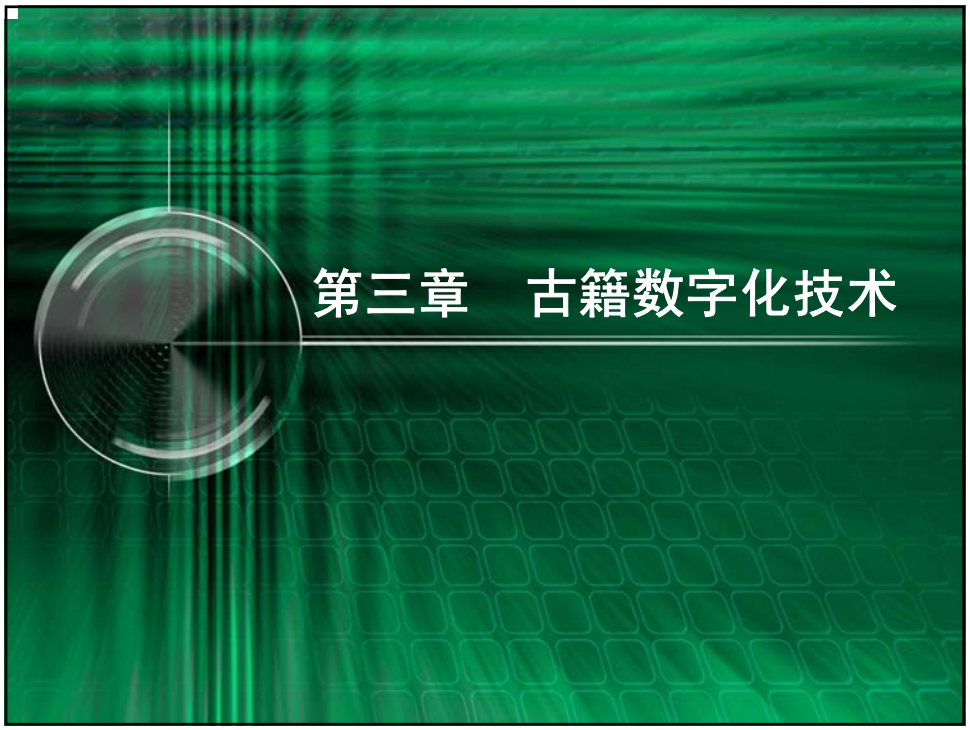
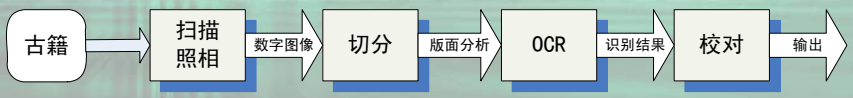

古典文献专业
欽定四庫全書
欽定四庫全書
欽定四庫全書
欽定四庫全書
古籍數字化
主講教師 葛懷東

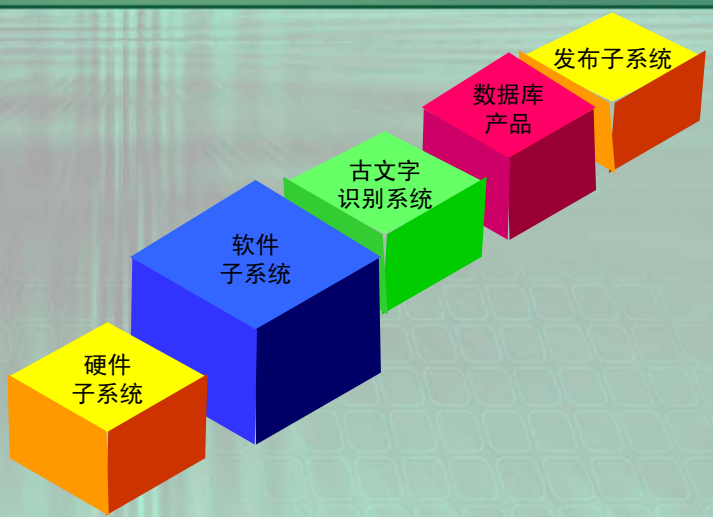

第三章 古籍数字化技术

古籍数字化基本处理流程



- 扫描或数码照相的主要工作是将纸质信息尽量无损地转化为数字图像，这已是成熟技术，现有的扫描仪就可以完成此项工作。
- 切分是将数字图像进行版面分析，分割成单字图，以利于OCR处理，同时记录了每个汉字的位置信息，使得无纸校对成为可能，并且为版面复原提供了基础。
- OCR是利用模式识别算法将单字的数字图像映射为字符。由于OCR无论如何完美，一定会存在差错，这些差错必须由人来修正，而利用版面分析记录的信息，提供软件进行无纸校对是整个流程的核心。

古籍数字化加工系统的搭建



古籍数字化系统平台的建设

- 系统主机
  - UNIX or Windows Server 的选择
  - 主机能量 （CPU 等级，主存储器）
  - 主机扩充与升级（能量与费用）
  - 相对稳定性
  - 相对安全性
  - 运作环境需求（机房）
- 储存设备
  - 空间需求的估算
  - 扩充方式与能量
  - 磁盘稳定性
  - 环境需求
  - 数据安全备份与还原

古籍数字化系统平台的建设

- 其它相关设备
  - 进行数字化工作所需要的设备
    - 数字设备
      - 扫描仪、数码照相机、数码摄像机、数字录音设备……
    - 播放实体对象所需的设备
      - 录放机（VHS）、录音机、唱盘……

古籍数字化系统平台的建设

- 功能规划 – 功能完整性
  - 资源建置与维护的完整性
  - 使用权限管理的完整性
  - 可处理数字内容种类的完整性
  - 可处理数字内容档案格式的完整性
  - 使用端查询阅览的完整性
  - 个人化信息的完整性

古籍数字化系统平台的建设

- 藏品之信息组织与检索的技术
  - 后设资料（metadata）
  - 数据库管理系统
  - 程序语言
- 虚拟实境的技术：2D、3D...
- 动画：Flash、Director....
- 图片：PHTOSHOP、PHTOIMPACT
- 影音：Streamline-Play或 Download-Then-Play...
- 标示语言：XML、HTML...
- 浏览器：IE、火狐（Firefox）
- 网页设计：Frontpage、Dreamweaver

古籍数字化系统平台的建设

- 标准规范 – 标准兼容性
  - 字码—UNICODE UTF8
  - ISO2709 (MARC) 数据交换
  - Metadata XML 数据交换
  - OAI & Z39.50 检索与传输协议
  - 与其它系统的整合与互动
- 安全管理
  - 电子商务
  - 账号/密码
  - 水印、加密技术

数字化业务的预设

- 确定数字化内容/对象与规范标准
- 藏品准备
  - 藏品了解与清点整理
    - 原件的保存状况与处理方式
  - 预估数据的数量
  - 决定数字藏品传送或取用方式
  - 制作拟数字化藏品的完整目录及相关表格

数字化业务的预设





数字化业务的预设

- 规范标准的确立
  - 决定数字对象之档案储存、组织及呈现的规范与架构
    - 决定储存数字档案的命名原则
    - 建立数据库的管理与连结架构
    - 决定数字藏品的组织方式和呈现架构
- 决定数字对象的规格
  - 数字化方式
    - 考虑因素：使用者需求及经费状况
    - Key-in、OCR、扫瞄、语音识别或影像处理
  - 储存媒体
    - 数据库主机硬盘、磁盘、磁带、光盘、激光视盘
  - 储存格式
    - 文字、影像、声音的储存规格

数字化业务的预设

- 人力规划
  - 人力类型
    - 内容专家、系统设计人员（信息人员）、网页/美工设计人员、信息组织人员、数据建文件人员……
  - 自制vs. 外包
  - 标准作业流程的制订
  - 教育训练

成本/预算分析

- 建置与维护费用
- 计划与筹备
- 建置费用
  - 系统建置成本
    - 数字典藏系统软硬件费用
  - 资源建置成本
    - 实体资源整理费用
    - 数字资源建立费用
      - 书目/摘要/全文/影音
- 后续费用
  - 系统保固期
  - 保存维护成本
  - 教育训练与推广成本
  - 扩充费用
    - 系统功能扩充成本
    - 硬设备扩充成本

古籍数字化技术支持

- 1. 汉字编码和字符集
- 汉字在计算机上的显示是通过编码来实现的，汉字编码的空间决定了计算机可显示汉字的多少。
- 我国1980年颁布的GB2321字符集只能表示6763个常用汉字，2000年颁布的GB18030共收录27484个汉字，而古籍中通用汉字就有4万，再加上异体字、避讳字、生僻字等，古籍用字可达8万之多。因此，有超大的字符集支持是字符数字化的基本要求。为了解决世界范围内的信息交换、处理和显示问题，国际标准化组织制定了ISO/IEC 10646国际标准编码，也称为Unicode统一编码。
- 由于Unicode的编码空间浩瀚，最新的Unicode 5.0可以定义71226个汉字，这就使古籍中大量生僻字、异体字数字化成为可能。Unicode标准解决了古籍数字化的字符编码不足和不同系统编码表不同而造成的信息交换问题。同时，Unicode标准也使各种不同的语言平台之间有了一个共同的编码系统，使得古籍的跨平台展现变得可能。

古籍数字化技术支持

- 2. 数字化输入技术
- 古籍的数字化输入主要是将古籍文献转化为数字符号保存在计算机中，其中涉及到图形图像处理技术、汉字字符识别技术、数字存储技术等。古籍资料的输入有两种形式，一种是图像格式，一种是电子文本格式。
- 古籍资料的图像化具有成熟的技术，如快速扫描、数字照相。两种技术各有特点，扫描的方式录入速度快，图像不会变形失真，但会对古籍造成一定的损伤；数字照相的方式可以减少古籍的损伤，但在高分辨率下图像容易变形失真。生成的古籍图像要进行相应处理，包括对图像进行裁边、纠偏、调色、去污、分辨率转换等。为保持古籍原貌，初始扫描或照相时，可能采用的是彩色模式，而后继的版面分析和识别等处理，都是在二值图像的基础上进行的。因此，图像的二值化是必不可少的，这就要用到图像处理技术，常用的软件如Photoshop、ACDSee等。

古籍数字化技术支持

- 3. OCR光学识别技术
- OCR （Optical Character Recognition光学字符识别）是一种文字自动输入方式，通过光学技术对文字进行识别，通过光电转换，获取纸张上的图像信息，利用各种模式识别算法分析文字形态特征，判断出文字的标准编码，并按通用格式存储到文本文件中。
- OCR系统是数字化加工处理的一个关键环节，它主要包括版面分析、汉字识别两个方面的内容。经过专业的OCR识别辅以人工校对可以生成准确率较高的古籍文本，较人工录入的方式具有很大的优越性，为数字化古籍的使用提了方便。

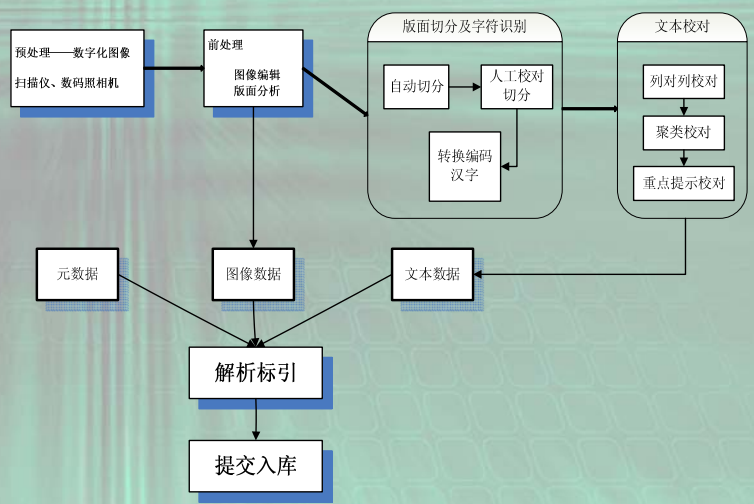
古籍数字化技术支持

- 4. 网络与数据库技术
- 网络和数据库技术在古籍数字化过程中主要应用于构建古籍存储与利用平台。数据库和网络是数字古籍进行数据存储、传输、利用的载体。根据古籍的文献组织结构和用户需求设计合理的数据库结构，为数据量巨大的数字古籍提供存储、检索、管理、利用的平台需要数据库技术的支撑。同时，为了更好的利用古籍数字化的成果，古籍数据库必须具有网络支持。实现超链接的阅读环境、古籍数据的远程传输利用、数据更新和版本升级、交流论坛和一些辅助功能的实现都需要有网络技术的支持。建立古籍全文数据库和利用的网络化已经是古籍数字化发展的主流。

古籍数字化技术支持

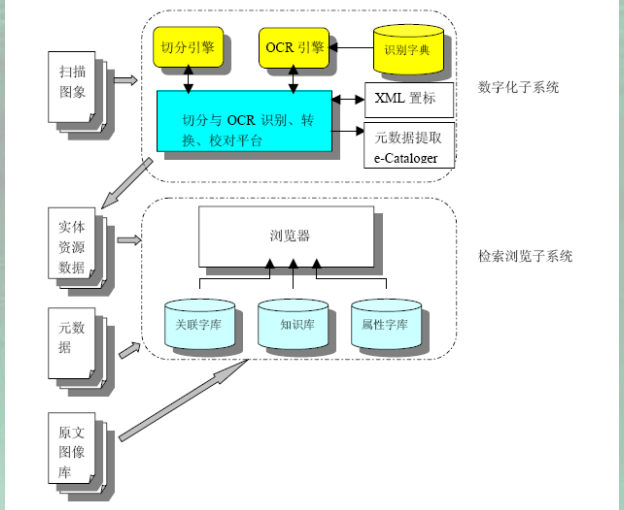
- 5．智能化处理技术
- 纸质古籍可为学术研究提供简单的服务，数字化后的古籍是纸本古籍高级形式，它突破了文献的物理形态，深入到了它所包含的信息单元，利用计算机算法对这些信息单元进行分合与重组，可以向读者和研究人员提供针对性更强、内容更丰富的服务。对古籍信息单元的研究是古籍资源开发智能化的途径。

古籍数字化处理流程与体系结构



古籍数字化处理流程与体系结构

体系结构：
一个基本的古籍数字化系统至少包括两个子系统：数字化子系统和检索浏览子系统。



本 章 结 束