数字推广工程·地方文献数字化项目建设经验交流

湖南图书馆：彭维

尊敬的各位领导、各位同仁：

大家上午好！

目录

一、基本流程

- **文献确认：**文献遴选—文献申报—文献确认
- **图像采集：**图像扫描或者拍照
- **图像处理：**图像命名—图像裁剪—图像纠偏—图像去污
- **OCR识别：**文字识别—文字校对
- **PDF制作：**单版PDF生成—命名—校对—合并单版PDF—书目制作
- **TXT制作：**单版TXT生成—命名—校对—描述—合并版TXT
- **数据表制作**
- **第三方验收**
- **数据提交**

二、人员配置

岗位	主要职责	人数	备注
项目管理岗	1、项目申报、验收、提交。 2、业务指导、质检、数据表制作 3、人员资源的调配	1	熟悉数字化工作流程，阶段性质检
图像采集与处理岗	1、图像扫描或拍照 2、图像的裁剪、纠偏、去污等 3、图像文件的命名	1	临聘人员，在进行图像处理的时候，须逐页处理，留意每一页扫描的质量， 发现扫描质量问题立即发还上一工序重扫。
文字识别岗	1、将处理好的图像进行全文识别 2、文字的校对 3、PDF 及 TXT 制作及命名等	1	临聘人员

三、具体方案

- **3.1文献遴选：** 文献的选择是项目建设的开端，一个好的文献选择直接关系后续工作量的大小以及制成品质量。陋以为，选书遵循以下原则，可让项目建设更加的顺利。

- 严格限定地域：域范围应该严格限定在本馆所辖区域内，不得越级或降级。

- 以内容为划分：地方文献概念上有狭义和广义之分，广义上的地方文献包括地方出版物、当地人著述，经常有区域交叉。尽量以文献内容是否是地方的来进行划选。

- 先易后难：从简单的文献开始着手，选一些纯文本的、字体较大、书貌较新的精装本文献对后续工作的开展比较有利。

- 版权优先:为方便该数字文献的传阅，已取得版权或已进入公有领域的文献无疑是最佳选择。

- 3.2、图像采集：图像采集主要有两种方式，扫描方式和拍照方式。
- 拍照采集方式：拍照速度快，但要保证页面与镜头垂直角度比较难处理，图像清晰但常有偏色问题，设备价格也更为昂贵。
- 扫描采集方式：更容易上手，图像清晰，选择文本模式能自动处理透字问题，设备成本较低，但速度较慢，扫描时的压框对文献二次伤害更大。
- 综合比较这两种图像采集方式，纸张质量较好的地方文献数字化则主要使用平板扫描仪进行扫描。

3.3扫描图像质量审校：

- 因扫描图像的质量直接关系到文字识别的准确率和PDF显示效果，往往很多质量不过关的成品都是因为其扫描原件质量不好。所以对扫描图像的质量审校是不可或缺的一环。图像审查首先是图像参数检查，看是否与标准相符，然后检查其图像质量，对照原书查看是否有缺页漏页等问题。如发现有漏扫、重张，图像黑边、折角、不清晰等情况，则进行重扫或补扫误，并及时进行调整，这样方能避免以后的连环出错。

- 3.4、图像处理：须采用图像处理软件对TIFF原图进行裁剪、纠偏、去污处理等处理，为保证图像能真实反映原件的风貌，一般不对其进行锐化等渲染性的处理。
- 图像裁剪：按建设标准是要图像保留到文献的外边缘，但如果是在不拆书的情况下采集的图像，因书脊处夹框的问题很难做到这一点。我们的做法是尽可能的保证正文内容、页眉、页脚、反面印章、附件、手写注释等信息完整，且正文部分在裁剪区域的中央。需要注意的是，同一种书各页裁剪后图像尺寸大小应保持一致。
- 图像纠偏：纠偏需要手工逐条进行处理，对出现偏斜的图像进行纠偏处理，图像歪斜度不可以超过一度，对方向不正确的图像进行旋转还原，以符合阅读习惯。

- 3.5、OCR识别：OCR识别是该项目中最重要的一环。需要借助OCR识别软件对tiff原图进行识别，然后在全文识别的基础上对文字进行人工逐一校对。校对后方能保证生成的双层PDF可实现无障碍的全文检索。
- 识别软件的选择：比较常用的桌面识别程序国内有汉王、清华文通、国外有款abbyy中文泰比，就识别效果来说感觉还是abbyy强些，前提是你的图片质量一定要过关质量太差，再好的识别结果也不会好了。建议选择正版的文字识别软件，从而可以提高文字的识别率。
- 文字的校对：但即算是最好的识别软件也不能达到项目建设标准差错率千分之三以内。初步统计，目前我们使用的是Abbyy FineReader，就算对质量较好的图像其文字识别率也仅仅95%左右。所以需要在全文识别的基础上逐行逐字进行人工校对。

3.6、PDF制作

文字校对完了就可以开始制作单版的PDF了。需要注意的问题是需要生成图在文上的格式，且要调整文字区与反显区的位置重合。反显区域与文字区域相差1毫米以内，这样才能保证全文检索时光标能停留在准确的位置。标准要求控制生成的单个PDF文件大小在80KB-200KB之间。但不能过度压缩，这样会导致PDF文字模糊，一般24K页面将压缩因子控制在20%左右，PDF文件大小120左右为最佳。

3.7、TXT制作

要注意TXT文本的排版，尽量保持文本格式和原书排版格式一致，尽量使其样式美观、大方，不能有错别字和漏字，方便读者阅览。同时经常在文献当中会有一些特殊情况：如表格、注释、插图、插图页等等，有些转换不了文本的，或需要特殊标记的。这些情况需要在TXT当中进行描述或转换。

四、存在问题

- 4.1单个PDF文件大小悬殊：按照要求，为了保证PDF质量，对单个PDF文件的大小是有一个区间限制的(80-200KB)之间，可是在实际加工过程当中，同一种书虽然页面大小一样、PDF压缩比例设置相同，但制作出来的PDF大小却依然悬殊。个人臆断造成这种情况的原因有可能是因为在进行OCR转换的过程中，每个文件所包含的内容不同所造成的，比如说：文字、图片、表格、注解小字体等等，但并不影响PDF的显示效果。是否可以将该项限制适当放宽，以更加方便加工操作。

- 4.2版权保护的问题：版权问题一直是制约图书馆资源数字化建设的瓶颈问题，同样在我馆今年的地方文献数字化项目建设中依然是个悬而未决的难题。目前还有很多地方文献均处于版权保护期之内，国家图书馆老师给的建议就是能够解决版权问题是最好的，采取签订协议等等办法。但是实际实施起来困难还是比较大的，期待从更高的层面有一个统一的解决方案。

五、结语

- 互联网+时代，文献资源的数字化联合建设将是图书馆文献资源建设的一条重要分支，是扩充图书馆馆藏、提高图书馆服务能力的重要手段。我们应当以文化部、国家图书馆的各项目为依托，以数字图书馆为发展纲要，不断推广和发展自身的数字资源建设。把珍藏的、有价值的、快被时间湮灭的传统文献替换为无形的数字文献，努力打造一个以国图为中心、省市地各级图书馆为节点的，开放、共建、共享、共营的数字文献云。


谢谢各位！