

敦煌遗书数字化流程概述

□ 刘 婷 国家图书馆古籍馆

国际敦煌项目(The International Dunhuang Project, 简称 IDP)是旨在使敦煌及丝绸之路东段其他考古遗址出土的写本、绘画、纺织品等文物文献的信息和图像能在互联网上自由地获取,并通过教育与研究项目鼓励使用者利用这些资源的国际合作项目^①。项目由英国国家图书馆发起,其他收藏机构以合作机构的方式参与,将所藏藏品以数字图像的方式通过 IDP 网站发布出来。中国国家图书馆于 2001 年 3 月加入国际敦煌项目,同年底成立数字化工作室开始对馆藏敦煌遗书数字化的工作。IDP 数字化工作室的基本工作主要分为两大部分,文献拍摄和图像处理、发布。

一、敦煌遗书的拍摄

(一) 拍照的设备

IDP 项目拍照使用的是玛米亚 Mamiya 数码相机,照明设备为三基色柔光灯^②。玛米亚相机为中画幅相机,搭载了飞思数码后背,成像更加细致清晰;而三基色柔光灯的光线柔和、色彩还原率高,尤其是灯管的表面温度低,属于冷光源,非常适合古籍文献的拍摄。

(二) 拍摄写卷的定名

IDP 项目发布的敦煌写卷正反两面都会拍摄,一般按内容分为三种,写卷的正面名称为 R1,背面名称为 VB,还有一些写卷的背面也有文字,这些有字的部分单拍,定名为 V1。

(三) 拍照后的文件转换

拍摄完成后的图像需要用相应软件进行处理, IDP 项目使用的是丹麦的飞思数码后背公司开发的 CAPTURE ONE 软件, CAPTURE ONE 功能强大,可以对数码相机拍摄出来的图像进行 RAW 图像格式的转换。通过 CAPTURE ONE 会把拍摄完的原始图像转换成两种存储大小的数据,第一种 8bit 图像数据,分辨率为 300dpi,图像大小为 15MB 左右,这就是后续需要进行图像处理的数据。第二种是 16bit 图像数据,分辨率也是 300dpi,每幅图像大小为 480MB 左右。需要转换的是写卷正面和背面有字部分的单拍,另外是背面无字部分的图像数据,也是 8bit、分辨率 300dpi,每幅图像大小为 240MB 左右。除 8bit、15MB 大小的图像数据外,其他数据后续不做任何处理,直接用于永久保存^③。

二、敦煌遗书的图像处理

^①国际敦煌项目中文官网:<http://idp.nlc.cn>

^②乌心怡:《古籍拓片及舆图数字摄影解析》,《图书馆学刊》2014 年第 8 期,第 116—118 页。

^③乌心怡:《国家图书馆敦煌文献数字化概述》,《山东图书馆学刊》2011 年第 4 期,第 88—89 页。

（一）图像处理的硬件和软件

IDP 工作室使用的图像处理硬件设备是苹果电脑和 Wacom 数位板, IDP 项目进行的 20 年间, 硬件设备历经几次更新换代, 现在的版本为苹果 iMac21.5 英寸一体机和 Wacom 的 Intuos5 数位板。Wacom 第五代 (PTH—650) 新推出的有支持直觉式输入的多指触控功能和快捷功能等, 有效地提高了工作效率, 使用起来也更为便捷。

软件方面使用的是 Adobe Photoshop, 简称“PS”, Photoshop 的版本有很多, 最早的 Photoshop 版本为 1—8, 2003 年 PS8 更名为 Adobe Photoshop CS, 直到 2013 年又推出了 Photoshop CC, 取代了 Adobe CS 系列的最后一个版本 Photoshop CS6。目前的最新版本为 Adobe Photoshop 2021。Photoshop 作为一款图像处理软件, 功能强大, 使用方式智能多样, 便于应用, IDP 的所有图像处理过程都是运用该软件完成的。

（二）图像处理的流程

1. 单拍图像处理

单拍的图像处理主要是抠图, Photoshop 的抠图方式有很多, 这里主要用的是两种方法, Photoshop3 之前可以使用“抽出”, Photoshop3 之后取消了“抽出”, 则可以使用“边缘调整”。

(1) 方法一

①图像抽出: 首先, 打开需要处理的图像, 选择“滤镜”下拉栏的“抽出”选项, 进入“抽出”界面, 勾选“智能高光显示”, 然后将图像放大, 选择“边缘高光器工具”, 使用高光器工具沿写卷边缘描画, 过程中一旦出现错误, 可用“橡皮擦工具”进行修改。描画完整后, 选择“填充工具”, 在需要保留的图像上点按, 这时选取的内容就会变为紫色, 然后可按“确定”取出图像并退出此界面。

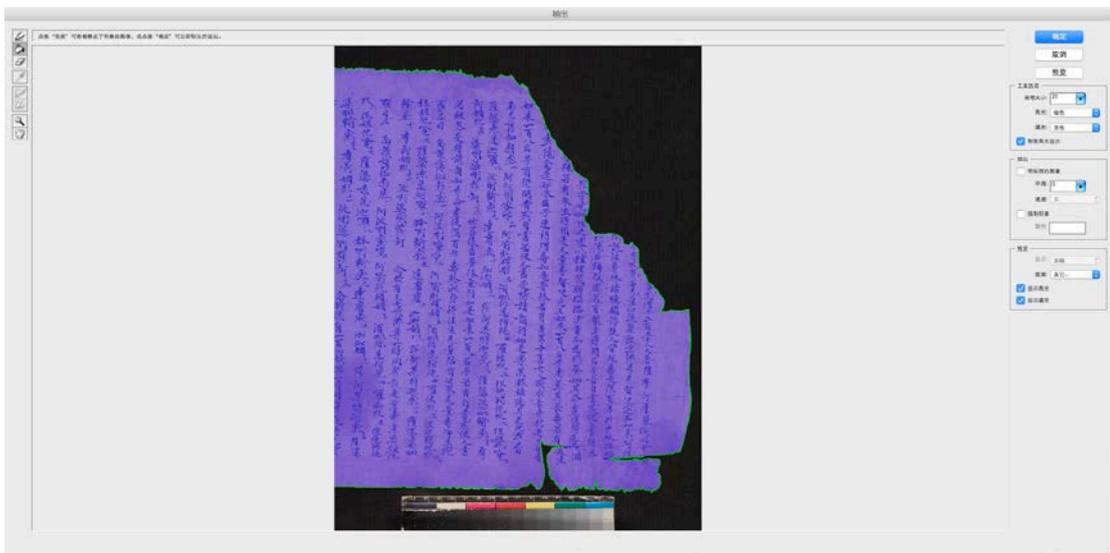


图 1 图像抽出界面

②设立背景图层: 在已有的“图层 0”的基础上创建一个新图层“图层 1”, 并将“图层 1”下拉至“图层 0”下方, 并使用“command+delete”快捷键将图层颜色设为黑色, 成为单拍图像的背景。

③图像修整：将图像适量放大，使用“橡皮擦工具”对图像的边缘进行修饰，去掉多余的部分（此处应注意一定是在“图层 0”上进行修整）。

④添加标尺及存储图像：将标尺拖拽至图像下方黑色背景处，左右上下移动至合适位置，最后将图像保存。至此，单拍图像的制作处理便全部完成了。

（2）方法二

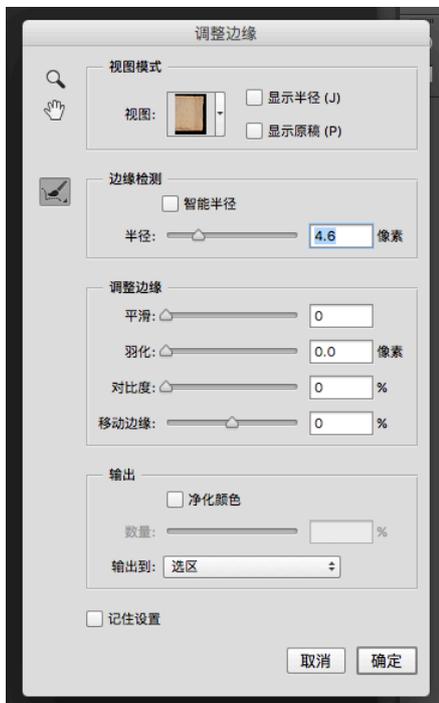


图 2 调整边缘界面

①打开图像，选取需要保留的内容（可以使用套索或者矩形工具），点击“属性面板”右侧的“调整边缘”，打开“边缘调整”窗口后，用“画笔工具”或“橡皮擦工具”调整需要抠图的区域，期间可适当增大“边缘半径”，完成后点击“确定”。

②“command+shfit+i”进行区域反选，使用“command+delete”变为黑色背景，再放大图像并用“橡皮擦工具”修饰边缘部分，最后添加标尺。

方法一和方法二主要的不同点就是抠图部分，之后的操作基本一致。方法一更为细致，具体操作起来步骤较少；方法二更加智能，其中对工具的选择也更具多样性，图像处理时可根据软件版本的不同和工作习惯进行选择。

2. 图像拼接处理

由于敦煌写卷的长度不同，大多数写卷无法仅用单拍拍摄完成，这时就需要对拍摄完的写卷进行拼接。拼接后的图像存储名称为“ST”。而一些写卷比较长，拍摄的拍数过多，无法拼接为一个完整的长图，则可制作多个拼接图，一般情况下 10—15 拍制作作为一个拼接图，另存为 ST1、ST2、ST3……以此类推，直至整个写卷全部拼接完成^①。

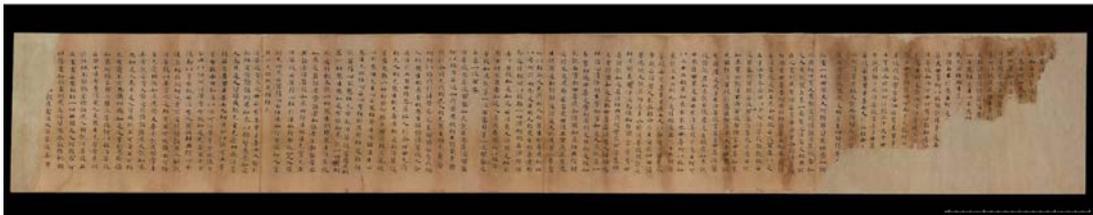


图 3 敦煌写卷连接图

（1）图像拼接

用 Photoshop 软件打开第一拍，点击“图像”下拉选项“画布大小”，点击“定位”右边第二个箭头，并将图像宽度加长（具体数值可根据写卷长度不同进行调整），把背景设置为黑色，然后将图像另存为“ST”，这是英文单词 Stitched（拼接）的缩写。将下一拍图像

^①乌心怡：《国家图书馆敦煌文献数字化图像梳理技术探要》，《图书馆学刊》2011 年第 4 期，第 118—121 页。

需要处理的图层拖拽进此图像中，后来的图像在工具栏中会成为“图层 2”，将图像放大，设置“图层 2”的不透明度为 60%左右，找到两拍之间的重复处，用“图层 2”覆盖，调整图像使两拍之间边缘、文字对齐。此处可使用“自由变换”命令（`command+t`），对图像的角度、方向等进行缩放，使两拍的上下边缘重合处对齐，再把“不透明度”调回 100%，然后用“橡皮擦工具”将两拍之间的连接痕迹擦去，图像拼接就可完成。以此类推，之后的每一拍都同样处理，直至这一拼接图完成。

（2）图像色差调整

由于受拍照时的环境等因素的影响，同一写卷前后内容也可能出现色差，这就需要对图像的颜色进行调整，尤其在进行图像缀合时，图像色差的调整可以使整个连接图看起来更加和谐统一。点击需要处理的图层，选择“图像”→“调整”→“曲线”（或者快捷键“`command+m`”），弹出“曲线”界面，界面中的“通道”有 RGB、红、绿、蓝几个可调整内容，对应下方的线条，根据图像的色差，先选择“通道”中要调整的内容，然后滑动下方的线条，直至颜色调整完毕，点击“确定”。

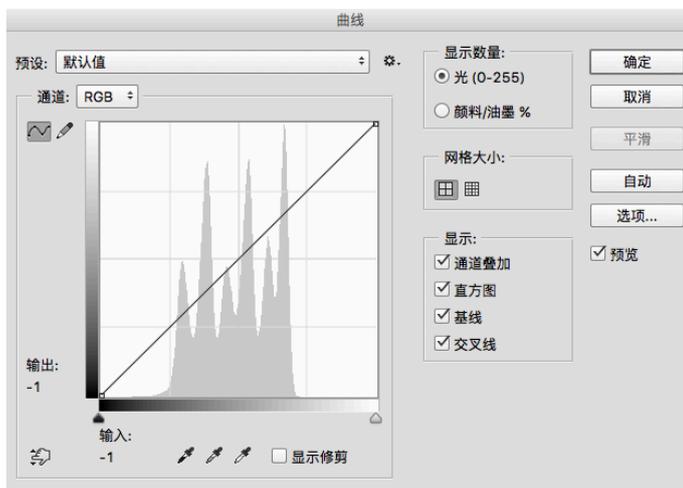


图 4 “曲线”界面截屏

3. 图像批处理

处理后的图像为 TIF 格式，容量为 30MB 左右，这样的图像仍然过大，不能直接上传到数据库，否则会造成网站打开速度变慢等问题，因此需要将图像转换成 JPG 格式。如果一步一步进行操作的话，大量的重复动作造成时间和精力的双重浪费，这时就可以采取批处理的方法。这里需要录制两个“批处理”，一个是“合层”，即将一个图像的几个图层合为一个，另一个是 JPG 格式转换。录制好后一定要先“合层”后再转换 JPG 格式，这个顺序不能错。批处理结束后，图像处理工作就完成了，之后才能进行图像的上传工作。

三、敦煌遗书的图像发布

（一）拍照前检索写卷清单，确保不会重复拍摄

由于一些工作项目的需要，以及应学者要求优先数字化某些写卷，我们开展敦煌遗书数字化的顺序并非完全严格按照编号进行，可能会出现某些编号靠后的写卷提前拍摄的情

况^①，这就要求在拍摄开始前要对准备拍摄的写卷进行检索，确保其之前并未拍摄并在平台发布，以免重复拍摄，造成工作失误。

（二）处理后的文献发布

文献的发布过程大体上分为两部分：4D 数据上传和 FTP 图像上传。

4D 是国际敦煌项目专门设计的用于敦煌遗书数字化的软件，其数据库是最早的复合型数据库，内容包括多国藏品的彩色图版影像和目录信息，可以通过对文献信息的录入进行数据上传。不过 IDP 数据库并非敦煌遗书专题库，库中还有许多中亚地区相关古代历史遗存的数字资源^②。

1. 4D 数据上传

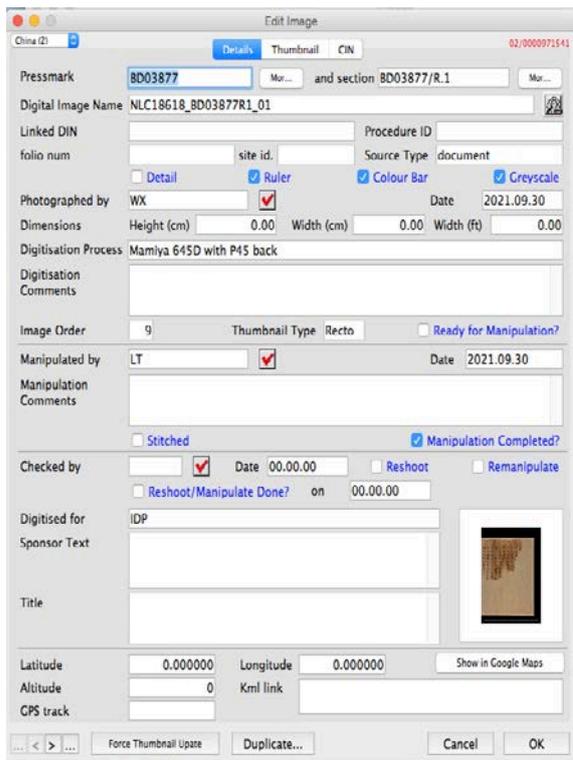


图 5 4D 数据上传截图

2. FTP 图像上传

IDP 项目使用 FTP 向服务器传送图像数据，配合 4D 软件在数据库中发布写卷图像。FTP 是用于在网络上进行文件传输的标准协议，IDP 对 FTP 上传工具没有硬性规定，一般可以进行相关文件传输的软件都是可以的。登录 FTP 的工具后，将已经批处理好的图像直接拖入即可。

还有一个正确显示图片的小技巧，那就是一定要先进行 4D 数据上传，然后再用 FTP 上传图像。刚上传完成时是不显示图像的，此时需要再次登录 4D，点开上传的数据条目，逐一点击条目界面左下角的向右浏览的箭头，全部点击后方可显示正确图片。如果先进行

登录 4D 数据库，点击“Data”→“Image Management”，出现新的界面，点击界面上的“add”，弹出的界面就是数据上传界面，里面包括写卷的馆藏号、尺寸、数据顺序、类型、拍照设备等条目，需要逐一录入。数据录入过程中要注意的有几点：第一，录入的内容需要与图像完全对应，尤其是文件名，除文件后缀不需要录入，其他必须完全相同，即使多一个空格最后图片都会无法显示。第二，图像顺序“Image Order”要计算清楚，以确保图像以正确的顺序在网页上展示，如正反面各 9 拍，正面第一拍的顺序号应为 9，背面第一拍的顺序号应为 18，连接图的顺序号统一为 999。第三，每一拍上传时一定要点取“Manipulation Completed”，否则录入完成后图像只在 4D 软件上显示，IDP 网站则无法显示图片，连接图则要同时点取“Stitched”。

^①刘波、乌心怡：《中国国家图书馆参与国际敦煌项目（IDP）二十年回顾》，《文津流觞》2021 年第一期，广西师范大学出版社，第 278—289 页。

^②韩春平：《敦煌遗书数字化演进史》，《中国社会科学报》2017 年 6 月 28 日，第 7 版。

FTP 上传,那么就会导致写卷的第一拍和第二拍都显示第一拍的图像,后面的第三拍显示第二拍的图像,之后所有图像显示的都是上一拍的内容,需要把每一拍刷新一遍。

（三）文献上传后的质检

文献上传后需要登录 IDP 网站,在检索栏输入编号,然后点开图像逐一检查,如果某一拍无法显示,就需要对照 4D 录入的数据和 FTP 上传的图像进行检查,看是否有错误之处。

四、图像拍摄处理过程中可能出现的问题

图像的拍摄和处理是个步骤繁杂的过程,Photoshop 等软件的操作是相对基础的工作技巧,其难点在于日复一日的面对相似的图片的枯燥过程和巨大的待完成的工作量,这就需要工作人员细致耐心地对待,仔细做好每一个步骤。相对于研究人员对于文献内容的关注,摄影师和图像处理工程师更多地会关注文献本身的破损程度、写卷长度、拍摄内容和原件的颜色差异等,这些都决定着数字化工作的效率。当然其中的疏漏也不可避免,一般来说可能出现的问题有几点:拍摄时编号容易混淆,敦煌遗书数量巨大,仅中国国家图书馆藏敦煌遗书就有 16000 多号,相近的编号,比如 BD03428、BD03482、BD03824、BD03842 等,在拍照时输入时需要格外仔细,绝对不能出错,避免有同一编号错误挂接两件写卷的情况。图像处理和发布时,要注意图像色差的调整、文献上传后的质检,尤其是 4D 数据库的录入过程中填写写卷编号、类型和顺序,一定要准确无误。

相对于图像处理,IDP 对于摄影师的要求显然要更严格,图像处理主要接触到的是以像素构成的数字图像,并不直接面对敦煌遗书,处理过程中出现错误的话可以直接在相关软件上进行修改。而拍摄一旦出现错误,则需要将原件重新取出进行补拍,这对于文献保护是很不利的。因此,摄影师在拍照时要格外注意,避免出现曝光过度、对焦不清甚至漏拍等情况。

五、IDP 数字化的意义

数字化其实就是将已知的各种实物信息转化成为二进制的信息代码的过程,文献的数字化,就是将已知的传统介质的各种文献资料用计算机技术转化成数字化的信息,通过互联网等平台可以供学者和普通民众研究阅览,实现资源共享。

同时,数字化相比其他的传统媒介的修改成本要低很多。以敦煌遗书为例,一些已经出版的关于敦煌遗书的目录资料等,一旦出现编写或者印刷错误,或者定稿时并未认定清楚的写卷,随着后续研究的深入可能被重新认识或定名。出现这样的情况,纸质等传统媒介修改十分不易,数字化的模式则显然要方便很多,一般情况下只需要在后台软件上进行修改即可。除了每个写卷上传时的质检,IDP 每经过一段时间就会对已经发布的内容进行一次整理和检查,以便能及时改正错误。

国际敦煌项目通过 IDP 网站将海量的敦煌遗书资源免费提供给需要的人,截止 2021 年 10 月,IDP 网站已经发布的图像数据超过 54 万幅,其中中国国家图书馆发布已超过 20.8 万幅,共 5400 多件。敦煌遗书数字化既方便了读者的浏览查阅,加快了学者对敦煌研究的进程,又有效地保护了敦煌遗书,避免了可能的二次损害。