# 藏经数据规范化工作文档

1. 现有的藏经数据命名不统一，需要加以规范
2. 数据使用者：
   1. 藏经切分
   2. 切分校对
   3. 藏经识别
3. 产出版本
   1. 原始图片
   2. 重命名后图片
   3. 修改大小后图片
   4. 切栏后图片
4. 命名规则
   1. 以经为单位组织的藏经按如下规则命名

藏经命名代码\_S#####\_R###\_T####.\*

例：高丽藏第1部经第22卷第28页：GLZ\_S00001\_R022\_T0028.jpg

* 1. 以册为单位组织的藏经按如下规则命名

藏经命名代码\_V###\_T####.\*

例：永乐北藏第1册正文第33页：YLBZ\_V001\_T0033.png

C）图片名（不含后缀）中不得带有字母、\_、数字以外字符

1. 藏经命名代码
   1. 乾隆大藏经：QLDZJ
   2. 永乐北藏：YLBZ 已上传
   3. 高丽藏：GLZ 已上传（网站清晰版）
   4. 碛砂藏：QSZ
   5. 洪武南藏：HWNZ
   6. 赵城金藏：ZCJZ
   7. 中华大藏经：ZHDZJ
   8. 宋藏遺珍：SZYZ
   9. 房山石经：FSSJ
   10. 高丽藏初雕：GLZCD
   11. 元官藏：YGZ
   12. 契丹藏：QDZ
2. 页类型代码
   1. 正文:T,
   2. 目录:I,
   3. 序:P,
   4. 封面、封底:C
3. 处理流程
   1. pdf文件
      1. 转换为图片

输入：pdf文件

输出：分册重命名图片（大小分辨率与pdf保持一致），对应目录信息

|  |  |  |  |
| --- | --- | --- | --- |
| 图片本地存储路径 | 转换前 | 转换后 | 是否上传成功 |
| ~/##册/\*\*\*\*.\* | ###册###页 | 见4.b |  |

* + 1. 调整分辨率、大小，以便切分校对平台使用

切分校对平台需要分辨率为宽1200像素，高根据宽进行调整，需安装Pillow库

核心代码

**from PIL import Image**

**import os**

**d**ef resize\_image():  
 for dir in os.listdir('/media/buddhist/智能信息中心/GLZ'):  
 for jpg in os.listdir('/media/buddhist/智能信息中心/GLZ/'+dir):  
 img\_path = '/media/buddhist/智能信息中心/GLZ/'+dir+'/'+jpg  
 c = Image.open(img\_path)  
 if c.width != 1200:  
 c.resize((1200, int(c.height \* 1200 / c.width)), Image.ANTIALIAS).save(img\_path, quality=85)  
 h.write(dir+'/'+jpg+'\n')  
 else:  
 continue  
 h.write('dir\n')

* + 1. 上传到亚马逊s3平台

分册上传，上传后检查是否上传完整，一册传完再传下一册

上传核心代码：

Import boto3

s3 = boto3.resource('s3')

s3.meta.client.upload\_file(图片本地存储路径, 'lqcharacters-images', key)

检查上传成功核心代码：

s3\_client.get\_object\_acl(Bucket='lqcharacters-images', Key=key)

如果不存在，会抛出异常，可以用

try:

s3\_client.get\_object\_acl(Bucket='lqcharacters-images', Key=key)

except Exception as e:

s3.meta.client.upload\_file(img\_path, 'lqcharacters-images', Key=key, Callback=call\_back)

重新上传

也可以直接在上传中加入回调函数：

**d**ef call\_back(c):  
 global yes\_or\_no  
 yes\_or\_no = c

s3.meta.client.upload\_file(图片本地存储路径, 'lqcharacters-images', Key=key, Callback=call\_back)

上传成功，会返回一个大于0的数字

因为网络中断会导致上传中断，所以还是需要检查，解决的方法是把上传成功的key名称写入一个文件，然后与所有key进行循环对比，直到全部写入成功为止。