地方文献数字化加工的新技术和新格式的运用

——浅谈资源联合建设中的方法突破

报告人：褚正东

推广工程数字资源联合建设的重要意义

• 文化插上科技的翅膀，让中小型图书馆分类表的“旧”书“新”起来。

《中小型图书馆分类表》自1957年开始试用，到1981年现行《中图法》正式试用，期间历经24年。镇江市图书馆藏有使用《中小型图书馆分类表》分编图书约8万多册，其中部分是反映镇江地区自新中国成立后的地方史料、文化志、方志和乡土作品等。这些图书深藏在提存书库里，得不到很好利用。地方文献数字化工程使得这类图书能够“变废为宝”，充分利用起来。

• 解决了“两端”资源不对称的问题，让信息资源渠道“扁平化”。

互联网科技和信息技术的发展，使得任意两个节点都能够存在理论上的连接，消除了分层冗余，信息流通和获取构建成最短距离的扁平化。然后，客观存在于两个节点间的资源不对称的情况将消失。理论上讲，流动于国家图书馆、省馆、市馆和县馆，甚至任一村镇基层点的海量资源实现了均等化服务。

• 资源和服务的共建共享，整体提升图书馆服务水平和事业的均衡发展。

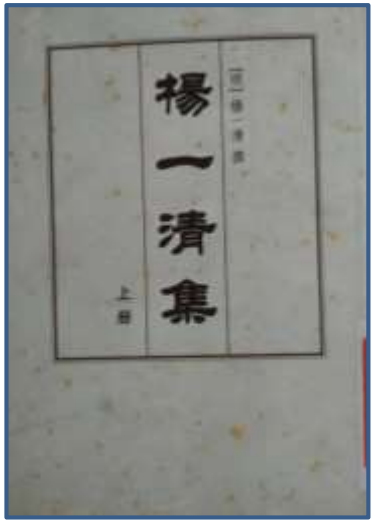
数字图书馆推广工程构建了一种新型的资源与服务的共建共享模式，客观解决了资源获取的难题。一方面基层图书馆具有鲜明地域特色、较高文献和历史价值的资源得到了资金支持能够加以保护和利用；另一方面国家图书馆获取了海量的信息资源，这不仅有利于对历史文化的传承利用，更是国家文化战略层面的迫切需要。

地方文献的遴选标准

- 《地方文献数字化加工规则（2016）》遴选馆藏资源的三要素：

一、**时间**：1949年以来出版
二、**资源载体**：图书、期刊、报纸
三、**类型**：方志、地方文史资料、民族语言资料

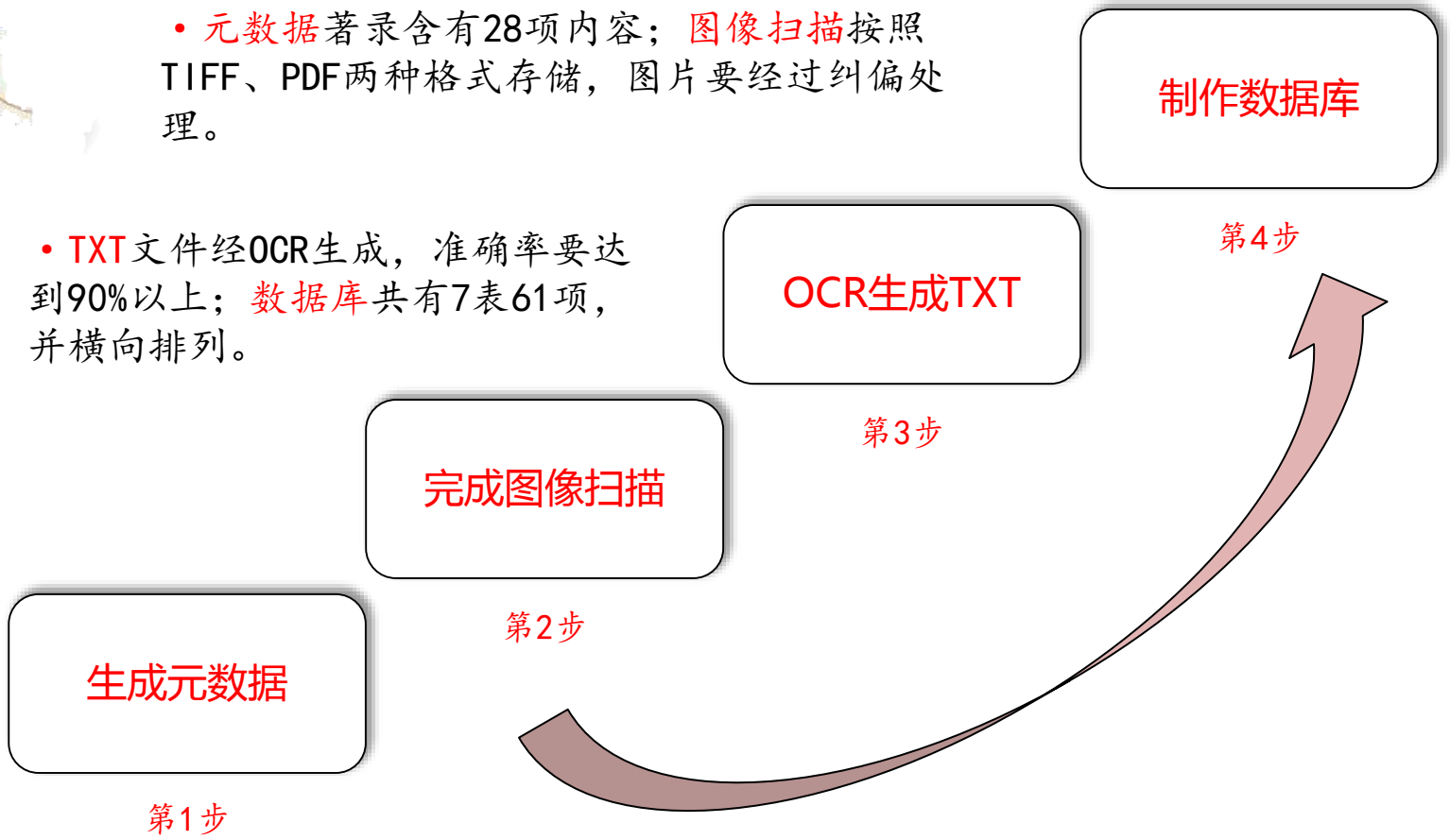
- 遴选的标准：1、特色馆藏；2、保护目的；3、研究需要；4、读者需求。



地方文献的加工规则

• 元数据著录含有28项内容；图像扫描按照TIFF、PDF两种格式存储，图片要经过纠偏处理。

• TXT文件经OCR生成，准确率要达到90%以上；数据库共有7表61项，并横向排列。



地方文献数字化加工的困难

• 技术力量薄弱，业务人员缺乏。地方文献数字化对图书馆采编业务和技术能力提出了较高要求，多数地级市图书馆在资源联建上缺乏必要技术力量和业务水平。

• “用”资源比“建”资源困难，版权问题成为瓶颈。地方文献包含的图书、报刊杂志以及各类其它资源因《著作权法》法律规定，具有相当长的保护期，因此，完成《版权证明》变成一件非常困难的事情，特别是版权人的不确定性使得开具证明几乎变成不可完成的任务。

• 数字化加工设施设备缺乏，加工能力有限，工业化水平很低。对多数馆来讲，数字化加工至少要有一台较好的高速平面扫描仪，最好能够自带纠偏功能，自动生成PDF文件，有OCR识别功能，才能基本满足加工需求。同时，因为设备和人手问题，加工的能力也不足，每人每天甚至做不到30页。考虑边际成本很不划算。

• 资源格式复杂，机器难以批量处理，人工作业成本较高。1949年建国初期多数图书、报刊仍采用繁体字、竖排版，这种格式类型现有OCR识别率极低，仅为20%~30%，由此需要大量人工录入、校验，编辑、贴合等工序，费时费力。

地方文献数字化加工的困难

竖排版（印刷）

多图版（印刷）

竖排版（印刷）

竖排版（手写）

镇江市图书馆数字化加工的成果

• 自2014年以来，镇江市图书馆把图书数字化，特别是地方文献古籍数字化保护和利用作为重要工程来抓，逐步探索出一条低成本、高效率、信息安全度高和专业技术强的可持续发展的全新商业运作模式。

完成馆藏地方文献古籍图像数字化36种275册32200页。

2014年

完成馆藏地方文献古籍图像数字化288册38260页，完成全文数字化38种246册31000页。机器合成双层PDF文件。
完成网事典藏200个网站，完成政府公开信息20000条。

2015年

已完成馆藏地方文献古籍全文数字化45种298册41000页双层PDF文件。
拟完成地方文献图书3万页，地方文献报纸1万版，网事典藏网站150个，政府公开信息1.2万条。
拟完成矢量PDF格式地方文献图书再造读本30部。

2016年

江苏文心数字产业有限公司

•2016年5月11日国务院办公厅转发文化部、国家发展改革委、财政部、国家文物局《关于推动文化文物单位文化创意产品开发的若干意见》，明确提出图书馆要依托馆藏文化资源，开发文化创意产品和服务。由此，我馆和江苏省“双创”团队江苏奥博洋信息技术有限公司合作成立了江苏文心数字产业有限公司。

•江苏文心数字产业有限公司成立于2016年3月，公司位于镇江市图书馆内。公司主营业务为内容数字化、数据处理、中文古籍数字化、数字内容发布管理、大数据分析等。公司负责人王旭是美国耶鲁大学博士，曾在美国担任Broadband Sports Inc的软件总设计师，idealab Inc的技术总监，美国商业网(business.com Inc.)副总裁，曾担任云南大学软件学院名誉院长兼软件工程系主任等职务，镇江市“331计划、江苏省“双创“人才引进计划和江苏省“双创团队”领军人才。公司研发团队现有3位博士，自主研发的图像切割录入校对系统已申请国家专利并得到业界广泛认可。

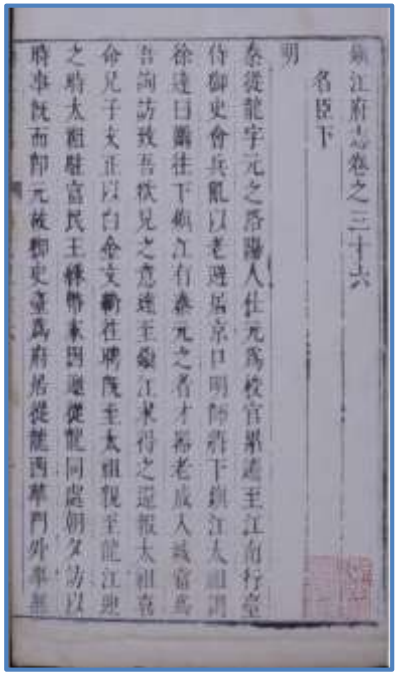
地方文献全文数字化必要性

• 1949年以来的图书作为脆弱的，珍贵的纸质文本，它易损坏、难传播、研读不便。无论是TIFF，还是PDF，图像格式仅能够反映图书原貌，而不能完善利用。

• 我们采用的全文数字化则可以使它转换成双层PDF等格式电子文本，具有可复制、易查寻、能检索的功能，既能如实再现纸质图书的原貌，还拥有原本所不具备的文档优点。

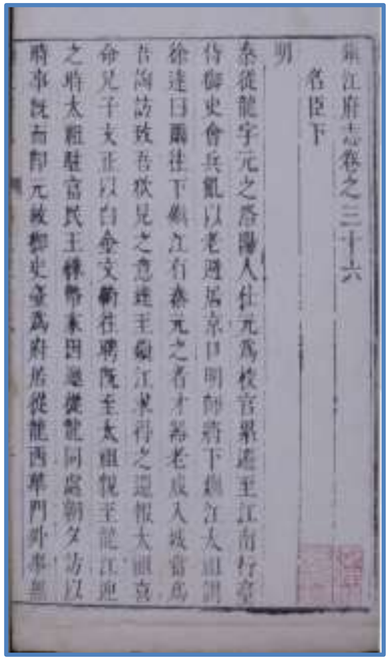
书籍不出馆，全文数字化

• 公司拥有先进的图像切割软件，提供客户端技术支持为图书馆在馆书籍图像直接切割，实现了重要图书资源“足不出馆”实现“碎片化”操作，图书完整的版权信息由图书馆自行掌握，信息安全得到完全保障。



书籍不出馆，全文数字化

- 使用图片碎片化技术，原图中的字符信息在客户端系统后台经分割处理，以碎片的形式上传服务器，出现在录入作业人员屏幕上的都是无信息关联的碎片画面，可完全保障信息安全，杜绝了泄密的可能。


鎮江府志卷之三十六
名臣下

书籍不出馆，全文数字化

• 目前，图书电子化的主要手段是采用OCR技术进行识别，但是早期地方文献图书、报刊等字体繁复，格式多样，插画较多，校色困难，因此OCR识别率一直不高，因此人工录入和校对仍然是保障全文数字化用以识别技术的重要手段。

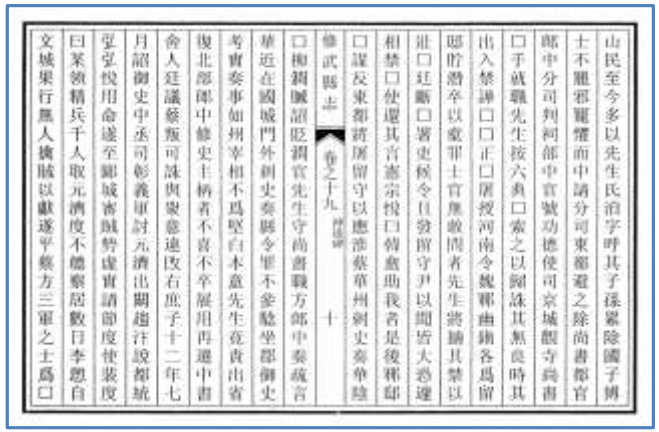
• 公司采用互联网众包模式，实行互联网在线作业，充分发掘社会各界人员的碎片化时间，依托敲宝网30万“敲友”，形成了巨大的交付能力。

书籍不出馆，全文数字化

书籍不出馆，全文数字化

纳品介绍：

- 图片档：精细加工处理后，生成24位彩色TIFF图像，最大可能的展现古籍的原始风貌；
- XML：具有清晰易读、传播方便、平台通用、检索迅速等优点，为数字图书馆提供基础；
- 双层PDF：上层是原始图像，下层是识别文本。阅读时看到的是图像，同时又可以像电子文件一样查询、复制。在图书原版图像上实现全文检索、全文定位、复制粘贴；
- 矢量PDF：根据图书原档的内容及格式，进行重新排版，为图书再造读本使用。



基于大数据众包的全新商业模式

• 将数字内容制作和加工的作业模式扩展到互联网的虚拟空间。充分发掘社会各界人士的碎片化时间，实行互联网在线作业，从而大大降低内容制作的成本；

• 运用云计算等最新的IT技术，将图片文本拆分为大量的单字，实现任务碎片化、业务流程化，并保障图书所有人的信息安全；

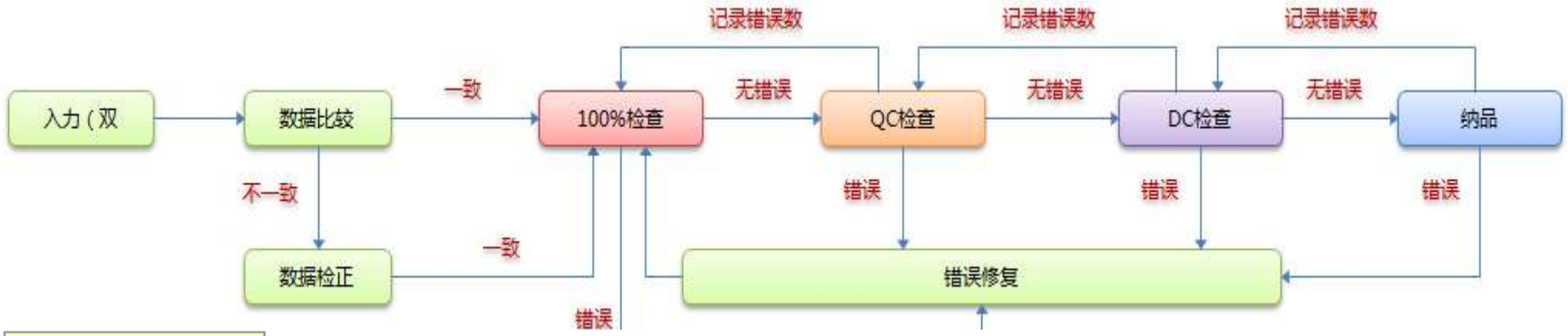
• 最后，平台会将作业者的内容制作和活动情况数据完整纪录下来，构建大数据平台，提供进一步挖掘和分析的数据，为进一步改善作业效率和纳品质量提供依据。

互联网众包模式形成巨大的交付能力

基于“双校”的纳品检验流程化管理

- 运用大数据、云计算等IT技术确保纳品交付品质

以作业流程改进为基础，每一个任务结点均建立数据收集和分析管理机制，确保交付品质。文字精度达到万级，除同行双录、连笔手写、污损模糊等特例，正确率基本在万分之三到万分之五之间。



作业指南：

1.0 数据比较发现错误时，需分析入力A和/或入力B注入的错误，然后按分析结果计入“入力人员数据记录”表单中的“入力错误数 - 比较”数据列中。

2.0 质量控制（QC）活动包括：数据比对，FC，QC检查，DC检查，纳品。

3.0 任何一个QC活动找出的错误数都需计入“入力人员数据记录”表单。

4.0 任何一个QC活动（除“数据比较”外）找出错误时，除了执行作业1要求之外，还需在上一步QC活动对应表单”遗漏错误数”列记录一笔。

5.0 图例说明 入力人员作业 FC人员作业 QC人员作业 DC人员作业 用户作业

基于“碎片化”识别的劳动力成本大幅降低

• 互联网（含手机移动端）在线作业的众包商业模式图像分割技术的运用，使碎片工作时间赚取收入成为现实。

• “敲宝网”将数字内容制作和加工作业模式扩展到互联网的虚拟空间。互联网在线作业可以大大降低内容制作的成本，实现新型的用工模式，大量闲散的社会人力资源得到了有效应用。劳动力成本大幅降低使得产品和服务成本的大幅降低。

劳动力成本降低

我们与传统图书加工的区别

	镇江图书馆	其他加工商
加工模式	免费提供切割软件，到馆碎片化加工，信息安全高	将扫描影像拷贝出馆，整页分发加工，信息安全系数较低
精度	针对不同版本的古籍，采用不同的加工流程，精度万分之三以上	千分之三至万分之五
效率	10本/天	1本/7天
成果物	Word、Txt、双层PDF	TXT

谢 谢！

ADD：江苏省镇江市解放路17号镇江市图书馆

Mob：138 5290 7182（褚正东）

132 7619 9999（潘　峻）