数字资源建设标准应用与实例解析

------国家图书馆征集数字资源建设规范

国家图书馆数字资源部　龙伟

2013年4月11日

目 录

1. 征集资源建设背景

◆ 2010年起，国家图书馆联合全国各图书馆共建国家数字图书馆，面向全国各省级公共图书馆等机构，广泛征集数字资源，得到各地图书馆的热切响应和积极参与。通过征集数字资源，带动了图书馆特色数字资源的建设和应用发展。

◆ 密切结合“数字图书馆推广工程数字资源联合建设”工作，重点地征集主题明确、特色鲜明的优秀数字资源，为国家数字图书馆和数字图书馆推广工程的建设与服务提供重要的资源保障。

◆ 重点征集地方文献、人文悦读、图林拾珍专题库内容要求的已经完成建设的数字资源。优先征集具有鲜明地方文化特色的精品数字资源，侧重征集与主题有关的全文型数字资源，包括文本全文、图像、音视频等。

制定《国家图书馆征集数字资源建设规范》，按照“建设规范”以及数字资源整合发布要求，对元数据、对象数据进行全面的数据检查、版权检查和数据整合，形成最终符合要求的不同加工级别的数据，并建立完备的数据说明文档，提交各类检查情况报告。

工作流程

截至目前为止，已有34家单位提交了6万余种资源数据，已完成了资源质检和整理，并提交发布。

序号	主题	数量（种）	序号	主题	数量（种）
1	馆藏特色资源	12732	6	专题视频资料	3007
2	地方志	1900	7	年画	224
3	民国文献	9795	8	非物质文化遗产	497
4	家谱	377	9	少年儿童资源	86
5	老照片	33708	10	少数民族资源	12

数字资源征集工作聚合了一大批类型丰富、覆盖广泛、规模庞大的数字资源，推动了图书馆数字资源的建设和应用。

首先，征集的成果数据最终在推广工程网站上统一揭示，提供给广大读者和研究人员使用，为其提供了数字化的便利手段，大批量、多类别的数据集中在一起，提升了数据的应用价值。

其次，通过对征集数据的统一处理和管理，实现了各地方馆资源的数字化、规范化、标准化保存，有利于未来的长期利用，具有重大意义。

最后，各地方馆也已通过本项目，借助国家图书馆的统一平台展示自身资源，实现全国图书馆范围内的资源共享。

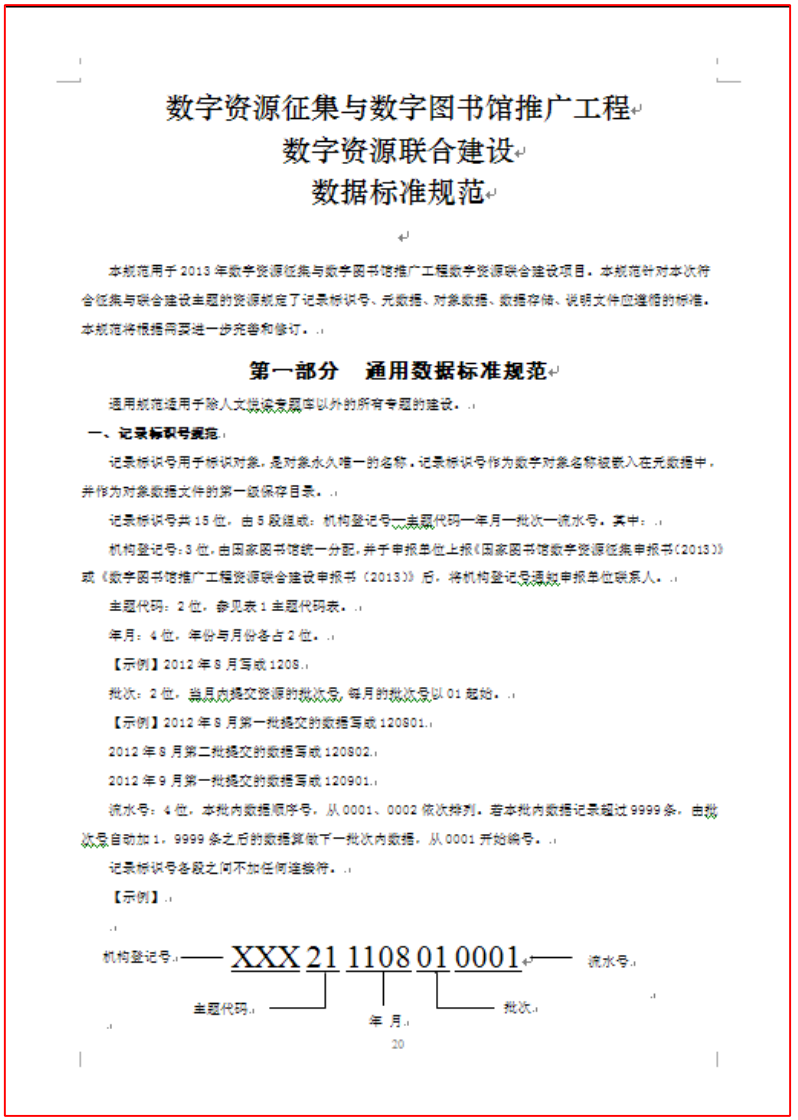
目 录

第一部分　通用数据标准规范

1. 规范概述---

《数字资源征集与数字图书馆推广工程数字资源联合建设数据标准规范》，规范针对征集资源（地方文献、民国文献、老照片、年画、非物质文化遗产、少数民族资源、碑帖、图林拾珍、网事典藏、其他特色资源等10个主题），规定了记录标识号、元数据、对象数据、数据存储、说明文件应遵循的标准。

规范将根据需要进一步完善和修订。



数字资源征集与数字图书馆推广工程
数字资源联合建设
数据标准规范

本规范用于2013年数字资源征集与数字图书馆推广工程数字资源联合建设项目。本规范针对本次符合征集与联合建设主题的资源规定了记录标识号、元数据、对象数据、数据存储、说明文件应遵循的标准。本规范将根据需要进一步完善和修订。

第一部分　通用数据标准规范

通用规范适用于除人文地理及环境以外的所有专题的建设。

一、记录标识号规范

记录标识号用于标识对象，是对象永久唯一的名称，记录标识号作为数字对象名称被嵌入在元数据中，并作为对象数据文件的第一级保存目录。

记录标识号共15位，由5段组成：机构登记号—主题代码—年月—批次—流水号。其中：

机构登记号：3位，由国家图书馆统一分配，并于申报单位上报《国家图书馆数字资源征集申报书（2013）》或《数字图书馆推广工程资源联合建设申报书（2013）》后，将机构登记号通知申报单位联系人。

主题代码：2位，参见表1主题代码表。

年月：4位，年份与月份各占2位。

【示例】2012年8月写成1208。

批次：2位，当月内提交资源的批次号，每月的批次号以01起始。

【示例】2012年8月第一批提交的数据写成120801。

2012年8月第二批提交的数据写成120802。

2012年9月第一批提交的数据写成120901。

流水号：4位，本批内数据顺序号，从0001、0002依次排列。若本批内数据记录超过9999条，由批次号自动加1，9999条之后的数据算做下一批次内数据，从0001开始编号。

记录标识号各段之间不加任何连接符。

【示例】

机构登记号 — XXX 21 1108 01 0001 — 流水号

主题代码　年 月　批次

20

2. 记录标识号规范

记录标识号用于标识对象，是对象永久唯一的名称。记录标识号作为数字对象名称被嵌入在元数据中，并作为对象数据文件的第一级保存目录。

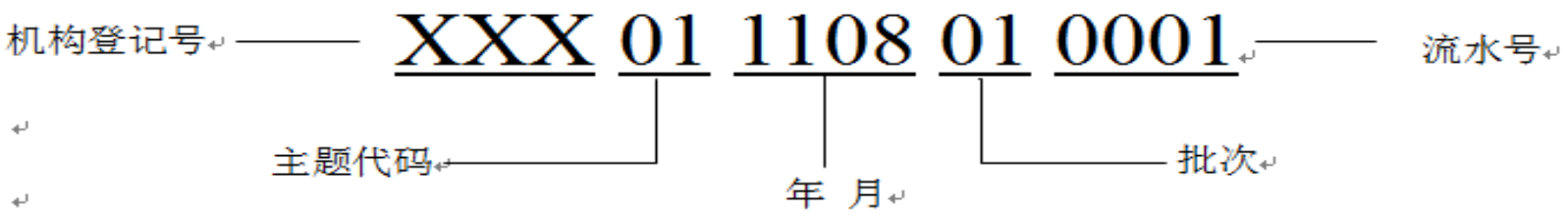
记录标识号共15位，由5段组成：机构登记号—主题代码—年月—批次—流水号。

- *机构登记号*：3位，由国家图书馆统一分配，并于应征单位上报《国家图书馆数字资源征集申报书》后，将机构登记号通知应征单位联系人
- *主题代码*：2位
- *年月*：4位，年份与月份各占2位

主题代码表

主题名称	代码
其他特色资源	01
地方文献	02
民国文献	03
家谱	04
老照片	05
年画	06
专题视频资料	07
动漫素材	08
非物质文化遗产	09
少数民族资源	10
少年儿童资源	11
历史文化	12
科普	13
网事典藏	14
碑帖	15
图林拾珍	16

【示例】



- *批次*：2位，当月内提交资源的批次号，每月的批次号以01起始。

【示例】2012年8月第一批提交的数据写成120801

2012年8月第二批提交的数据写成120802

- *流水号*：4位，本批内数据顺序号，从0001、0002依次排列。若本批内数据记录超过9999条，由批次号自动加1，9999条之后的数据算做下一批次内数据，从0001开始编号。

记录标识号各段之间不加任何连接符。

3. 元数据规范

元数据：数字资源对应源文献的内容及特征的描述。本规范元数据推荐选用CNMARC（ISO2709）格式或“都柏林”元数据核心元素集。

- CNMARC（ISO2709）

遵循《新版中国机读目录格式使用手册》及其它相关编目著录手册。

001字段内容为“记录标识号”，指由机构登记号—主题代码—年月—批次—流水号5段组成的15位记录标识号。

- “都柏林”元数据核心元素集

遵循GB/T25100-2010信息与文献:都柏林核心元数据元素集。

“标识符”元素内容为“记录标识号”

- 文化行业标准元数据规范

针对不同的资源类型，遵循文化行业标准《专门元数据元素集及著录规则》系列元数据规范。

➢ 必备元素/字段

元数据中必须包含**7**项元素/字段：名称、责任者、主题、类型、格式、唯一标识号、馆藏信息。

1、名称：著录资源的名称，内容为可以概括著录资源内容的词、词组、符号等；

元素修饰词（有则必备）：并列题名、交替题名、其他题名信息、系列正题名

【示例1】

名称：程乙本红楼梦

其他题名信息：桐花凤阁批校本

交替题名：红楼梦

【示例2】

名称：英语求职信写作指南

并列题名：A guide to writing English letters of job-application

➢ 必备元素/字段

2、责任者：创建文献资源内容的主要责任者；

元素修饰词（有则必备）：并列名称 责任方式 创建者说明

【示例】

责任者：安德义

责任方式：主讲

创建者说明：湖北省孔子学会副会长

➢ 必备元素/字段

3、主题：依照CLC（中国图书馆分类法）、CCT（中国分类主题词表）著录描述资源主题内容的受控或非受控的词汇

元素修饰词（必备）：主题词、关键词、分类号

【示例】

主题词：历史发展

关键词：藏书

分类号：G250

➢ 必备元素/字段

4、类型：征集资源的主题名称，等同于全国公共图书馆自建数字资源元数据登记项目规定的“类型”元素中的“内容类型”

【示例】

类型：地方文献

主题名称	代码
其他特色资源	01
地方文献	02
民国文献	03
家谱	04
老照片	05
年画	06
专题视频资料	07
动漫素材	08
非物质文化遗产	09
少数民族资源	10
少年儿童资源	11
历史文化	12
科普	13
网事典藏	14
碑帖	15
图林拾珍	16

➢ 必备元素/字段

5、格式：数字资源对象数据，文本、图像、音频、视频文件的数据格式；

【示例】

格式：TIF

flv

MP3

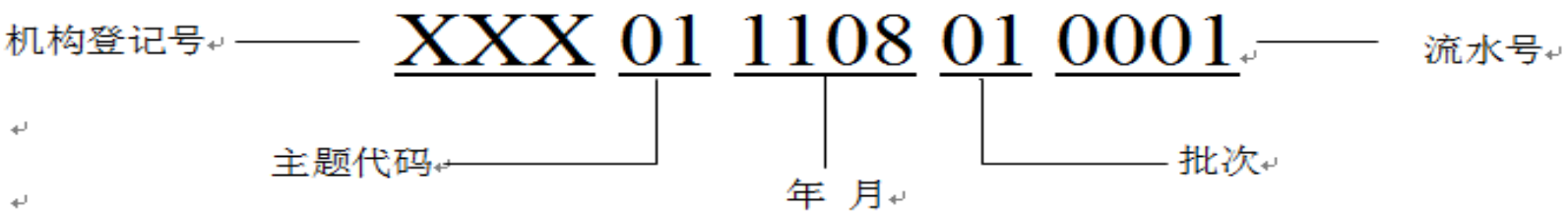
➢ 必备元素/字段

6、唯一标识号：15位记录标识号；

【示例】

唯一标识号： 001021208010001

【示例】



- 必备元素/字段

7、馆藏信息：资源提交方的规范机构名称。

【示例】

馆藏信息： 长春市图书馆

若采用的元数据的格式无法嵌入上述必备元素/字段，可采用元数据补充元素/字段表的形式进行元素/字段补充，该表与规范格式元数据采用唯一标识号（15位记录标识号）关联，采用EXCEL格式文档提交，命名规则与元数据文件名一致（11位，机构登记号（3）—主题代码（2）—年月（4）—批次（2），扩展名为.xls）

元数据补充元素/字段表

唯一标识号	类型	格式	馆藏信息

说明：

1、表头字段可扩展；

2、“唯一标识号”一栏，填写资源的15位记录标识号，同种资源的唯一标识号在规范格式元数据及其补充元素/字段表中要保持一致；

3、“类型”一栏，填写资源所属的主题名称，如“地方文献”；

4、“格式”一栏，填写数字资源对象数据包含的格式，多种格式之间用分号“；”隔开，如“TIF；PDF”；

5、“馆藏信息”一栏，填写机构规范名称，如“长春市图书馆”。

- 国家图书馆在征集项目中使用的DC格式元数据基本框架

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>¶
<DC>¶
    <item>¶
        <名称 并列题名="" 交错题名="" 其他题名信息="" 系列正题名=""></名称>¶
        <创建者 并列名称="" 责任方式="" 创建者说明=""></创建者>¶
        <主题 分类法=""></主题>¶
        <描述 载体形态项="" 附注="" 装订方式=""></描述>¶
        <出版者 出版地=""></出版者>¶
        <其他责任者 责任方式="" 责任者说明=""></其他责任者>¶
        <日期 创建日期="" 获取日期="" 数字化日期=""></日期>¶
        <类型></类型>¶
        <格式></格式>¶
        <唯一标识号></唯一标识号>¶
        <来源></来源>¶
        <语种></语种>¶
        <关联 关联说明=""></关联>¶
        <时空范围 空间范围=""></时空范围>¶
        <权限 授权用户="" 使用权限=""></权限>¶
        <版本></版本>¶
        <价格></价格>¶
        <馆藏信息></馆藏信息>¶
    </item>¶
</DC>¶
```

4. 对象数据加工规范

- 对象数据分为长期保存级和发布服务级，应征单位需一并提交两种级别的数据。若只有一种级别的数据则优先提交长期保存级数据，其次为发布服务级数据。

长期保存级：达到加工技术参数，符合格式要求。用于数字资源长期保存。

发布服务级：长期保存级数字资源的衍生文件。用于数字资源整合发布。

- 对于连续性资源，提交时应尽量保证内容的完整性，避免出现卷册或分集缺失等问题。

4. 对象数据加工规范

➢ （1）文本类

文本类资源数字化规范

字符编码	文件格式	后缀名（小写）
Unicode （无法录入的生僻字、公式、符号等内容用“〓”表示）	txt / doc/html/pdf	txt / doc/htm/pdf
		

- 文本数据内容应忠实于原文献，完整有序；
- 字符的错误率不超过0.3‰

➢ （2）图像类

图书地图字画类图像资源数字化规范

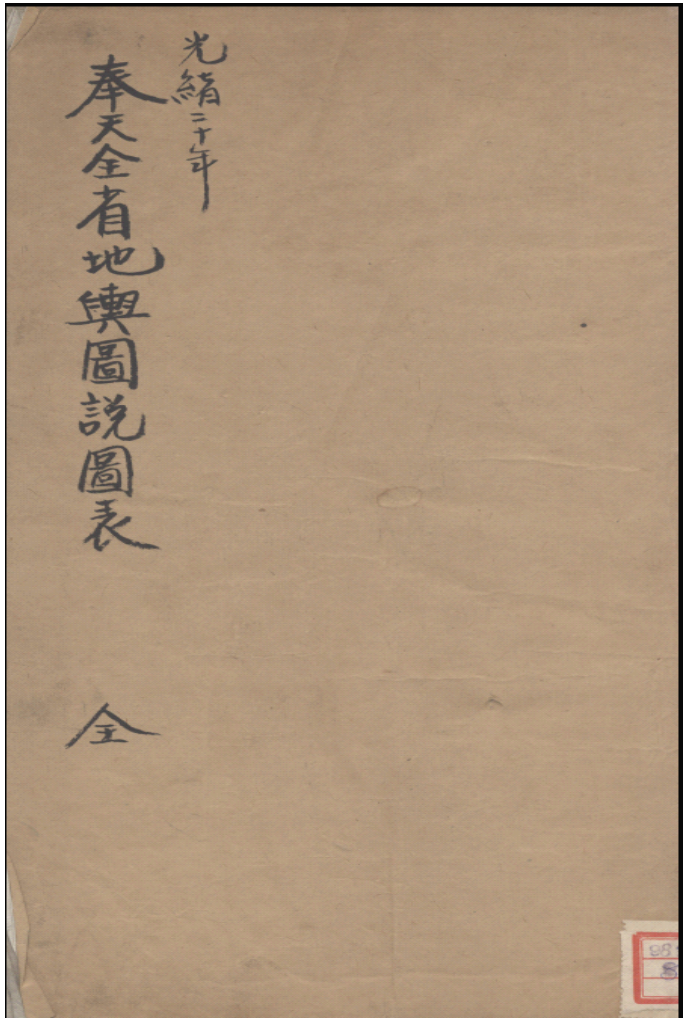
文献类型	应用级别	图像分辨率（dpi）	色彩位深	允许的编辑加工	文件格式	后缀名（小写）
普通书刊报/古籍/字画	长期保存级	>=300	黑白8位，24位	保持原始文件技术参数不变的基础上适当进行纠偏等处理	tiff	tif
	发布服务级	300，图像尺寸不小于800×600		除分辨率调整，其他技术参数保持不变，进行相应图像处理	jpeg pdf	jpg pdf
小幅照片	长期保存级	>=800	8位24位或更高	保持原始文件技术参数不变的基础上适当进行纠偏等处理	tiff	tif
	发布服务级	<=800，图像尺寸不小于800×600	24位	除分辨率调整，其他技术参数保持不变，进行相应图像处理	jpeg	jpg

补充说明：

（1）书、刊、报的发布服务级数据采用pdf格式，字画、图片、照片类的发布服务级数据采用jpg格式

（2）小幅照片可以根据不同的原载体规格选择不同的分辨率标准，

- ✓ >=800dpi （原载体规格<=4’X5’）
- ✓ >=500dpi（原载体规格为4’X5’或8’X10’）
- ✓ >=400dpi（原载体规格>=B5）



File name:	0001.tif
Directory:	D:\工作相关\数字资源征集质检\征集成
Full path:	D:\工作相关\数字资源征集质检\征集成
Compression:	None
Resolution:	300 x 300 DPI Change
Original size:	2040 x 4007 Pixels (8.17 MPixels) (1.96)
Current size:	2040 x 4007 Pixels (8.17 MPixels) (1.96)
Print size (from DPI):	17.3 x 33.9 cm; 6.8 x 13.4 inches
Original colors:	16,7 Millions (24 BitsPerPixel)
Current colors:	16,7 Millions (24 BitsPerPixel)
Number of unique colors:	25166 Auto count
Disk size:	23.39 MB (24,523,020 Bytes)
Current memory size:	23.39 MB (24,522,880 Bytes)
Current directory index:	1 / 14
File date/time:	2012-5-17 / 16:27:50
Loaded in:	63 milliseconds

OK

压缩方式

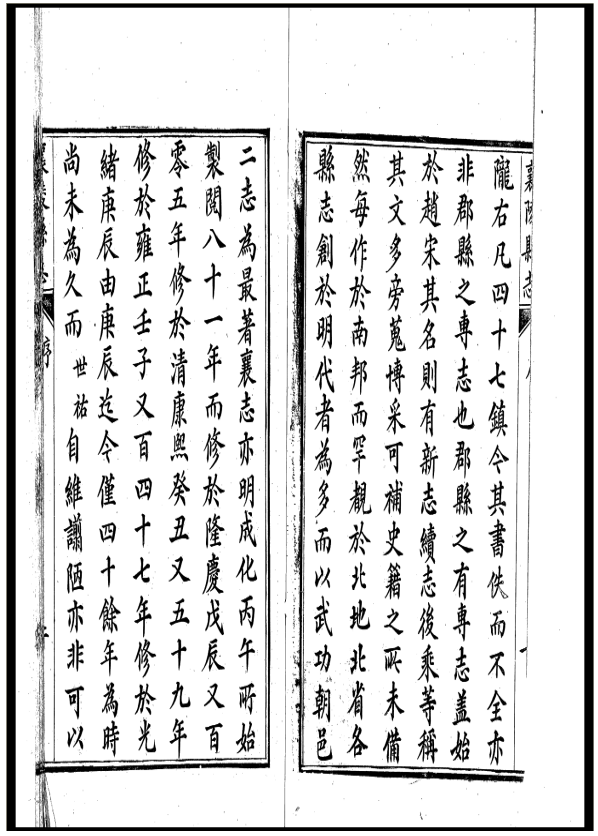
分辨率

色彩位深

胶片类图像资源数字化规范

文献类型	载体规格	资源级别	主要参数					考虑因素及说明
			分辨率 dpi	色彩位深	文件格式	后缀名（小写）	压缩率	
缩微胶片	开窗卡片	长期保存级	300	黑白 8位 24位	tiff	tif	不压缩或无损压缩	注意原始对象的大小，并根据内容选择图像分辨率；本标准所给图像分辨率参考专业缩微胶片扫描仪的各项参数。
		发布服务级	150～300		jpeg pdf	jpg pdf	有损、适度压缩	成比例扩展，锐化，裁切，拼接，纠偏，去噪。浏览服务级的图像压缩后可转换为PDF格式进行发布。
摄影胶片		长期保存级	>=800	8位 24位	tiff	tif	不压缩或无损压缩	
		发布服务级	300～800		jpeg	jpg	有损、适度压缩	允许锐化，裁切，拼接，纠偏，去噪，色调调整等处理

补充说明：所有通过胶片数字化的图像数据均参考本表



File name:	0003.tif
Directory:	D:\工作相关\数字资源征集质检\征集成
Full path:	D:\工作相关\数字资源征集质检\征集成
Compression:	Group 4 Fax Encoding
Resolution:	300 x 300 DPI Change
Original size:	4960 x 3504 Pixels (17.38 MPixels) (1.42)
Current size:	4960 x 3504 Pixels (17.38 MPixels) (1.42)
Print size (from DPI):	42.0 x 29.7 cm; 16.5 x 11.7 inches
Original colors:	2 (1 BitsPerPixel)
Current colors:	2 (1 BitsPerPixel)
Number of unique colors:	2 Auto count
Disk size:	121.75 KB (124,676 Bytes)
Current memory size:	2.07 MB (2,172,528 Bytes)
Current directory index:	3 / 12
File date/time:	2010-4-6 / 10:57:52
Loaded in:	16 milliseconds

OK

压缩方式

图像分辨率

根据分辨率和像素尺寸计算出来的文档尺寸

色彩位深

实物资源（平面成像）的数字化规范

文献类型	处理方式	资源级别	主要参数			文件格式
			短边像素	色彩位深	图形大小	
实物	数码相机拍摄	长期保存级	>2000	24位	>=8M	TIFF
		发布服务级	<=1024	24位	约40K—500K	JPG

➢ （3）音频类

音频资料数字化规范

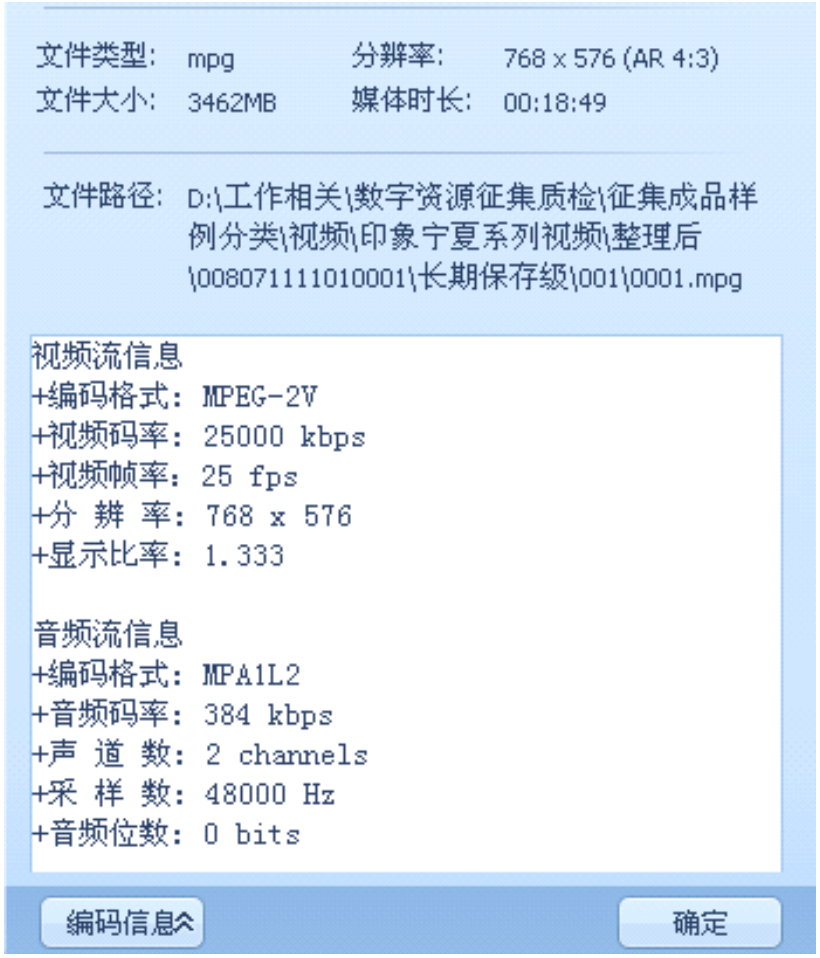
资源类型	资源级别	主要参数			文件格式	后缀名（小写）
		采样率	量化级	通道数		
音频	长期保存级	22 / 128 kHz	16 /24 bit	由原始资料特性决定	wav	wav
	发布服务级	22 / 44 kHz	8 / 16 bit	双声道/单声道	mp3	mp3

➢ （4）视频类

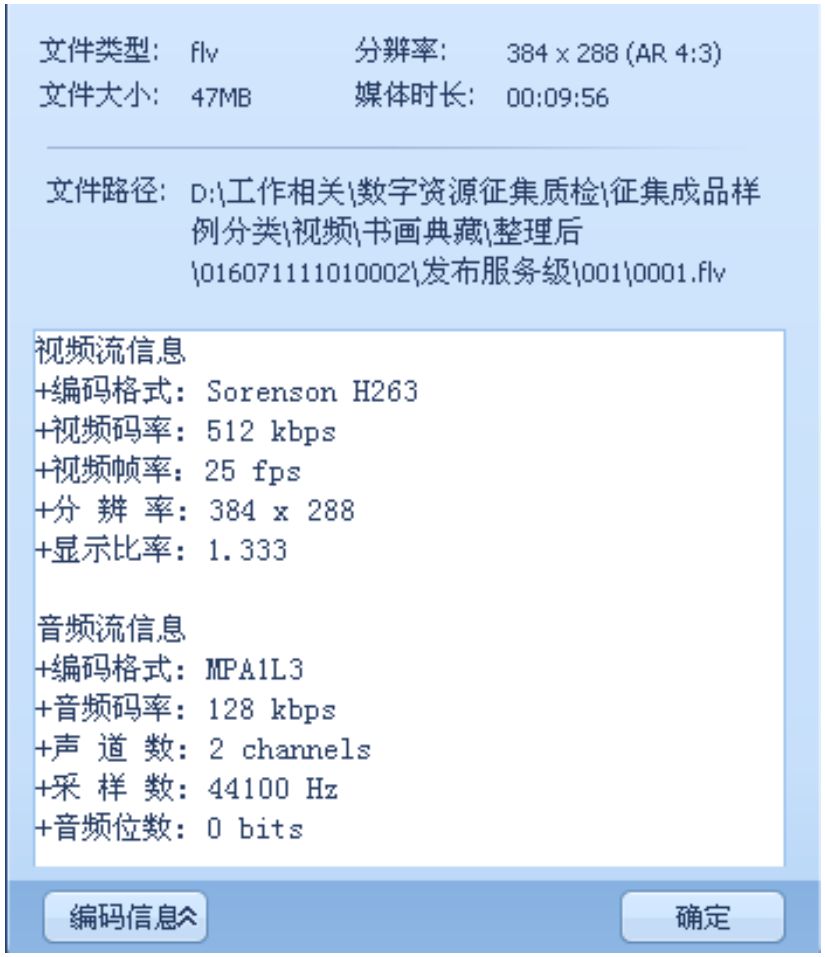
视频资料数字化规范

资源类型	资源级别	主观质量描述	主要参数						
			分辨率	帧数（帧/秒）	视频速率（kbps）	音频速率（kbps）	音频采样	文件格式	后缀名（小写）
视频	长期保存级	高清质量	1920 x 1080	25/30/60	固定码率50Mb/s 或25Mb/s 的可变码率	384	立体声 48 khz	MPEG2编码	mpg avi
		标清质量	720 x 576	25/30	15Mb/s 的固定码率或可变码率	384			
		最低质量	原采集窗口尺寸但不小于352x288	25	5Mb/s 的固定码率或可变码率	224	立体声 48 khz		
	发布服务级	最低质量	不低于352x288	15/25/30	不低于512kb/s	64 ～384k	立体声 44.1/48 khz	mpg4 flv	mpg flv

长期保存级样例参数



发布服务级样例参数



5. 文件存储结构规范

- 元数据及元数据补充元素/字段表文件结构

文件名由4段组成共11位：机构登记号—主题代码—年月—批次，各段的意义详见本文第一部分记录标识号。

文件存储路径为：根目录\文件夹名\

文件夹命名与元数据文件命名规则一致。

【示例】某图书馆（机构登记号为001）2012年8月提交的馆藏特色资源的第一批数据，元数据及元数据补充元素/字段表的存储路径为：

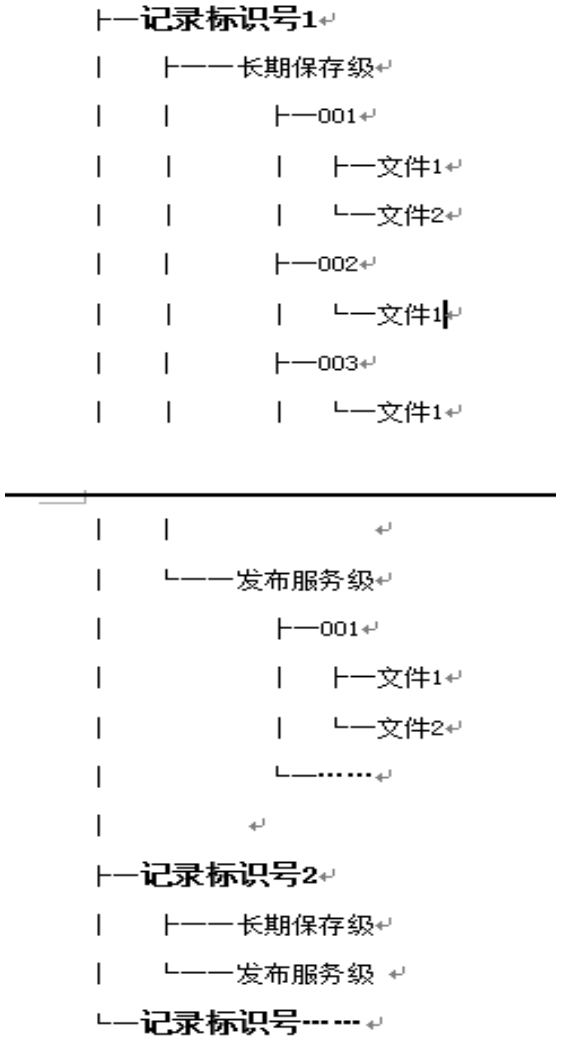
根目录\00102120801 ├──00102120801.iso
└──00102120801.xls

- 对象数据文件结构

对象数据存储路径为：根目录\记录标识号\加工级别\卷册流水号\

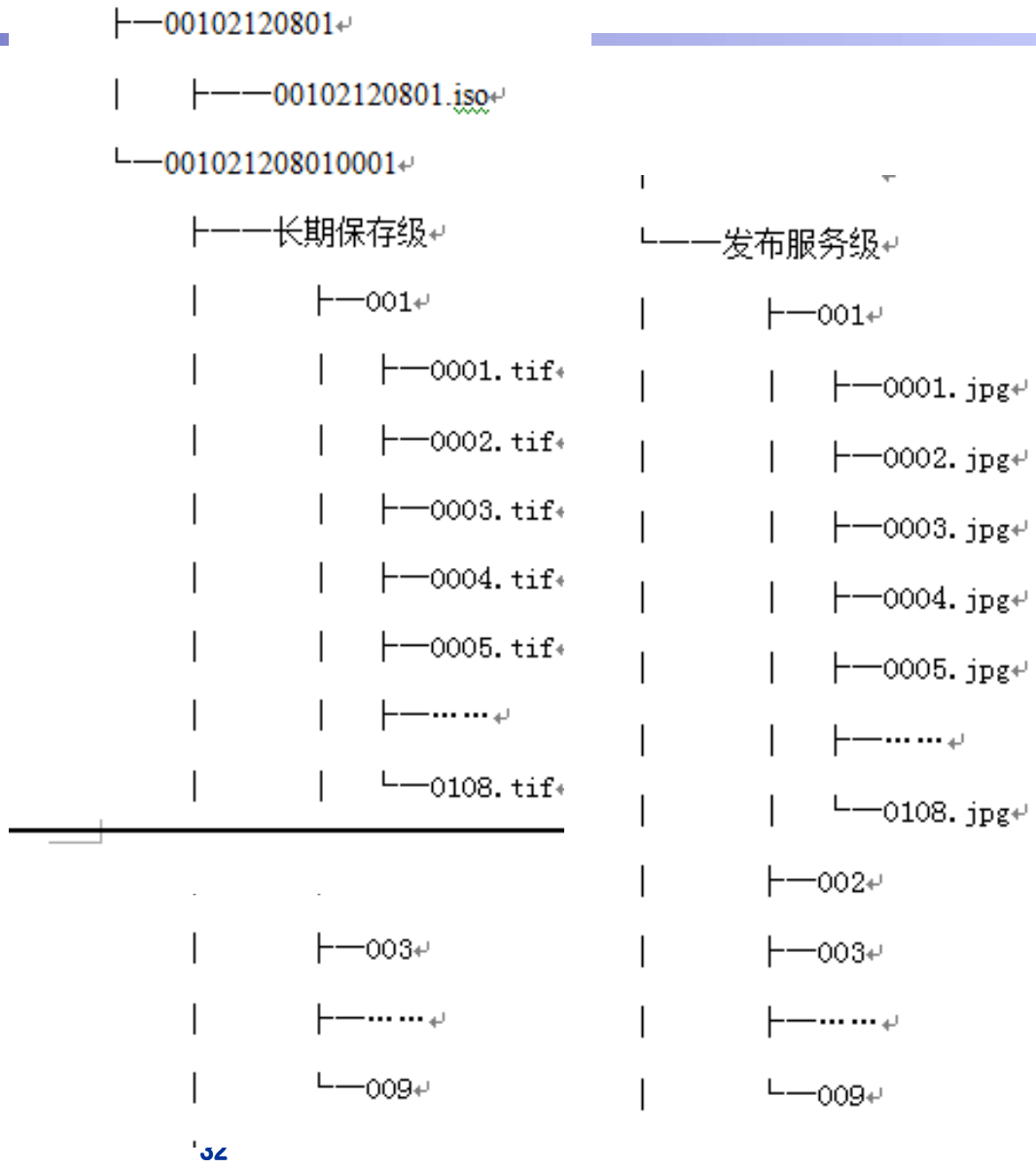
其中对象数据第一级目录为记录标识号，加工级别有两种：长期保存级和发布服务级，卷册流水号3位，从001开始顺序排序。如果该资源非多卷册，那么在加工级别下只有001一个文件夹，文件夹下对应存放着数据文件。

对象数据文件结构示意图如下：



【示例】以地方文献中《汶川縣縣志》为例说明对象数据的存储结构。

地方文献主题代码为02，提交日期为2012年8月，批次为第一批，其中《汶川縣縣志》为本批次中的第一种书，共9卷。则《汶川縣縣志》对应的记录标识号为001021208010001，存储结构如下：



6. 数据提交说明

数据制作单位在提交数据的同时应提交数据总体说明表和数据明细说明表，存放在一个EXCEL文件的多个工作表中，EXCEL文件的命名方式为“机构名称+数据提交说明表+提交年月”，存放在提交介质的根目录下。

【示例】某图书馆（机构登记号为001）2012年8月提交的数据提交说明表的存储路径为：

根目录\某图书馆数据提交说明表201208.xls

“数据总体说明表”为1个工作表，工作表命名为“数据总体说明表”；每个资源包对应1个“数据明细说明表”工作表，工作表命名采用资源包名称。

6. 数据提交说明

数据总体说明表			
制作单位：			提交日期：
序号	资源包名称	主题	来源
1			
2			
3			

说明：

1、“制作单位”一栏，填写机构规范名称，如“首都图书馆”；

2、“提交日期”一栏，填写数据提交国家图书馆的时间，如“2013年4月2日”；

3、“资源包名称”一栏，填写资源包正式名称，并与《国家图书馆数字资源征集协议》中的资源包名称保持一致；

4、“主题”一栏，填写《规范》中“主题代码表”中与该资源包对应的主题名称；

5、“来源”一栏，选择填写下列序号：①普通书刊报；②古籍；③字画；④小幅照片；⑤缩微胶片；⑥摄影胶片；⑦实物；⑧音频；⑨视频；⑩其它（自定义注明）。资源遵循的数字化标准应符合其注明的来源，单一数据类型资源包选填一项，多种数据类型资源包可选填多项。

数据明细说明表																
资源包名称：							制作单位：							提交日期：		
序号	记录标识号	题名	元数据	对象数据长期保存级						对象数据发布服务级						备注
				文件格式	加工参数	册（件）数	文件数量（个）	存储量（MB）	时长（分）	文件格式	加工参数	册（件）数	文件数量（个）	存储量（MB）	时长（分）	
1																
2																
3																

说明：

1、“记录标识号”一栏，填写资源的15位记录标识号，同种资源的唯一标识号在规范格式元数据及其补充元素/字段表中要保持一致；

2、“题名”一栏，填写每种资源的名称，并与元数据中“名称”项保持一致；

3、“元数据”一栏，填写该种资源是否具备元数据，是则画“√”；

4、“文件格式”一栏，参考《规范》里对象数据规范一节中各类资源的文件格式要求，填写本种资源包含对象数据的格式，如TIF、PDF等；当格式不唯一时，应根据不同格式分别统计文件数量、存储量或时长信息，并进行填写；

5、“加工参数”一栏，图像类资源可填写“分辨率”、“图像尺寸”信息；视频类资源可填写分辨率、码流信息；

6、“册（件）数”一栏，图像类资源以册/件为单位填写资源对应实体的册/件数，视频类资源填写场次或集数；

7、“文件数量（个）”一栏，以个为单位分格式填写资源有效文件的数量；

8、“存储量（MB）”一栏，以MB为单位分格式填写资源的实际存储量；

9、“时长（分）”一栏，以分钟为单位填写音视频类资源的有效播放时间长度，其他类型资源可不填。

【示例】某图书馆于2013年4月2日向国家图书馆提交资源包2个，其数据总体说明表和数据明细表示例如下：

数据总体说明表			
制作单位：　某图书馆			提交日期：2013年4月2日
序号	资源包名称	主题	来源
1	史志类线装古籍	地方志	②
2	非物质文化遗产资源	非物质文化遗产	④⑨

②古籍④小幅照片⑨视频

数据明细说明表																
资源包名称：史志类线装古籍						制作单位： 某图书馆									提交日期：2013年4月2日	
序号	记录标识号	题名	元数据	对象数据长期保存级						对象数据发布服务级						备注
				文件格式	加工参数	册（件）数	文件数量（个）	存储量（MB）	时长（分）	文件格式	加工参数	册（件）数	文件数量（个）	存储量（MB）	时长（分）	
1	***02130 4010001	潘陽縣志	√	TIF	400 DPI	6	743	3389		PDF		6	743	399		

7. 数据完整性规范

征集数字资源应包含完整的元数据、对象数据和数据提交说明表，存储结构规范。要求元数据和对象数据对应关系清晰明确，不可出现元数据与对象数据无法对应问题。

元数据与对象数据的对应关系清晰明确包含2层含义：

- 每条元数据和对应的对象数据之间存在明确的计算机可识别的对应关系，即一条元数据中嵌入的“记录标识号”应与其对应对象数据的一级目录保持一致；
- 通过记录标识号关联的元数据与对象数据对应关系正确，不可出现元数据描述不准确或“张冠李戴”的情况。

第二部分　人文悦读专题库数据标准规范

- 为确保中文图书数字化图像品质、标引数据质量和成品数据的管理，制定《人文悦读专题库数据标准规范》。本规范规定了中文图书数字化的工作内容、质量要求、技术规格和验收标准。制作单位应依据本规范所规定内容和要求进行中文图书数字化加工；我馆在验收数字化成果时，将依照本规范进行检验。

- 本规范针对当前中文图书的一般情况制定。由于我国目前图书出版情况复杂，选择印刷、排版等标准不统一，会存在本规范未涉及内容。在中文图书数字化加工过程中如遇到超出本规范内容，加工单位需及时与本馆沟通，双方协商解决。

1、图书整理

1.1　应征馆提交人文悦读专题库资源建设申报书。

1.2　数据查重

为避免重复建设，国图负责对申报的中文图书加工书目进行数据查重处理，并将确认的申报书反馈给申报单位。

1.3　MARC数据导入：生成中文图书数据库之图书基本信息表（book表）。（附件一 1）（附件指本规范附件，以下同）

序号	中文名称	字段名称	对应书目数据（MARC）内容
1	加工编号	book_id	
2	分类	cat_id	第一个 690 字段$a
3	书名	book_name	200 字段$a $h $i $e
4	作者	author	200 字段$f
5	出版社	pub_house	210 字段$c
6	出版时间	pub_date	210 字段$d
7	ISBN 号	isbn	010 字段$a
8	001	record_id	001 字段
9	条码号	barcode	

注：图书基本信息表（book 表）除加工编号和条码号外，各字段内容原则上均取自书目数据（MARC）。对于相同“001”、不同“条码号”的图书，要进一步标注“书名”字段，标注内容用“（）”括起来。如：（上册）、（下册）。

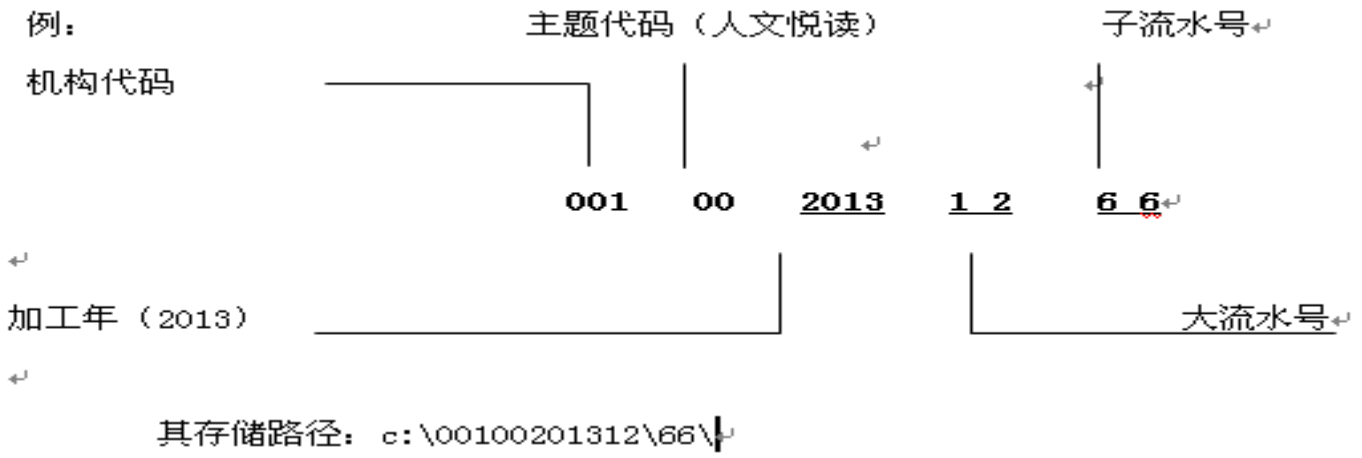
➢图书加工编号（book_id）

图书数字化加工过程中一册图书的唯一标识，它由13位数字和下划线组成。

机构登记号（3位）、主题代码（2位）、加工年（4位）、大流水号（2位）、子流水号(2位)。

本规范针对人文悦读专题库，其机构登记号为国家图书馆统一分配，主题代码为00，加工年为公元年（如2013），大流水号为2位数字01-99，子流水号为为2位数字01-99。

注意：一种多册，按单册扫描，在book表中：书名+（单册信息，如上册，或分册信息）



2、数据标引

2.1 目录信息标引。图书目录页内容及图像文件的字母和数字部分

图书目录信息表

序号	中文名称	字段名称	备注
1	加工编号	book_id	
2	标引序号	serial_num	
3	章节号	chapter_num	
4	章节名	chapter_name	
5	作者	author	
6	页码	page_num	客观著录，如实反映目录页原貌（可为空）
7	绝对页码	ppage_num	文件名数字部分
8	页位置	page_place	文件名字母部分
9	属性	page_prop	1）“目录”属性为“1”；2）“无目录”属性为“2”； 3）每册书除第一条目录外，其余记录的属性默认为“0”

➢图书目录信息表示例

book_id	serial_num	chapter_num	chapter_name	author	page_num	ppage_num	page_place	page_prop
00100201312_66	1	目录						1
00100201312_66	2	第1章	绪论		1	1	T	0
00100201312_66	3	1.1	软件的发展		1	1	T	0
00100201312_66	4	1.2	软件的基本概念		3	3	T	0
00100201312_66	5	1.2.1	软件的特点		4	4	T	0
00100201312_66	6	1.2.2	软件的分类		5	5	T	0
00100201312_66	7	1.2.3	软件构件		6	6	T	0
00100201312_66	8	1.2.4	软件应用的分类		6	6	T	0
00100201312_66	9	1.2.5	软件技术的发展趋势		7	7	T	0
00100201312_66	10	1.3	软件工程		9	9	T	0

2、数据标引

2.2 图书版权页标引。图书版权页内容。

图书版权信息表

序号	中文名称	字段名称	备注
1	加工编号	book_id	
2	书名	book_name	
3	作者	author	
4	001	record_id	
5	版权页位置	copyright_place	记录版权页文件名

【示例】

book_id	book_name	author	record_id	copyright
00100201312_01	2011年国家司法考试	杨艳霞，徐金桂，杨雄编著	005273254	B00002_00

2、数据标引

2.3　记录图书不带页号插图信息，生成中文图书数据库之图书插页信息表。

图书插页信息表

序号	中文名称	字段名称	备注
1	加工编号	book_id	
2	插页前正文页号	prior_text_page	图书印刷页码
3	插页数量	inset_num	

【示例】

book_id	prior_text_page	inset_num
0010020131 2_66	128	4
0010020131 2_66	352	4
0010020131 2_66	512	4
0010020131 2_66	672	4
0010020131 2_66	832	4
0010020131 2_66	960	4

2、数据标引

2.4 记录图书缺页信息，生成中文图书数据库之图书缺页信息表。

图书缺页信息表

序号	中文名称	字段名称	备注
1	加工编号	book_id	
2	缺页前正文页号	start_text_page	图书印刷页码
3	缺页数	lostpage_num	

【示例】

book_id	start_text_page	lostpage_num
00100201312 66	41	1
*		

2、数据标引

2.5 记录图书封面、前附页、目录、正文等结构信息。

序号	中文名称	字段名称	备注
1	加工编号	book_id	
2	封面页数	fore_cover_num	
3	目录前，前附页数	preface1_num	
4	目录前，前附页起始页号	preface1_start_page	
5	目录页数	content_num	
6	目录起始页号	content_start_page	
7	目录后，前附页数	preface2_num	
8	目录后，前附页起始页号	preface2_start_page	
9	正文页数	text_num	
10	正文起始页号	text_start_page	
11	后附页数	appendix_num	
12	后附页起始页号	appendix_start_page	
13	封底页数	back_cover_num	

注：起始页号均为图书印刷页码

【示例】

book_id	fore_cover_num	preface1_num	preface1_start_page	content_num	content_start_page	preface2_num	preface2_start_page	text_num	text_start_page	appendix_num	appendix_start_page	back_cover_num
0010020 1312_01	2	4		2	1	0		692		6		2

2、数据标引

2.6 记录扫描分辨率、压缩因子、文件数量、存储量等信息

序号	中文名称	字段名称	备注
1	加工编号	book_id	
2	书名	book_name	
3	扫描分辨率	dpi	
4	压缩因子	comp_factor	
5	灰度页数量	grey_num	
6	彩色页数量	col_num	
7	TIFF 数量	tiff_num	
8	PDF 数量	pdf_num	
9	TIFF 存储量	tiff_mb	存储单位： MB
10	PDF 存储量	pdf_mb	存储单位： MB
11	典藏级硬盘位置	hdA_place	典藏级硬盘财产号
12	服务级硬盘位置	hdB_place	服务级硬盘财产号

【示例】

book_id	book_name	dpi	comp_factor	grey_num	col_num	tiff_num	pdf_num	tiff_mb	pdf_mb	cdA_place	cdB_place	hdA_place	hdB_place
00100201312_01	2011年国家司法考试名师教案 卷二	300	灰度=33-34,彩色=99	706	2	708	708	4464.8	133.42	0112-1-A100953、0112-1-A100954	0112-1-B100018	200933817	200933871

3、扫描制作

➢**3.1** 扫描前根据国际色彩协会ICC标准，做基本的色彩校正，及针对各类型图书进行色彩校正。

➢**3.2** 图书为全书逐页扫描方式，依照“扫描规格”和命名规则进行数字加工。

灰度方式扫描

色彩位深：8 位

◆分辨率：300 dpi；小于5号字体用400 dpi

◆JPEG2000压缩因子：20（根据图像规格、颜色、数据量适当调整）

◆档案典藏级格式：TIFF 不压缩

◆发布服务级格式：PDF（经过JPEG2000压缩后，再做格式转换）

◆彩色方式扫描

色彩位深：24 位

◆分辨率：300 dpi；小于5号字体用400 dpi

◆JPEG2000压缩因子：120（根据图像规格、颜色、数据量适当调整）

◆档案典藏级格式：TIFF 不压缩

◆发布服务级格式：PDF（经过JPEG2000压缩后，再做格式转换）

3、扫描制作

3.3 扫描文件命名规则

➢文件名后缀为小写字母

- ✓前封（含封一、封二），扫描文件名为Axxxxx_00，其中xxxxx为5位数字，按原书顺序依次排序
- ✓前附页，目录页之前的前附页扫描文件名为Bxxxxx_00，其中xxxxx为5位数字，按原书顺序依次排序。
- ✓目录页之后的前附页扫描文件名为Dxxxxx_00，其中xxxxx为5位数字，按原书顺序依次排序。
- ✓目录页，扫描文件名为Cxxxxx_00，其中xxxxx为5位数字，按原书顺序依次排序。
- ✓正文，有页码的正文扫描文件名为Txxxxx_00,其中xxxxx为5位数字,与原书页号一致，按原书顺序依次排序。
- ✓正文中插页扫描文件名为Txxxxx_yy，其中xxxxx为5位数字,表示插页的前一页顺序号,yy为数字，表示插页，并按原书顺序依次排序。
- ✓后附页，扫描文件名为Yxxxxx_00，其中 xxxxx 为5位数字，按原书顺序依次排序。
- ✓后封（含封三、封四），扫描文件名为Zxxxxx_00，其中xxxxx为5位数字,按原书顺序依次排序。

4、数据检查

- **4.1** 图像文件（各种格式）放大到1：1状态，逐页检查。检查文件是否有太淡、太浓、黑边、污点、歪斜、模糊（马赛克等）或图像内容不完整等现象，不符合图像质量要求应进行图像校正或重新扫描。
- **4.2** 发现文件漏扫时，应及时补扫并正确插入图像。
- **4.3** 检查是否符合扫描规格要求。
- **4.4** 所有文件保存位置正确，可以正常地打开、显示。
- **4.5** 检查图像页码是否连续，不得跳页。
- **4.6** 文献以册/件为单位检查标引（描述和管理）数据是否完整、准确。
- **4.7** 按照命名规则，检查目录、文件、数据库、文档、介质等名称是否正确。
- **4.8** 检查各类说明、统计、验收等文档是否齐全。

5、数据提交（交接）

提交内容详见“规范”主要文件有：

- **元数据（书目数据）**
- **数据库文件，**Microsoft Access

  图书基本信息表（book表）

  图书目录信息表（catalog表）

  图书版权信息表（copyright表）

  图书插页信息表（inset表）

  图书缺页信息表（lostpage表）

  图书结构信息表（struct表）

  图书加工信息表（process表）

- **中文图书数据说明文件，**Microsoft Excel

  数据总体说明，保存级对象数据硬盘存储清单，发布级对象数据硬盘存储清单，图书单册数据量统计表

- 图像文件以及**每册图书的说明文件 （bookinfo.txt）**

6、质量要求

- 1　中文图书均采用ASCII码进行标引，无法录入的生僻字、公式、符号等内容用“〓”表示。同时将“〓”所对应图像文件保存在档案典藏级数据内，以“〓”命名的文件夹内。
- 2　标引信息应严格按照原书实际内容进行描述，真实反映图书原貌。各类链接准确无误。
- 3　图书封面和各种内页的扫描方式正确，不得随意改变。
- 4　每本图书相同扫描方式生成的图像保持相同的清晰度，不得有失真现象。
- 5　图像歪斜度不可以超过一度；去除与文字、图片、版式无关的杂点、黑边、污迹等信息。
- 6　拼接图像接缝处无错位、无缝吻合，不应出现白边和内容缺失，没有明显的歪斜。

6、质量要求

- 7　长期保存级（档案典藏级）图像，数字文件保持原采集信息，以无损压缩和不压缩标准格式存档。发布服务级的图像，为有损压缩图像格式，在转换工作中应在图像轮廓清晰可读的前提下（可放大到实际尺寸检查判定），尽量减小数据量。
- 8　图像名称必需正确，同一数据流水号不得有跳号情况，按顺序排列命名；图像文件的排列顺序应与原书一致。
- 9　数据库字段、说明文件、各类表格等内容严格按照附件规定和样例版式，加工方不得擅自更改。
- 10 介质中不得存放与备份内容无关的文件、严禁携带病毒、严禁浪费介质空间。
- 11在加工过程中如遇特殊情况，应及时与我馆沟通，协商解决并做备忘记录。

7、数据验收

➢**验收内容**

中文图书数据库，标引数据，扫描图像文件，说明文件等成品数据的质量和数量；保存介质的品质、数据结构合理以及内容的完整。

➢ **验收标准**

✓**1** 数据验收将采取抽样检验，抽检样本数为送检成品的 3%。验收人员随机抽选。

✓**2** 送检数据内容与《中文图书验收数据提交单》相互匹配，各种格式数据和文档一一对应，不可夹杂无关文件。

✓**3** 各种标引、说明文件的文字、符号、版式、位置和文件名称准确，其综合错误率不超过0.3‰。

✓**4** 图像数据扫描方式、扫描规格、文件格式、文件命名、图像处理、压缩方式等符合要求，其综合错误率不超过1‰。

7、数据验收

➢ 验收标准

✓5　成品数据备份数量、保存介质命名、数据存放方式、数据内容符合规范要求，且各类型保存介质内无坏死文件、不准携带病毒，错误率为0。

✓6　达到验收标准的数据视为合格，合格范围内检查出的问题进行修正；未达到验收标准的数据由加工单位重新对进行数据的检查、修改、重扫等返工工作。

✓7　验收人员撰写《数字资源验收记录单》，将验收结果及时通知加工单位，并监督处理过程和结果。

目 录

- 加强数据管理工作
  - ✓ 做好本单位的数据管理工作，避免资源重复申报
    例：某机构在两年中申报的连续性资源包有约30%的重复数据
  - ✓ 做好记录标识号的管理工作，避免重复
    例：存在多种数据使用同一记录标识号的情况
- 在正确理解规范的前提下制作数据
  - ✓ 避免进行无意义的数据转换
    例：将低质量的数据转换为高质量的数据，将发布服务级数据转换为长期保存级数据
  - ✓ 避免进行错误的格式转换操作
    例：直接修改文件后缀名，如将.jpg后缀直接改为.tif
  - ✓ 相同资源包的数据加工方式和标准要统一
    例：一个资源包的同种加工级别之间存在加工方式、技术参数、格式等方面的差异

- 在提交数据前做好相关的自检工作
  - ✓ 检查元数据、对象数据、说明文件是否齐备，其互相之间关联关系是否正确
    - 例：元数据错误，如无内容、著录错误、必备字段缺失、元素命名不规范
    - 例：对象数据错误，如内容缺失或重复、混入无关内容、多册图书未分册存储、逻辑颠倒
    - 例：关联关系错误，如元数据和对象数据无数据层面对应关系，依赖先后顺序关联；元数据所描述内容与对象数据内容不一致
  - ✓ 检查是否存在无关冗余文件

谢　谢！