

信息化时代的数字图书馆的 技术与应用

田捷 研究员、博士生导师
中国科学院自动化研究所

<http://www.3dmed.net>

Email: tian@dr.com

学时： 40
方式： 采用计算机多媒体方式授课
目的： 使学生了解掌握当今数字图书馆技术的动态、进展、多学科融合的高新技术

内容简介：以数字图书馆/数字档案为主线，介绍互联网环境下信息文献技术的最新发展，以及多文种处理、文献结构和网际互联等领域的标准与应用，同时包括涉及到的图像压缩、全文检索和信息安全等技术的应用，以及数字图书馆的建设实例讲解。

大纲：

一、数字图书馆的背景与需求

- 信息技术的冲击
- 互联网高速发展的带动
- 内容为王对图书馆专业的影响
- 中国的数字图书馆建设的发展

二、数字图书馆的概念和结构

- 数字图书馆的概念
- 数字图书馆的构成
- 数字图书馆的功能
- IBM数字图书馆系统的案例

三、数字图书馆技术专题：文献信息与多媒体应用

- 文本、图像、声音、视频等信息特点及处理，传统信息的移植
- 多文种的处理：ISO 646
GB2312-80 ISO/IIIC 10646
- 纸媒介、胶片、光盘、网络等媒介的应用

四：数字图书馆技术专题：文献信息的组织

- 介绍海量信息的存储结构，信息的基本检索方式，关于元数据的标准和发展，传统图书索引方式的演变：MARC数据、Z39.50标准等，XML结构
- 分布式结构及CORBA技术应用

五：数字图书馆技术专题：图像压缩及处理一

- 通用图像压缩：CCITT G4和JPEG
- 黑白二值图的压缩：JBIG和JBIG2
- 进一步的压缩技术：小波变换
- 新的图像压缩标准：JPEG2000
- 图像的优化技术：去污、去噪、纠偏等等

六：数字图书馆技术专题：视音频压缩及处理二

- 视频动态压缩：MPEG、MPEG2、MPEG4、MPEG7、H.261
- 音频动态压缩：MP3

七：数字图书馆技术专题：信息检索与全文搜索技术

- 信息检索技术
- 全文搜索引擎

八：数字图书馆技术专题：版权保护与信息安全技术

- 网络安全
- 访问权限控制及信息加密技术的应用
- PDF MERCHANT技术原理
- 水印技术

九：数字图书馆技术专题：PDF与电子商务

- 电子书的发展：ADOBE和MS READER，PDF技术特色
- ADOBE的epaper解决方案
- PDA和手持ebook
- 基于电子书的商务模式，B2C和C2B模式
- 在线支付手段

十：数字图书馆技术专题：基于Web的信息发布与传播

- WEB基本技术
- 面向对象的技术 C++ & Java
- 数据库技术 JDBC等

十一、数字图书馆的技术案例介绍

- 以DIGIARK和IBM方案为例

十二、数字图书馆的应用案例及讨论

- 以北图（国家图书馆）、首都图书馆为例。

⌘ IBM方案介绍

■ 概述

IBMDB2数字图书馆系统集成了信息捕获，存储管理，查询检索以及安全发布等四大块技术，IBM具有目前为最大的媒介机构实施数字图书馆的经验，其实施的项目包括国会图书馆、CBS新闻中心，台湾故宫博物院，梵帝岗图书馆，俄罗斯Hermitage博物馆。

■ 创建及捕获

对开放性的承诺是IBM DB2数字图书馆系统的基石。

IBM DB2 DL不仅能够创建原始信息，并且能够捕获开放环境中的已有信息，支持包括声音，视频，图形，图象，文本等的多媒体的对象。

IBM认为：创建DL的动机是要保存模拟材料，如书、手稿、图片、胶片、视频及录音等易损坏的，或者是价值不菲的物品。为此目的IBM在其DL中集成了多项先进的图象获取与增强技术，如IBM与梵帝岗图书馆合作开发，集

成了高分辨率扫描及颜色矫正技术，使得能用近乎完美的质量和颜色捕获或重显影像。通过软件处理，可以在因特网上清晰地阅读。

数字化后可以修复原始材料因年代久远及保护不当的损坏。IBM DB2 DL使你可以同时看到先前破坏和修复后的情景。

■ 存储及管理

对于数量巨大的多媒体信息，存储和管理的关键是有效地组织并便于搜索。IBM DB2 DL的信息管理特征包括自动索引，建文件夹、相关性、特征提取和翻译。

IBM 另外两个特征是，开发性和可扩展性，其支撑平台包括NT，AIX和OS390。

IBM DB2 DL开发了独特结构，用于存储对象的管理。IBM DB2 DL存储和管理内部结构的核心是一台数据库服务器，它管理着目录信息，并提供到对象的链接。对象服务器存储着数字图书馆实际的数据文件，例如，一段视频。数字图书馆存放的数据不允许用户随机访问，而必须由数据库服务器向终端用户提供访问路径。数据库服务器采用这种三角结构，防止未被授权的用户访问数据库服务器。

IBM DB2 DL提供分层的数据管理，人们可以将数字对象存储在希望的地方，最有可能被访问的High profile对象可以存在高速硬盘上，而不常访问的对象可以存储在磁带、或光介质上，这样可以降低存储费用。

■ 搜索和访问

现在大多数图书馆都在使用美国国会图书馆的分类系统，但是数字图书馆的混合媒介对象需要另外一种方法去检索。目录分类系统是向用户指名的信息的位置，但是IBM DB2 DL使用强大的搜索和访问技术，帮助用户直接检索内容。

IBM DB2 DL方案提供不依赖于内容的数据库存储，允许多种对象可扩展地存储在一起。多种对象的分类方式不同于单个对象，例如书可以按题目、作者和主题分类，一个电影可按制片、导演或主体分类。如果用户想要查询某一部数据定义，如作者、主题、标题、长度等。

传统技术只允许用户按关键字检索，并返回一串含有该文字的索引。IBM 检索技术包括文本和图像的分析，自然语言查询，允许用户用简单、自然的风格表达查询，而忽视确切的字母位置。这种查询返回分级列表，高相关度的列在首位。文字分析，如“白宫”和“白房子”的区别，同时能识别“IBM”和“国际商用机器”之间的关系。

QBIC，图像内容检索，是IBM的获奖技术。通过颜色百分比、分布、位置和图像描述，可以检索到需要的图像。IBM DB2 DL技术扩展到移动图像检索。

■ 发布

IBM DB2 DL可以通过内部网、企业网、WWW或交互电视直接向用户发布。IBM开发了ATM技术和网管软件来辅助数字和模拟信息的传递。

■ 版权管理

IBM DB2 DL通过电子签名、数字水印保护作品版权，用电子信封保护作品的发布。任何人必须用密钥才能想打开电子信封。用户可以通过预缆电子信封的内容，来决定是否购买密钥。对于安全级别高的内容，电子信封使用若干个密钥进行保护。

⌘ DIGIARK方案介绍

■ 数字图书馆的系统组成

■ 数字图书馆系统应完成信息资源的生产加工、存储、检索、发布、保护、以及共享等功能环节。其系统结构如下图所示：

其中：



用户

数字图书馆系统

图书
杂志
图片
录音
录像
缩微

加工系统

存储系统

元数据

数字对象

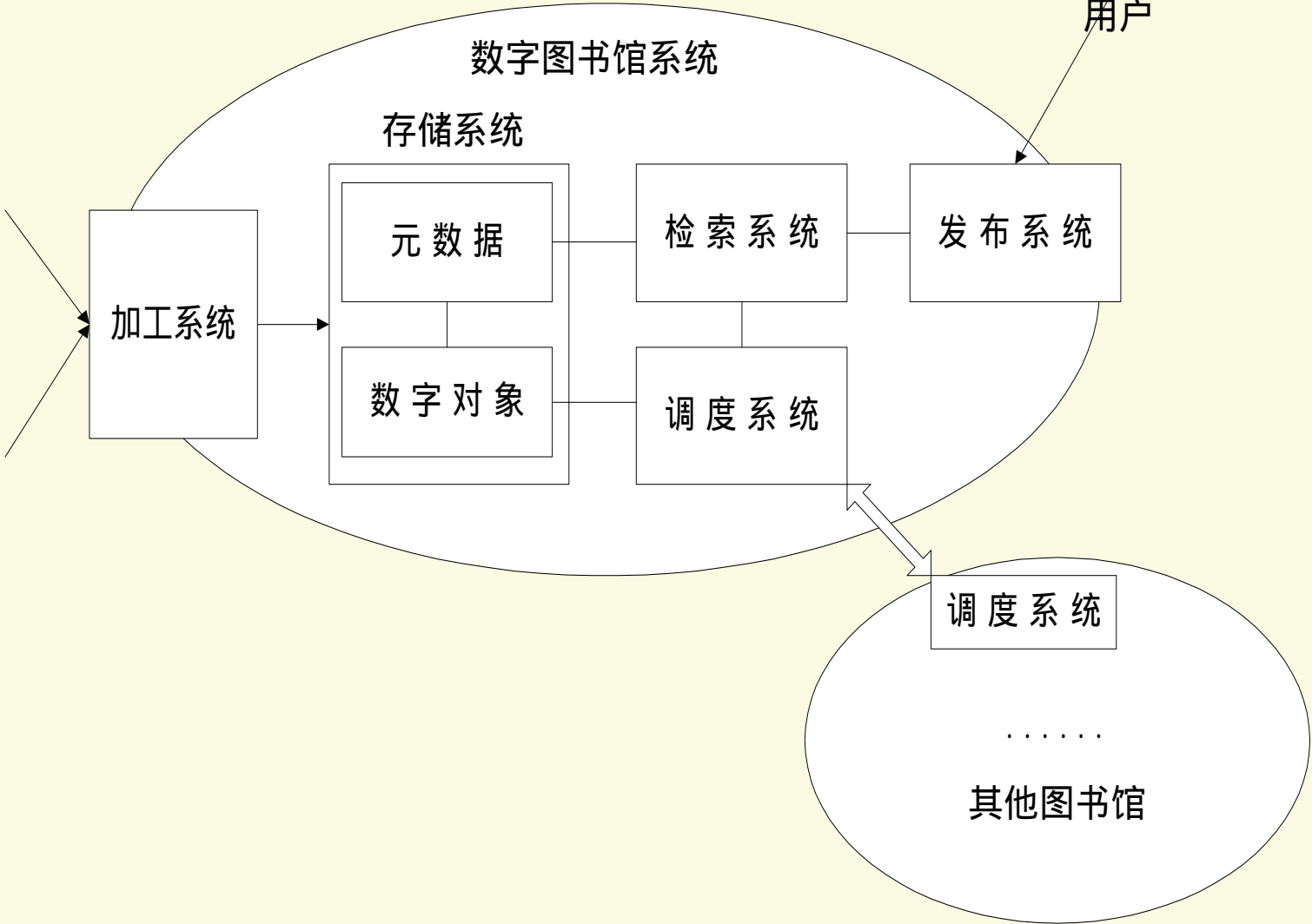
检索系统

调度系统

发布系统

调度系统

.....
其他图书馆



- 📖 加工系统：各种非数字化对象的数字化处理，同时完成数据对象的索引（index）和置标（markup）处理
- 📖 存储和管理系统：解决海量数据的存取、备份、权限控制等管理
- 📖 查询检索系统：通过基于SGML的搜索引擎，实现元数据检索及内容的全文检索
- 📖 发布系统：解决数据对象的流通、传播和增值，以及安全和版权的保护
- 📖 调度系统：解决异地跨库的数据共享