 CopiBook（I2S）书刊扫描：主要适合 A3 以内的文献。

（4）德国赛数 Zeutschel 书刊扫描仪：适合扫描 A3 以内的文献。

从工作效率来看，每台扫描仪 7 小时可采集 600ppi 有效影像数据 500—1200 拍[1]，效率受图像分辨率高低要求、扫描对象保存状况及操作员熟练程度影响。虽然速度上扫描仪远低于拍照，但图像质量高，受环境因素影响小，为目前主要的加工方式。

2011 年，推广组还使用过照相机翻拍的方式。其工作速度快，每分钟可拍照 5—8 张，一台相机 7 小时可加工 2000 拍。其不足在于受光线、温湿度等环境因素影响非常大，相机本身的自动对焦功能不甚完善，对摄影师的专业技术要求很高，拍照生成的数据质量相对扫描较低；不同品牌相机的数据传输、软件兼容有时会出现问题。因此除非极易碎、难以翻动却急需加工的文献，邀请专业摄影师进行拍照外，其余基本以扫描方式进行加工。

扫描加工的作业规范，主要包括电源管理、机器保洁、色卡与标尺的摆放、扫描区域的设定、文件命名、存储路径、图像质检、扫描登记、藏品管理等内容。

[1] 按：在图像采集过程中，常用 PPI 作为描述图像分辨率的单位（pixels per inch），来表示输入设备的输入精度，如扫描仪，数码相机等，意即每英寸长度上有多少个像素。图像 ppi 值越高，画面的细节就越丰富。DPI 指输出分辨率，是每英寸长度上有多少个打印点（dot per inch），是针对于输出设备而言的，一般的激光打印机的输出分辨率是 300—600dpi，常见的冲印一般在 150 到 300 dpi 之间。

其中，色卡与标尺，对书籍而言，一般放在左侧或右侧（同一种书或同一项目用书最好统一位置），与书籍边缘保留 1 厘米左右间距。如果是长轴手卷，则置于藏品上方，同样距离边缘 1 厘米左右。灰度卡与彩色卡可以并列摆放，开本稍大者，可将灰度卡压在彩色卡的长度尺上。目前我们所使用的是柯达色卡。扫描区域一般指线装书籍的一个筒子叶（个别项目要求半个筒子叶或双半叶），大尺寸藏品则需要划分扫描区域（如从左至右、从上至下），以便于数字化图像质量检查、拼图及数据存储管理。图片文件命名一般采用藏品号+题名+册序+图像流水号的方式。多册古籍，最好分级建立存储文件夹，按册序存储，以便未来查找数据。图像质检，指对照古籍原件，逐个检查图像数据，包括是否缺叶（漏扫），图像清晰度、完整度、是否歪斜变形、色彩还原情况，是否有杂物（碎屑、毛发等杂质），文件命名是否规范等，质检情况要认证登记在《质检工作表》中。扫描登记，指在《扫描登记表》上登记当日扫描古籍文献的藏品号、题名、版本、册数、扫描采集的像素解析度、拍数、扫描日期、扫描员、质检员、登记员、数据存储位置等信息。藏品管理，主要是数字化工作区内的专门人员负责，包括提书、归书，数字化流程内藏品安全监护，衬纸修复等前整理协调管理，非加工流程内书籍锁入保险柜等。藏品管理员掌握保险柜钥匙，不得随意交给他人。

在整个加工过程中，藏品管理员、扫描员、质检员相互配合，及时沟通，遇有藏品缺叶或残损，及时交给修复师修整，确保藏品安全。对于暂时无法修复的文献，要从扫描队列中移除，同时请示上级领导和藏品所在科组的组长，申请是否专门修复后再加工。修复台最好是带有抽屉，可以放置简单的修复工具，如剪刀、镊子、启子、铅砣等，同时也可以按尺寸规格放不同大小的衬纸。注意衬纸的尺寸以略窄于书叶为宜，可留出书口的鱼尾所在位置，不要衬的过满、过紧，以避免撑裂书口。尽量不要使用书画纸，而是要用宣纸，以免衬纸撤纸过程中损坏书叶、伤害原书。

需要特别注意的是，为保证数据安全，加工区电脑一律不得连接互联网。藏品不在加工或修整过程时，必须锁入保险柜。加工区工作台只能放铅笔、软尺或塑料尺，不能放个人书包等杂物、水杯、食品、尖锐文具等。个人物品放入更衣柜，饮水要在指定区域，远离操作区。秋冬干燥季节，在接触藏品前一小时内不能涂护手霜等化妆品，以防粘留在藏品上。加工区内人员不能穿细跟鞋，不能披

散头发，不得留长指甲。其余扫描仪参数设置、托稿台背景布置等具体作业规范，根据不同项目的要求设置，其标准就高不就低，以使当下采集的数字资源在一定时期内能够满足存档、出版、发布等不同级别的需求，从而减少对古籍的反复加工。

3. 数据管理

一般而言，数字化对象数据的管理，主要基于之前数据采集和质量控制流程中所产生的各种表格。用书量小的项目，数据管理的问题和优势并不明显，但用书量逾百册或更多的情形下，再加上馆藏地不同、藏品类型不同，势必要求数据管理的及时、有效。2014 年以前，推广组主要是设专人逐个项目核对数据，再统一登记到《数字化资源总表》上，《总表》包括之前各环节工作登记单的信息，是内容最全面的表格。随着数据量的增长以及相关岗位人员的流动，一人一表的工作方式难以满足异地协作和大量数据管理的需求。为此，2015 年，推广组尝试多人一表、实时更新的工作方式，以 access 配合快盘，由每个工作室的专人，将《总表》分别登记，并实时同步，一定程度上满足了数据著录和管理的需求。但长远来看，设计并建设古籍数字化资源管理系统是非常必要的，以这个系统为工具，扫描员、质检员、修复师、登记员各自在相应模块下登记，可以实现无纸化实时更新和按不同需求分类汇总，减少从纸本登记到电子表格中产生的错误，也节省时间。同时，该系统可以预留接口，便于以后与书目系统、展览系统、修复系统、借阅系统之间挂接，实现古籍文献的全流程数字化管理与服务。

三、古籍数字化工作者的业务素质要求

就数字化加工而言，看似只有提归书、扫描、质检、图像处理、文件存储等步骤，但实际操作起来并不简单。为了保证文件加工的质量，还需要具备相关的文史知识。最基本的知识，如古籍的结构，卷序、装帧、版式、目录著录规则，都对提高工作效率有诸多帮助。以多册本古籍而言，常用的册序，除了“上下”、“上中下”外，还有“元亨利贞”、“春夏秋冬”、“宫商角徵羽”等，还有以天干、地支等名称作为册序的情况。而大部头古籍，还有以千字文等形式来排序的情况，例如大藏经，即用千字文来排列各函。再如古籍的装帧形式，除了常见的线装书，还有经折装、包背装、蝴蝶装、卷轴装等，其透字情况的处理、文献

翻动方式、扫描或是拍照加工方式的选择，都需要根据实际情况提前进行调整、确定，并仔细叮嘱扫描员和质检员。对这些装帧结构的基本了解，有助于正确地选择和实施数字化加工方式。

除了文史知识，常用的软件，如 excel，access，photoshop 等，是数字化工作人员必须掌握的工具，藉之而实现数据表登记、统计，图像拼接、裁切、格式转换等相关工作。此外，还要结合工作需求，探索和掌握适用的小软件，如批量加水印、批量建立文件夹工具等，才能不断地提高工作效率。

古籍数字化从业人员对于古籍数字化的未来趋势应有所认识，才能扎实做好每一步基础工作，适应未来发展的需要。例如，虽然目前尚未开展数字化资源的编目工作，但是应该对古籍编目的原则有所了解，并思考能够适应不同书目数据的通行标准；元数据的制作要尽可能统一，以便未来建立可以整合的数据库；图像采集信息应该尽可能完整保留，永久存档文件应该不做任何数字化处理；古籍影像资源的目录建立与知识标引应如何实现；新的信息技术的发展将如何突破古籍数字化产品服务的局限而实现传播对象的最大化等等。这些问题，不再是依靠信息技术进步就能够解决的，而是要有一个明确的观念，要对古籍资源建设有统一的规划和全局的考量，要将古籍数字化的具体工作标准化，实现编目标准、图像数据处理标准、元数据标准、多语言术语标准等方面的统一，从而使古籍数字化这个高投入的行业实现利用和效益的最大化，以满足学界和大众的不同需求。