



数字图书馆推广工程2014年度数字资源联合建设培训班

推广工程数字资源联合建设 地方文献数字化项目建设方案解读

国家图书馆 数字资源部

韩萌

提 纲

文献数字化相关知识

2014年项目方案解读





数字图书馆推广工程2014年度数字资源联合建设培训班

推广工程数字资源联合建设 地方文献数字化项目建设方案解读

国家图书馆 数字资源部

韩萌

一、文献数字化相关知识



文献数字化简介

- 文献数字化是利用现代信息技术对传统介质的文本(包括印刷图书)、图形、图像、声音、视频等信息进行格式转换和压缩处理,使其转化为数字信息并存储,然后以计算机技术和互联网为依托进行信息传递,实现检索,为读者提供信息服务。

实体文献



电子文献





实体文献

实物文献

- 甲骨
- 竹简
- 印章
- 石碑
- 木版
- 铜板
-

纸本文献

- 印刷型
- 手工型

缩微文献

- 缩微胶卷
- 缩微平片

视听文献

- 录音带
- 唱片
- 录象带
- 视盘

电子文献

- 网络资源
- 实体资源



电子文献

文本数据

- TXT
- XML
- HTML
- PDF

图像数据

- TIFF
- JPEG
- JPEG2000
- PDF

音频数据

- MP3
- WAV

视频数据

- MPEG-1
- MPEG-2
- MPEG-4



文献数字化简介

常用设备

扫描仪



数码相机



数码摄像机



声卡



视频卡





文献数字化简介

高清扫描仪



高速扫描仪



零边距扫描仪



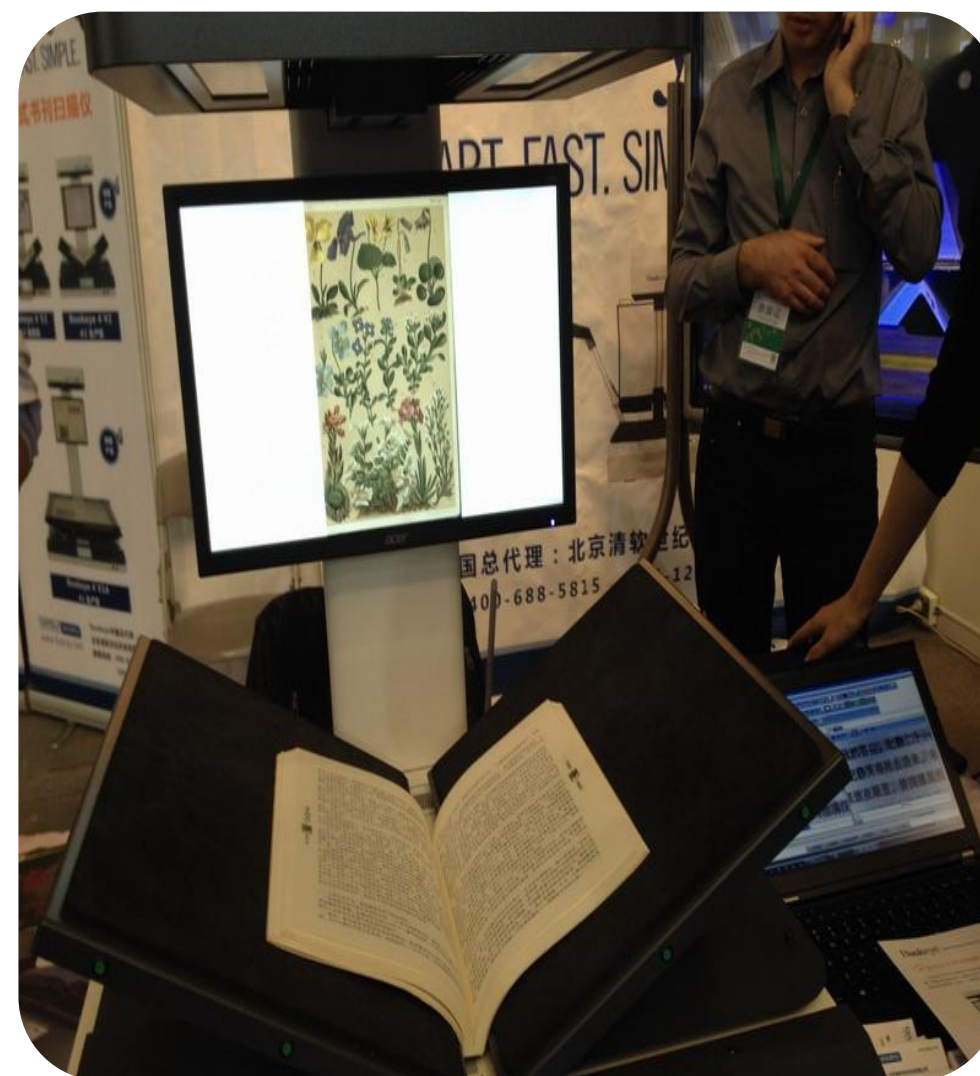


文献数字化简介

缩微胶片扫描仪



缩微胶片扫描仪





文献数字化意义

保存需求

- 残破的文献
- 具有历史意义的文献
- 具有收藏价值的文献

浏览使用需求

- 扩充数字资源存储量
- 方便用户浏览
- 需求量较大的文献



文献数字化方案制定

方案制定

自主型

外包型

合作型

外购型



数字图书馆推广工程2014年度数字资源联合建设培训班

推广工程数字资源联合建设 地方文献数字化项目建设方案解读

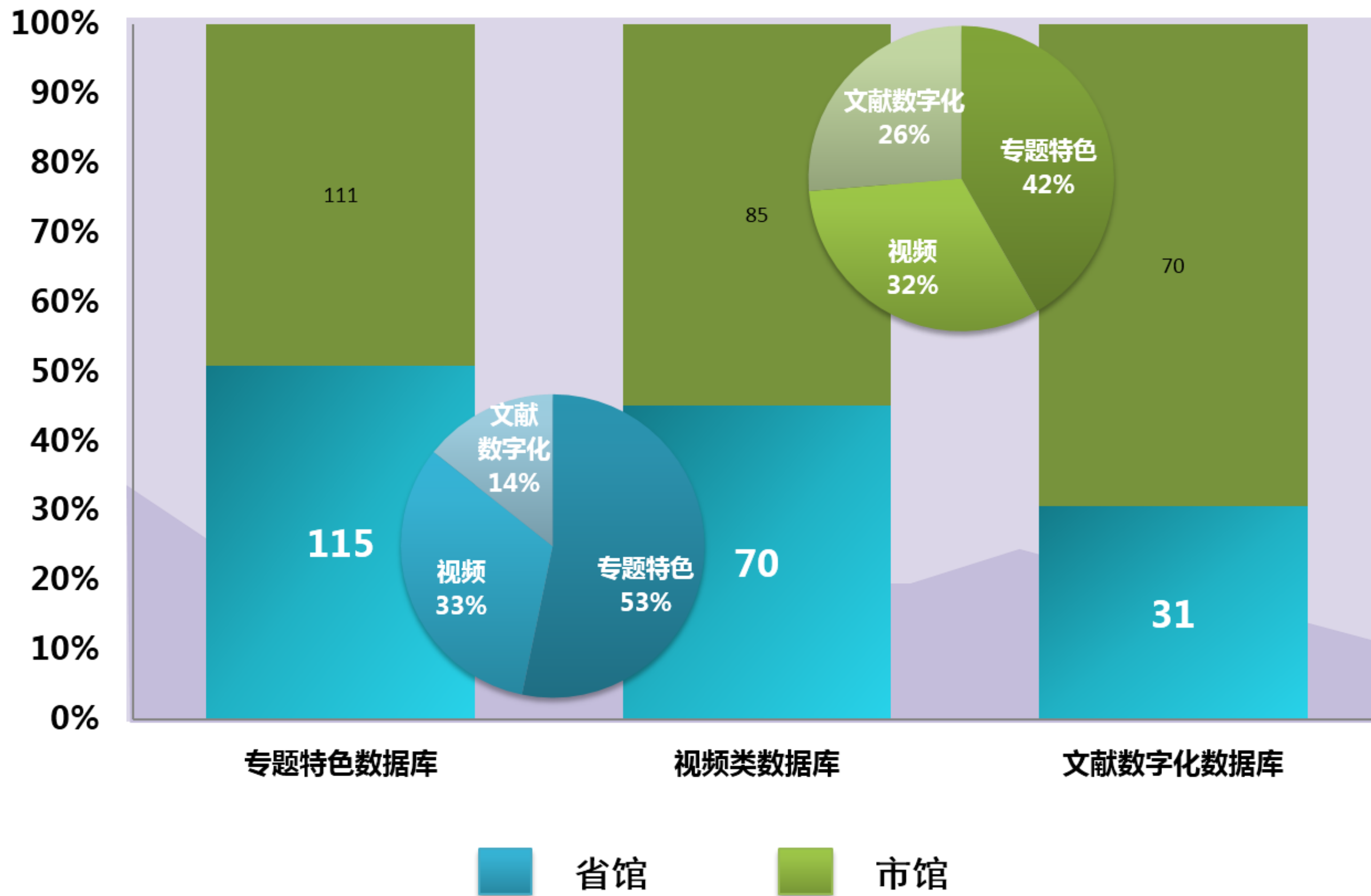
国家图书馆 数字资源部

韩萌

二、2014年项目方案解读



地方文献数字化项目建设背景





地方文献数字化项目工作流程





地方文献数字化项目工作流程



文献目录确认

文献选取

时间：1949年之后

内容：地方特色文献
包括不限于地方志，
行业志。

版权：公有领域或者
自有版权优先。



文献详单 提交



查重处理



返回确认书



地方文献数字化加工

- 重点建设1949年以来的地方文献，包括但不限于行业志、地方志等资源。每个省馆数字化加工不少于10000页。

数字化内容

- 精选体现本地区地域特色，有较高文献价值和历史价值的精品馆藏资源。

版权

- 筛查公有领域版权作品，积极与利益相关方协商，充分利用项目经费，尽可能解决版权问题。



- 一般地说，一个数字对象就是一个具有唯一标识的数字资源。一个数字对象的组成包含三个要素：





应用级别

文本数据	图像数据
<ul style="list-style-type: none">■ TXT■ XML■ HTML■ PDF	<ul style="list-style-type: none">■ TIFF■ JPEG■ JPEG2000■ PDF

● 长期保存级 (A)

✓ 建议以TIFF格式存储，不压缩或无损压缩

● 复制加工级 (P)

✓ 建议以JPG或JPEG2000，有损压缩

● 发布服务级 (D)

✓ L (Large, 大图)

✓ M (Medium, 中图)

✓ S (Small, 小图)





地方文献数字化项目工作流程



- 数字资源对应源文献的内容及特征的描述。建议选用CNMARC (ISO2709) 格式或DC元数据核心元素集。

名称	标签	名称	标签
Title	题名	Creator	创建者
Contributor	其他责任者	Publisher	出版者
Subject	主题	Description	描述
Date	日期	Type	类型
Format	格式	Identifier	标识符
Source	来源	Language	语种
Relation	关联	Coverage	时空范围
Right	权限		

Right	权限		
Relation	关联	Coverage	时空范围
Source	来源	Language	语种

CNMARC

新版中国机读目录格式使用手册

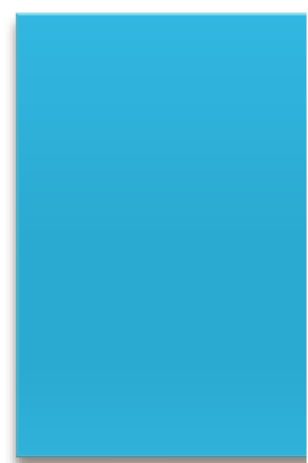
DC元数据

GB/T25100-2010信息与文献：都柏林核心元数据元素

元数据



地方文献数字化项目工作流程



图书整理



扫描制作



数据标引



质量检查



数据验收



成品移交





地方文献数字化项目工作流程



图书整理

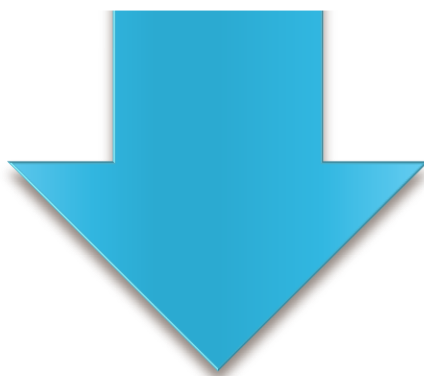
扫描制作

数据标引

质量检查

数据验收

成品移交



图书基本信息表

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	<u>cat_id</u>	第一个690字段\$a
3	书名	<u>book_name</u>	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	



地方文献数字化项目工作流程



图书整理

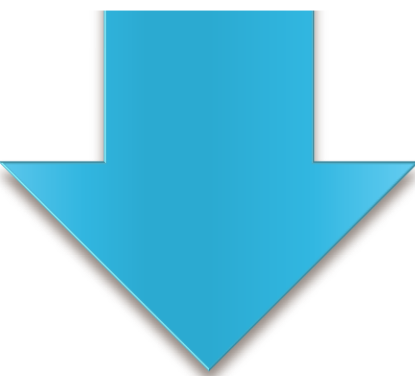
扫描制作

数据标引

质量检查

数据验收

成品移交



色彩校正

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。

文献扫描

- 黑白页和灰度页用灰度方式扫描。
- 彩色页用彩色方式扫描。

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）



图书整理

扫描制作

数据标引

质量检查

数据验收

成品移交

- **前封（含封一、封二）**
 - 扫描文件名为Axxxxxx_00，其中xxxxxx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Bxxxxxx_00，其中xxxxxx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Dxxxxxx_00，其中xxxxxx为5位数字，按原书顺序依次排序
- **目录页**
 - 扫描文件名为Cxxxxxx_00，其中xxxxxx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Txxxxxx_00,其中xxxxxx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Txxxxxx_yy，其中xxxxxx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Yxxxxxx_00，其中xxxxxx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Zxxxxxx_00，其中xxxxxx为5位数字,按原书顺序依次排序。



地方文献数字化项目工作流程



图书整理

扫描制作

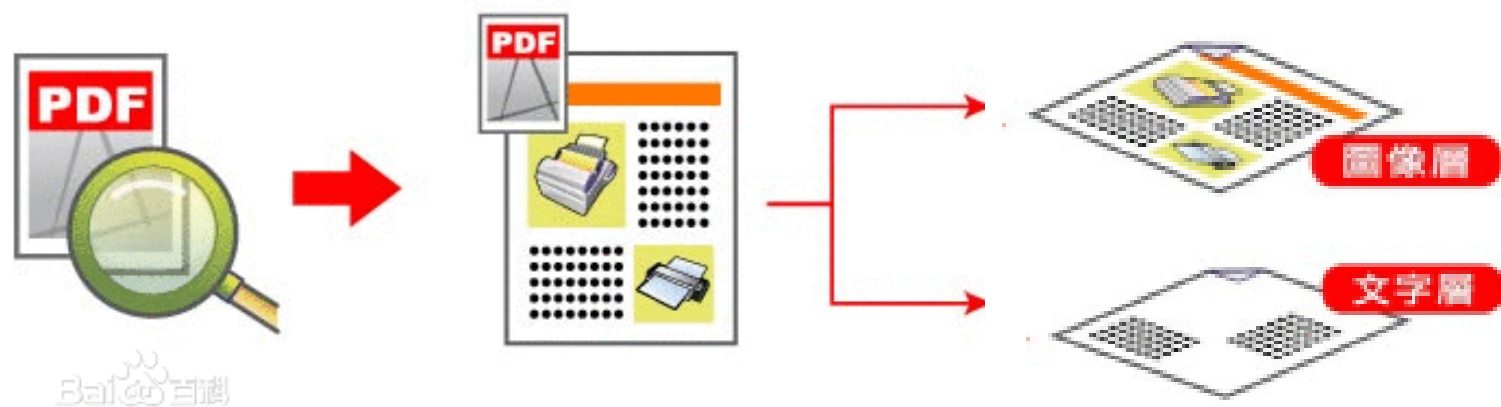
数据标引

质量检查

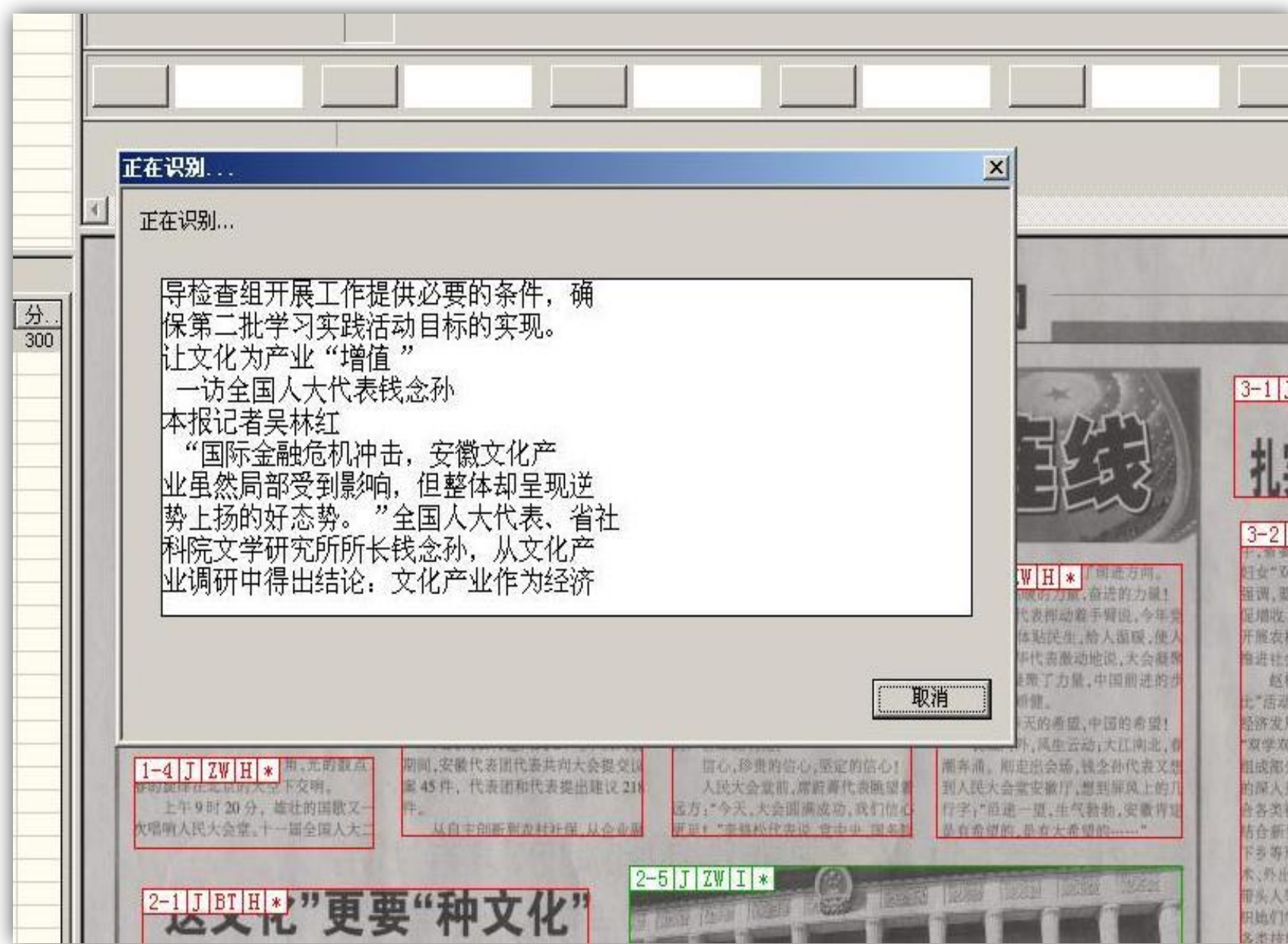
数据验收

成品移交

双层PDF



图像文本转换





地方文献数字化项目工作流程



图书整理

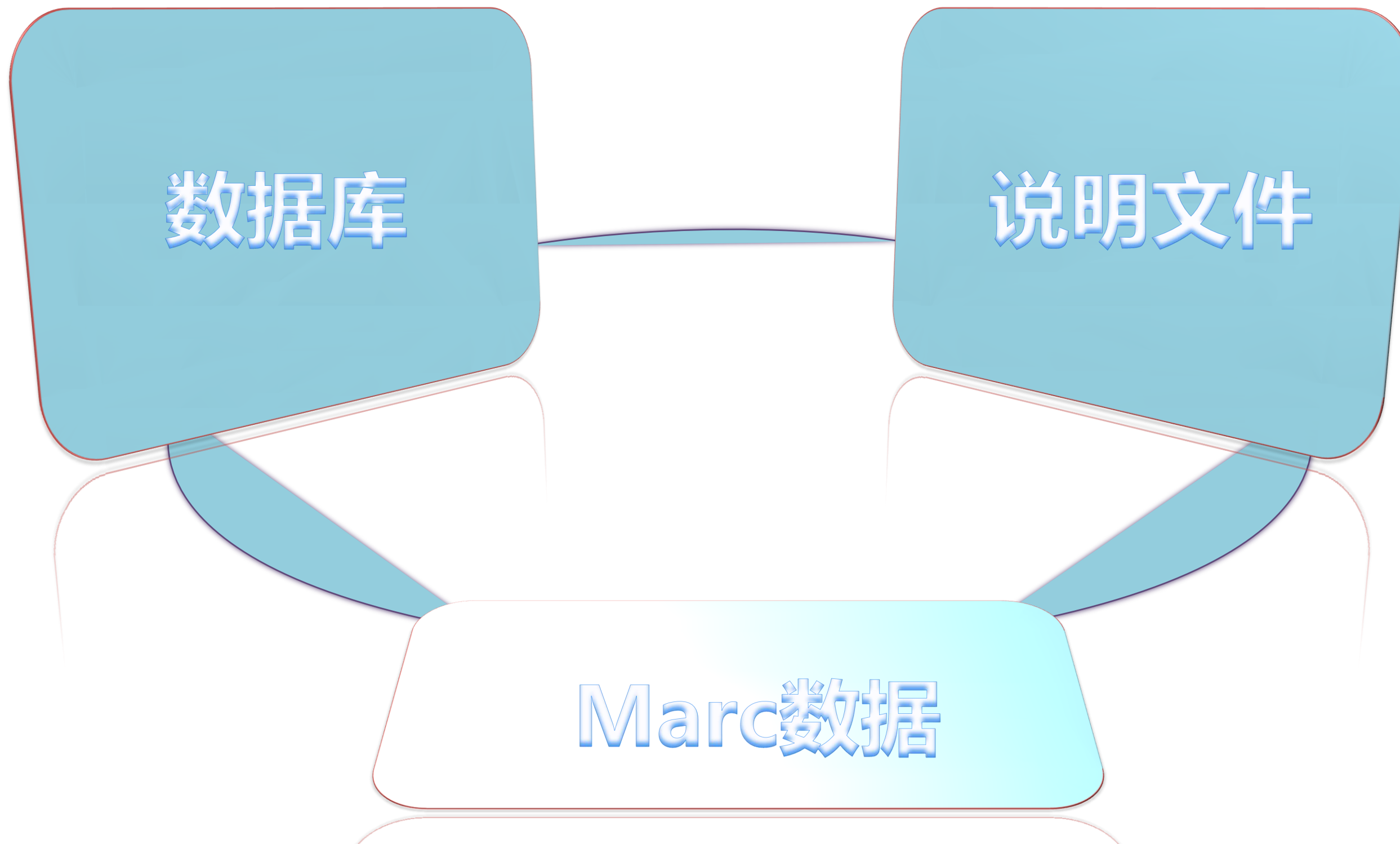
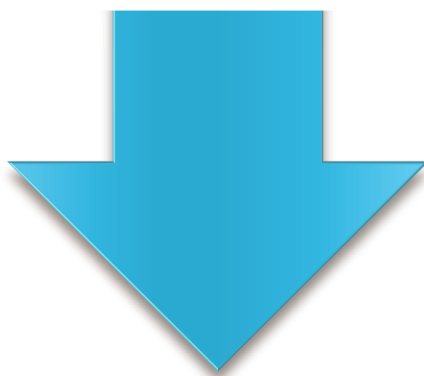
扫描制作

数据标引

质量检查

数据验收

成品移交

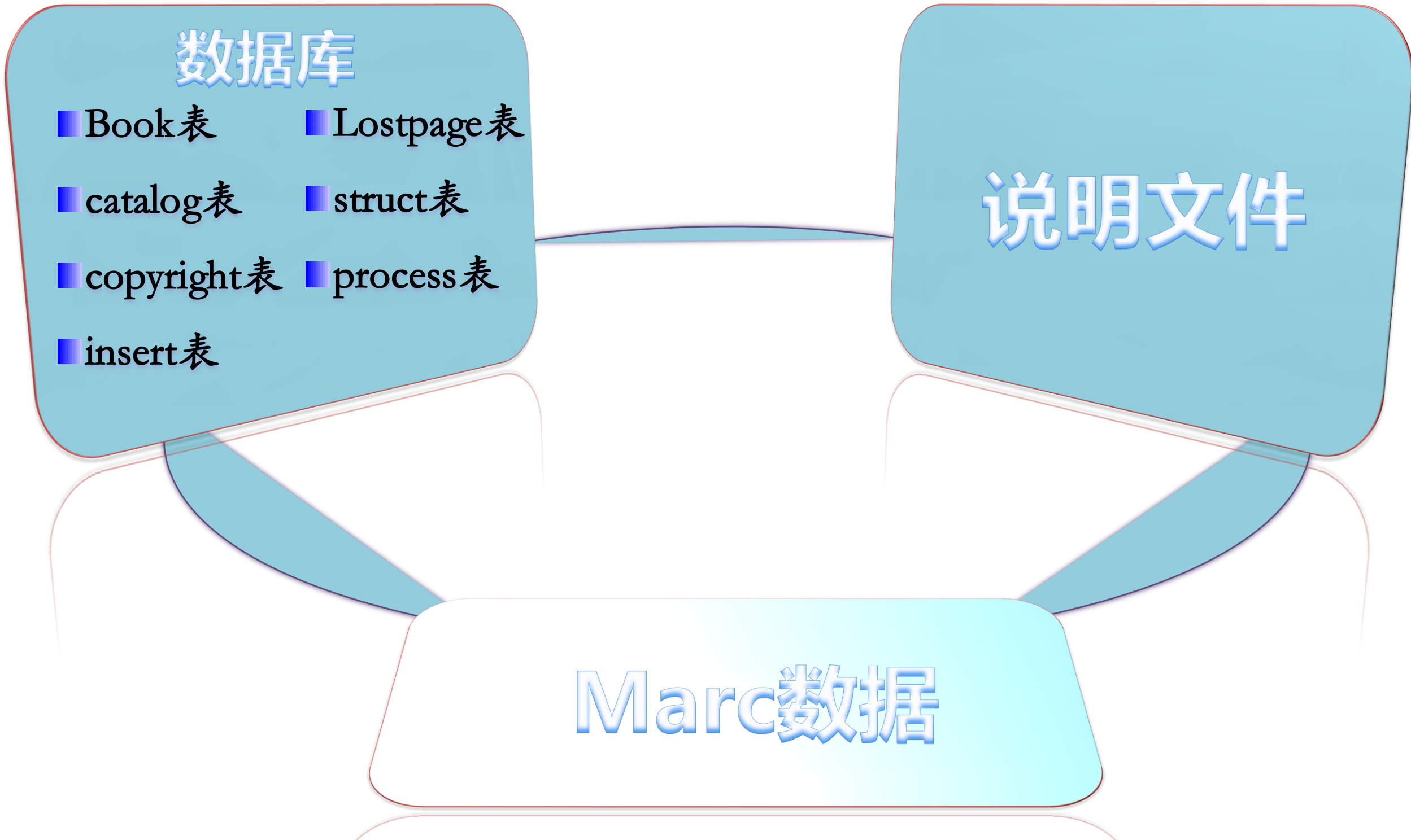




地方文献数字化项目工作流程



-  图书整理
-  扫描制作
-  **数据标引**
-  质量检查
-  数据验收
-  成品移交





地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

质量检查

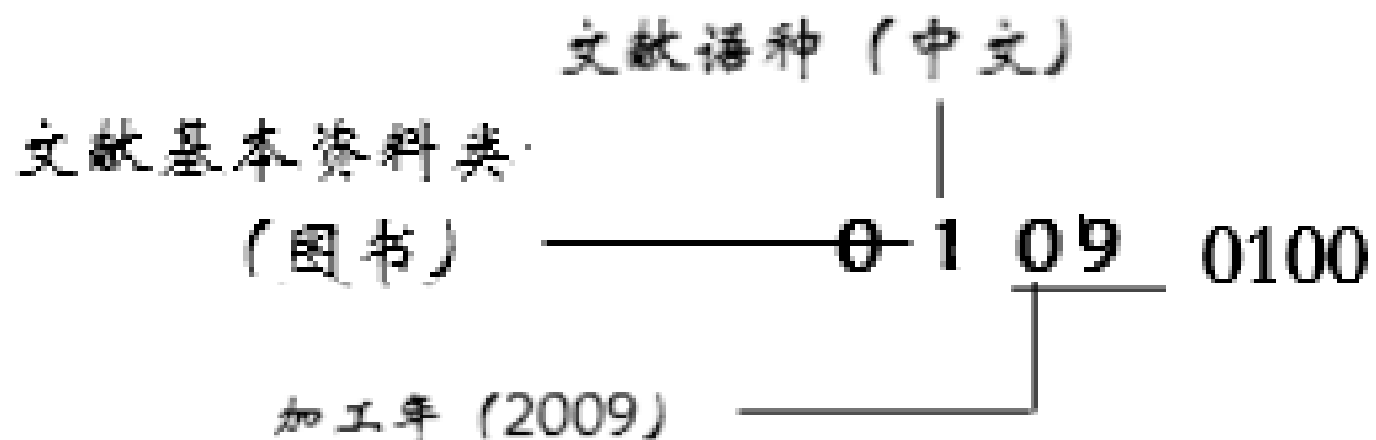
数据验收

成品移交



数据库命名

例：





地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

Book表

质量检查

数据验收

成品移交

book_id	cat_id	book_name	author	instructor	specialty	realm	degree	publish_h	publish_d	page_num
01613042_01	A811.5	马克思的哲学	丁立卿著	胡海波指导	马克思主义理	马克思主义基	博士论文	东北师范大学	2008	125页
01613042_02	A811.6	马克思的理论	韩玲著	袁银传指导	马克思主义理	马克思主义理	博士论文	武汉大学	2007	266页
01613042_03	A811.6 D082	马克思民主思	王国宏著	叶庆丰指导	科学社会主义		博士论文	中共中央党校	2007	100页
01613042_04	A811.6 D616	马克思社会形	黄斌著	荣长海指导	科学社会主义	中国特色社会	博士论文	天津师范大学	2008	175页
01613042_05	A811.6 D095	欧洲近代政治	杨晓东著	阎孟伟指导	马克思主义哲		博士论文	南开大学	2007	304页
01613042_06	A811.6 D990	论马克思恩格	董运珍著	李爱华指导	马克思主义基		博士论文	山东师范大学	2008	159页
01613042_07	A811.6	自由的意蕴及	曾宇辉著	徐大同指导	政治学理论	中西政治思想	博士论文	天津师范大学	2007	166页
01613042_08	A811.6	马克思阶级意	苗泉竹著	侯惠勤指导	马克思主义哲		博士论文	南京大学	2007	195页
01613042_09	A811.6	马克思恩格斯	李征著	刘建军指导	思想政治教育	思想政治教				
01613042_10	A811.6 B03	马克思的空间	孙江著	任平指导	马克思主义哲	发展哲学				
01613042_11	A811.6 F319	农村公社与东	熊振均著	盛邦和指导	马克思主义哲					
01613042_12	A811.6 B516	马克思和海德	吴书林著	王南?指导	马克思主义哲	马克思主				
01613042_13	A811.6 B03	从“世界历史	刘招明著	张盾指导	马克思主义哲					
01613042_14	A811.6 B03	论马克思恩格	李娟著	刘永富指导	马克思主义哲					
01613042_15	A811.6 B0-0	现代性的反思	李重著	张再林指导	马克思主义哲					

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	001字段



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

catalog表

质量检查

数据验收

成品移交

book_id	serial	chapter_num	chapter_name	author	page	ppage	page	page_1
01613042_01	1	目录						1
01613042_01	2		中文摘要		I		2B	0
01613042_01	3		英文摘要		III		4B	0
01613042_01	4		目录		V		1C	0
01613042_01	5		引言		1		1D	0
01613042_01	6	(一)	为什么研究《1844年手稿》?		1		1D	0
01613042_01	7	1.	我们是否“读懂”了《1844年手稿》?		1		1D	0
01613042_01	8	2.	从当代中国的理论需求来研究《					
01613042_01	9	(二)	转变研究《1844年手稿》的观					
01613042_01	10	1.	回到《1844年手稿》本身					
01613042_01	11	2.	研究《1844年手稿》的“经济					
01613042_01	12	一、	《1844年手稿》是马克思主					
01613042_01	13	(一)	马克思的第一部“经济学—					
01613042_01	14	1.	第一部“经济学—哲学”					
01613042_01	15	2.	马克思“新哲学”的研究方					
01613042_01	16	(二)	马克思对“理论之谜”与					

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	数值型
3	章节号	chapter_num	
4	章节名	chapter_name	
5	作者	author	
6	页码	page_num	
7	绝对页码	ppage_num	客观著录, 如实反映目录页原貌 (可为空)
8	页位置	page_place	数值型, 文件名数字部分 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”数值型



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引 copyright表

质量检查

数据验收

成品移交

序号	中文名称	字段名称	备注
1	加工编号	<u>book id</u>	
2	书名	<u>book name</u>	
3	作者	author	
4	001	record_id	
5	版权页位置	copyright_place	记录版权页文件名

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



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

insert表

质量检查

数据验收

成品移交

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

book_id	prior_text	page_plac	inset_num
01613046_52	208	T	1



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

lostpage表

质量检查

数据验收

成品移交

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	数值型

book_id	start_text	page_plac	lostpage_
01613043_16	205	T	1
01613043_16	139	T	1
01613043_16	111	T	1
01613043_16	71	T	1
01613043_16	13	T	1
01613043_16	13	L	1
01613043_16	13	L	1



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

struct表

质量检查

数据验收

成品移交

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	数值型
3	目录前, 前附页数	preface1_num	数值型
4	目录前, 前附页起始页号	preface1_start_page	
5	目录页数	content_num	数值型
6	目录起始页号	content_start_page	
7	目录后, 前附页数	preface2_num	数值型
8	目录后, 前附页起始页号	preface2_start_page	数值型
9	正文页数	text_num	
10	正文起始页号	text_start_page	数值型
11	后附页数	appendix_num	
12	后附页起始页号	appendix_start_page	数值型
13	封底页数	back_cover_num	

book_id	fore_cove	preface1_	prefa	content	conter	preface2_	preface2_	text_num	text_star	appendix_	appendix_	back_cove
01613042_01	2	5		2	V	9	1	116	10	0		1
01613042_02	2	9		3		21	1	245	22	0		1
01613042_03	2	5		2		6	1	94	7	0		1
01613042_04	2	4		3	I	10	1	165	11	0		1
01613042_05	1	6		3	V	30	1	274	31	0		1
01613042_06	2	1	1	2		18	I	149	11	0		1
01613042_07	2	4	I	2	i	9	1	157	10	0		1
01613042_08	1	0		3	1	4	1	191	5	1		1



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

struct表

质量检查

数据验收

成品移交

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	数值型
3	目录前, 前附页数	preface1_num	数值型
4	目录前, 前附页起始页号	preface1_start_page	
5	目录页数	content_num	数值型
6	目录起始页号	content_start_page	
7	目录后, 前附页数	preface2_num	数值型
8	目录后, 前附页起始页号	preface2_start_page	数值型
9	正文页数	text_num	
10	正文起始页号	text_start_page	数值型
11	后附页数	appendix_num	
12	后附页起始页号	appendix_start_page	数值型
13	封底页数	back_cover_num	

book_id	fore_cove	preface1_	prefa	content	conter	preface2_	preface2_	text_num	text_star	appendix_	appendix_	back_cove
01613042_01	2	5		2	V	9	1	116	10	0		1
01613042_02	2	9		3		21	1	245	22	0		1
01613042_03	2	5		2		6	1	94	7	0		1
01613042_04	2	4		3	I	10	1	165	11	0		1
01613042_05	1	6		3	V	30	1	274	31	0		1
01613042_06	2	1	1	2		18	I	149	11	0		1
01613042_07	2	4	I	2	i	9	1	157	10	0		1
01613042_08	1	0		3	1	4	1	191	5	1		1



地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

process表

质量检查

数据验收

成品移交

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	数值型
4	压缩因子	comp_factor	
5	灰度页数量	grey_num	数值型
6	彩色页数量	col_num	数值型
7	TIFF数量	tiff_num	数值型
8	PDF数量	pdf_num	数值型, 存储单位: MB
9	TIFF存储量	tiff_mb	数值型, 存储单位: MB
10	PDF存储量	pdf_mb	第一套典藏级光盘编号
11	典藏级光盘位置	cdA_place	第一套服务级光盘编号
12	服务级光盘位置	cdB_place	典藏级硬盘财产号
13	典藏级硬盘位置	hdA_place	服务级硬盘财产号
14	服务级硬盘位置	hdB_place	

book_id	book_name	dpi	comp	grey	col_num	tiff_num	pdf_num	tiff_mb	pdf_mb	abst_num	abst_mb	cdA_place	cdB_place	hdA_place	hdB_place
01613042_01	马克思的哲学	300		133	2	135		987.71		1	1	01613A10130		200801135	200801100
01613042_02	马克思的理论	300		279	2	281		1986.88		1	1	01613A10130		200801135	200801100
01613042_03	马克思民主思	300		108	2	110		794.63		1	1	01613A10130		200801135	200801100
01613042_04	马克思社会形	300		183	2	185		1339.01		1	1	01613A10130		200801135	200801100
01613042_05	欧洲近代政治	300		313	2	315		2255.96		1	1	01613A10130		200801135	200801100
01613042_06	论马克思恩格	300		171	2	173		1242.94		1	1	01613A10130		200801135	200801100



地方文献数字化项目工作流程



图书整理

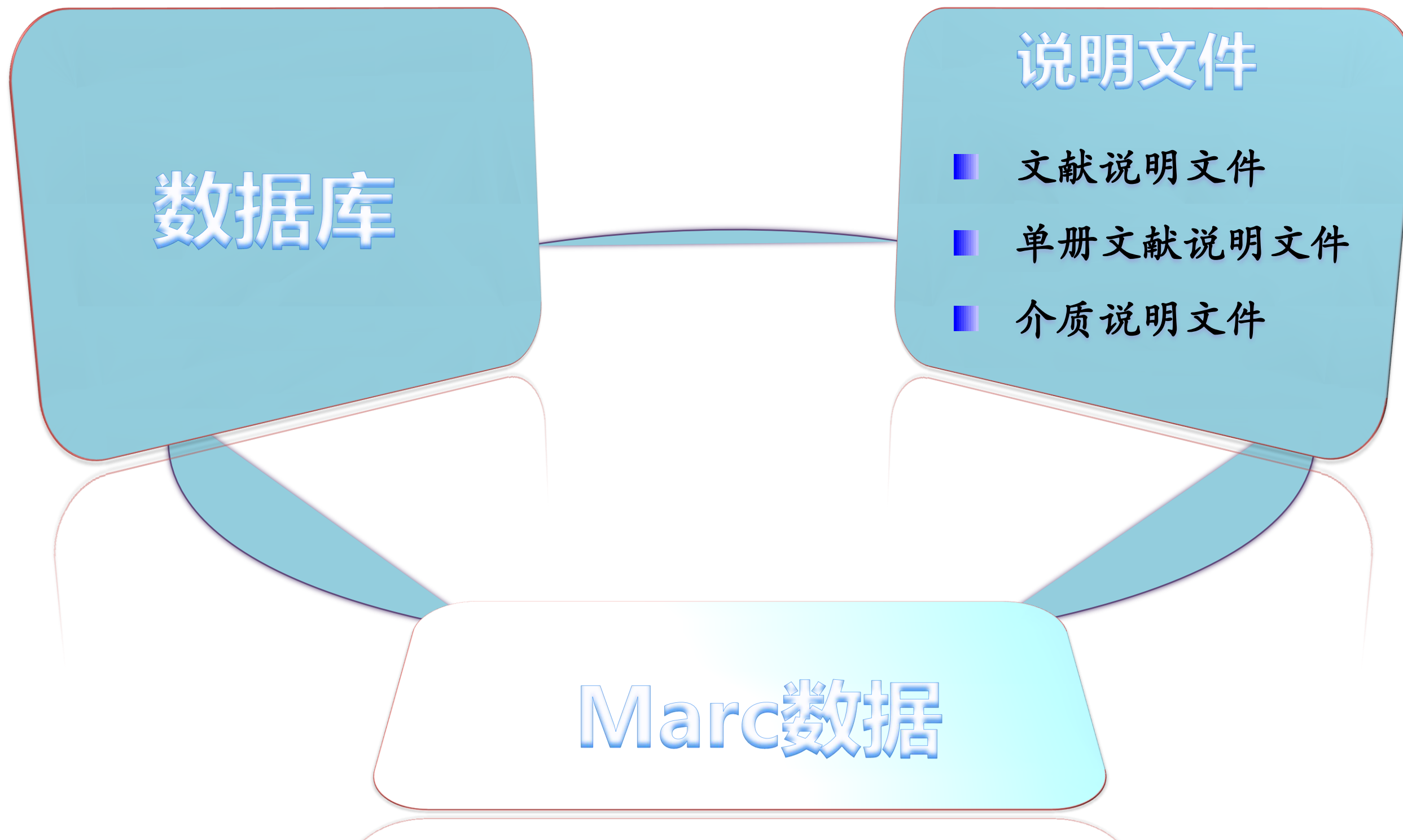
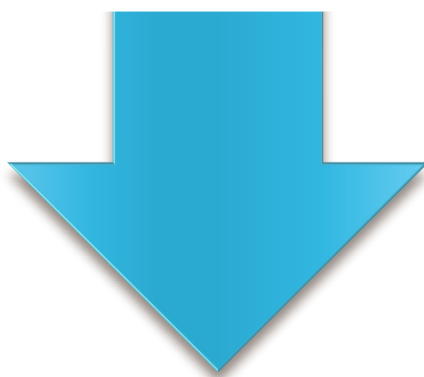
扫描制作

数据标引

质量检查

数据验收

成品移交





地方文献数字化项目工作流程



图书整理

扫描制作

数据标引

质量检查

数据验收

成品移交

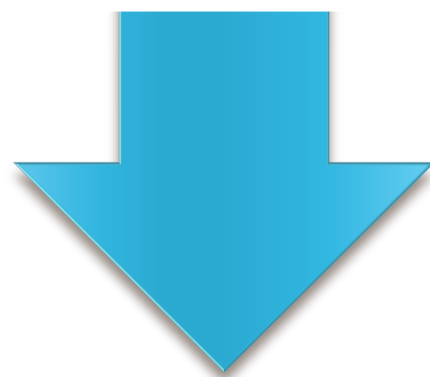
规范性

一致性

正确性

完整性

有效性





地方文献数字化项目工作流程



数据验收

图书整理

扫描制作

数据标引

质量检查

数据验收

成品移交

- Tiff图像
- txt文件 (单版和合并版)
- 双层PDF (单版和合并版)
- 对应数据库
- 介质说明文件
- 地方文献数据说明
- 单册文献说明文件
- 文献书目数据
- 第三方质检报告
- 地方文献验收数据提交单

质量检查

数据检查

验收合格



地方文献数字化项目工作流程



数据验收

图书整理

扫描制作

数据标引

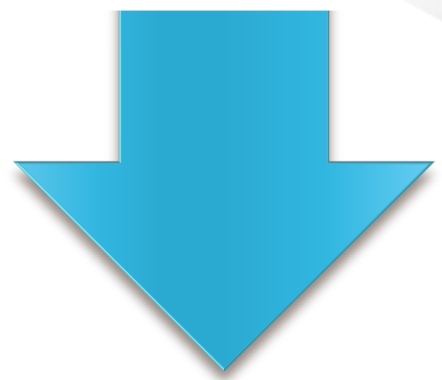
质量检查

数据验收

成品移交

- Tiff图像
- txt文件 (单版和合并版)
- 双层PDF (单版和合并版)
- 对应数据库
- 介质说明文件
- 地方文献数据说明
- 单册文献说明文件
- 文献书目数据
- 第三方质检报告
- 地方文献验收数据提交单

所有 Access 对象	book_id	book_name	author	chiabst_p	chiabst_n	engabst_p	engabst_n
表	01613042_01	马克思的哲学	丁立卿著	B00002_00	2	B00004_00	2
abstract	01613042_02	马克思的理论	韩玲著	B00003_00	3	B00006_00	4
book	01613042_03	马克思民主思	工国宏著	B00003_00	1	B00004_00	2
catalog	01613042_04	马克思社会形	黄斌著	B00002_00	1	B00003_00	2
inset	01613042_05	欧洲近代政治	杨晓东著	B00003_00	2	B00005_00	2
lostpage	01613042_06	论马克思恩格	蔺运珍著	D00001_00	3	D00004_00	5
process	01613042_07	自由的意蕴及	曾宇辉著	B00001_00	2	B00003_00	2
struct	01613042_08	马克思阶级意	苗泉竹著	D00001_00	2	D00003_00	2
查询	01613042_09	马克思恩格斯	李征著	B00002_00	4	B00006_00	6
	01613042_10	马克思的空间	孙江著	B00002_00	3	B00005_00	3
	01613042_11	农村公社与东	熊振均著	B00001_00	2	B00003_00	2
	01613042_12	马克思和海德	吴书林著	B00003_00	1	B00005_00	2
	01613042_13	从“世界历史	刘招明著	B00004_00	1	Y00003_00	2
	01613042_14	论马克思恩格	李娟著	B00001_00	2	B00003_00	2
	01613042_15	现代性的反思	李重著	B00001_00	2	B00003_00	2
	01613042_16	论马克思主义	张少军著	B00002_00	2	B00004_00	2
	01613042_17	列宁农民问题	许蓉著	D00001_00	1	D00002_00	1
	01613042_18	列宁司法思想	王建国著	D00001_00	1	D00002_00	2
	01613042_19	毛泽东中国社	牛田盛著	B00003_00	2	B00005_00	2





数字图书馆推广工程2014年度数字资源联合建设培训班

推广工程数字资源联合建设 地方文献数字化项目建设方案解读

国家图书馆 数字资源部

韩萌



谢谢各位！