# 基于古文献特藏的数字图书馆系统的设计与实现

肖琥　冯英

（北京大学图书馆）

## 介绍

正如在传统图书馆中存在有大量特藏一样，数字图书馆同样需要收藏特色资源，这部分资源我们称为数字特藏（digital special collection），是某一数字图书馆单独收藏的资源，具备特殊收藏和利用的价值。

正在建设中的北京大学数字图书馆，除收藏有数据库、电子期刊、电子图书、网络资源等数字化资源外，还拥有大量特色资源，包括：

- 北京大学博硕士学位论文；
- 北京大学著名学者教授的手稿、照片等各类收藏；
- 北京大学课程教学参考资料；
- 北京大学古籍数字特藏。

在上述特色资源中，数量最大、最具特色的当属古籍数字特藏，它是在北京大学图书馆纸本古籍特藏的基础上建设的，与专业出版商出版的电子版《四库全书》、《四部丛刊》、《二十五史》等资源一起，共同构建成北京大学古籍数字图书馆。它的建立，将使北大图书馆馆藏古籍突破时空的限制，在全世界的范围内得到广泛的利用，并能够长久妥善地留存于世。

古籍数字图书馆的建设代表了北京大学数字图书馆的一个方面，完整地体现了北京大学数字图书馆的建设与服务思想。本文将从资源建设、标准规范、系统结构与新技术的应用、服务建设等方面对北京大学古籍数字图书馆进行全面介绍。

## 一、资源建设

北京大学图书馆目前收藏中国古籍约 1,600,000 册（件）、12 万种，其中孤本、珍稀本比比皆是，并有相当数量是在公元 16 世纪以前印行的；被辟为特藏的敦煌卷子、家谱、舆图、戏曲小说、地方志、少数民族文字古籍、金石拓片等类型藏书，都在海内外收藏界占有重要的地位。特别是金石拓片，收藏异常丰富，计 30,000 种、约 60,000 份，拓印对象包括商周甲骨、青铜器，秦汉至明清的碑刻以及砖文瓦当等中国历代金石文献，许多拓片是举世罕见或北大独有的。

建设中的北京大学古籍数字图书馆将选择其中一部分作为自己的收藏：

1，古籍特藏，包括：

(1) 敦煌卷子 240 余件；

(2) 宋元版书 350 多种，5,000 多册；

(3) 明代嘉靖（1566 年）以前的版本约 3,000 种，25,000 册；

(4) 古代舆图 500 余种，名人书画近百种；

(5) 写本系列：包括手稿本、名人信札、日记，影抄本、旧抄本、名人抄本，圣训、玉牒、奏折、文书、档案、地契等，在 6,000 种以上；

(6) 手绘本 100 多种，近千册；

(7) 家谱 1，000 余种，近万册；

(8) 古代戏曲小说约 4,000 种，35,000 册；

(9) 地方志共 4,000 多种，60,000 册。

总计约 20,000 余种，均为传本稀少、版本珍贵、学术价值较高的收藏。

2，金石拓片

包括清代缪氏艺风堂、张氏柳风堂等两藏拓大家的完整收藏，以及其他著名学者、收藏家的旧藏，数量多，版本好，价值高，这些拓片将逐步经过数字化加工收入到古籍数字图书馆中，约 30,000 种、60,000 份。

基于上述收藏，北京大学古籍数字图书馆将包括以下数据库：

- 对象数据库：

  包括古籍拓片图像数据库、古籍拓片全文数据库，主要通过数字扫描加工、OCR 识别转换和人工录入方式进行建设。

  初期建设将对古籍和拓片进行扫描加工，建成图像数据库；之后逐步通过 OCR 技术转换（古籍）和人工录入（拓片）等方式，进行全文数据库的建设，最终实现基于内容的全文检索。

- 元数据库：

即按照专为古籍和拓片设计的元数据格式，对古籍和拓片进行描述和揭示，便于读者浏览和检索，同时在该数据库中通过标识建立与对象数据库的链接。

## 二、标准规范的设计与应用

数字图书馆大规模开展建设之前，首要解决的两个问题就是版权和标准规范。在北京大学古籍数字图书馆的建设中，鉴于古籍拓片年代已久，并不存在版权问题。但由于收藏类型众多，古籍印本、写本、家谱、舆图、敦煌卷子、拓片的情况各不相同，标准规范方面的工作相当复杂。

古籍数字图书馆涉及的标准规范可以从下列系统的流动过程中体现出来：

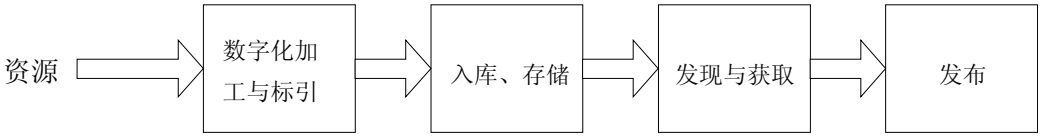


图 1：古籍数字图书馆系统流动过程

具体包括：

1. 数字化加工标引过程中：

- 数字化加工标准：对不同类型的资源，如印本、写本、舆图、拓片、敦煌卷子等进行扫描加工的标准，包括加工用途、加工级别、精度及色彩要求、存储格式等；

- 元数据标准：用于描述对象，并对对象进行定位、管理，且有助于它的发现与获取的数据，即为元数据；元数据标准则是如何描述某类资源的某个对象的所有规则的集合，如拓片元数据标准、古籍元数据标准等。每一类元数据标准又可分为描述元数据、管理元数据、应用元数据三种类型，其中描述元数据是针对不同类型资源设计的，管理和应用元数据则在整个数字图书馆中通用。
- 知识管理相关标准：包括知识分类体系等相关标准，以及人名、地名规范等；
- 数据封装标准：将元数据与对象数据封装在一起的标准规范，以便日后保存与存取。

2. 数据存储标准，如存储格式等。

3. 当信息对外提供服务，被外界通过检索手段发现、定位和获取时，涉及的标准有：

- 元数据交换标准：规定用于交换的元数据格式；
- 支持互检索的协议：如 OAI（Open Archive Initiative）协议等；
- 结果集的整合排序规则：如结果集的整合规范、排序规则等；
- 权限管理相关标准：权限描述与管理的规则与规范；
- 对象库的存取协议；
- 对象发送的封装格式、标准；
- 电子商务的相关标准等。

4. 应用服务标准：如文献传递、信息推送、在线参考咨询等方面的标准。

下面简单介绍一部分已经制定的标准：

1. 元数据标准

在《北京大学数字图书馆中文元数据标准框架》指导下，制定了古籍元数据标准和拓片元数据标准两个描述元数据标准，以及管理元数据标准、应用元数据标准。其中描述元数据标准的组成如下：

| | **拓片** | **古籍** |
|---|---|---|
| **核心元素** | 1、 题名 | 1. 题名 |
| | 2、 责任者 | 2. 主要责任者 |
| | | 3. 其他责任者 |
| | 3、 关键词 | 4. 主题词与类名 |
| | 4、 附注 | 5. 附注说明 |
| | 5、 金石年代 | |
| | 6、 金石类型 | |
| | 7、 资源形式 | 6. 资源形式 |
| | 8、 拓片标识 | 7. 古籍标识 |
| | 9、 铭文语种 | 8. 语种 |
| | 10、相关资源 | 9. 相关文献 |
| | 11、时空范围 | 10. 时空范围 |
| | 12、馆藏信息 | 11. 馆藏信息 |
| | | |
| **本馆核心元素** | 1. 版刻/版本 | 1. 版本 |
| | 2. 外观特征 | 2. 外观形态 |

| | | |
|---|---|---|
| 个别元素 | 1. 拓片收藏历史（Collection history） | 1. 收藏历史 |
| | 2. 书法特征（Handwriting） | |
| | 3. 金石所在地（Location） | |
| | 4. 金石材质（Materials and techniques） | |

在各元数据标准中，按照一定的定义规则和方法分别对每个元素及其子元素、限定词下了定义，并与 Dublin Core 做了对映。为检验元数据标准是否能够准确、充分地揭示资源，并广泛听取同行对元数据的意见，先行开发了古籍、拓片的著录实验系统，目前已有国家图书馆、上海图书馆、辽宁图书馆、台湾中央研究院图书馆等参与了试验著录。

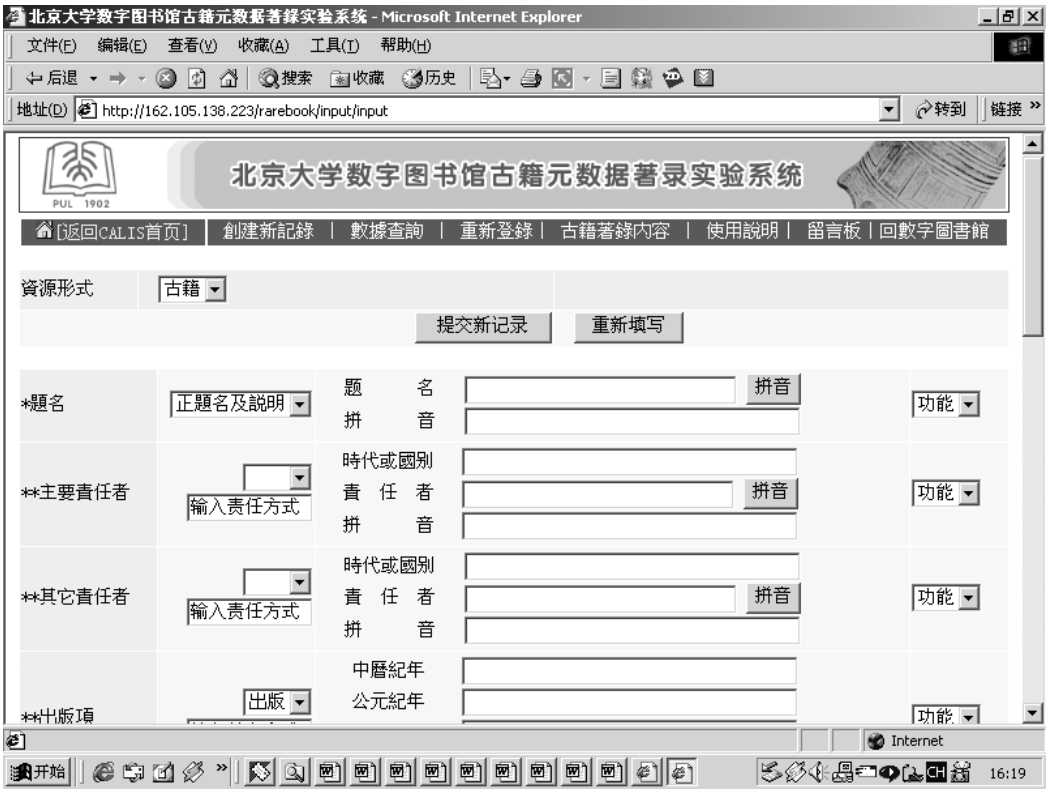


图 2：北京大学古籍数字图书馆著录试验系统

管理元数据：

即北京大学数字图书馆的管理元数据标准。借鉴 OAIS（Open Archive Information System）的分类机制，主要由上下文信息（context information）、出处信息（provenance information）、验证信息（fixity information）、内容表述信息（content representation information）、评价信息（remark/comments）等五个方面的元素组成。

应用元数据：

为便于通过地理信息系统（Geographical Information System）来访问时空属性很重要的拓片、古籍，特别设立地理信息系统元数据（GIS metadata）项，用来描述资源对象的地理时空属性，包括 2 个元素：空间项（spatial），即数字对象所涉及的空间信息；时间项（temporal），

即数字对象所涉及的时间信息。

2. 非数字化资源的数字加工标准

对每种资源的加工级别、色彩要求、保存格式、精度均做了详细规定，例如对善本的加工标准：

| 品种 | 类别 | 用途 | 级别 | 色彩要求 | 格式 | 最低精度要求 | 备　注 |
|---|---|---|---|---|---|---|---|
| 珍善本 | 原书 | 珍藏，精密印刷，网上浏览 | A | 24 位 彩色 | TIFF | 600PPI | 页面向上扫描，如做 OCR 则参照对普通古籍的要求 |
| | | | P | | JPG | 600PPI | |
| | | | L | | JPG | 300PPI | |
| | | | M | | JPG | 150 PPI | |
| | | | S | | GIF | 72PPI | |

3. 古籍拓片知识组织体系

以元数据内容为基础，初步搭建了古籍数字图书馆的知识组织系统，并在建设过程中逐步完善。

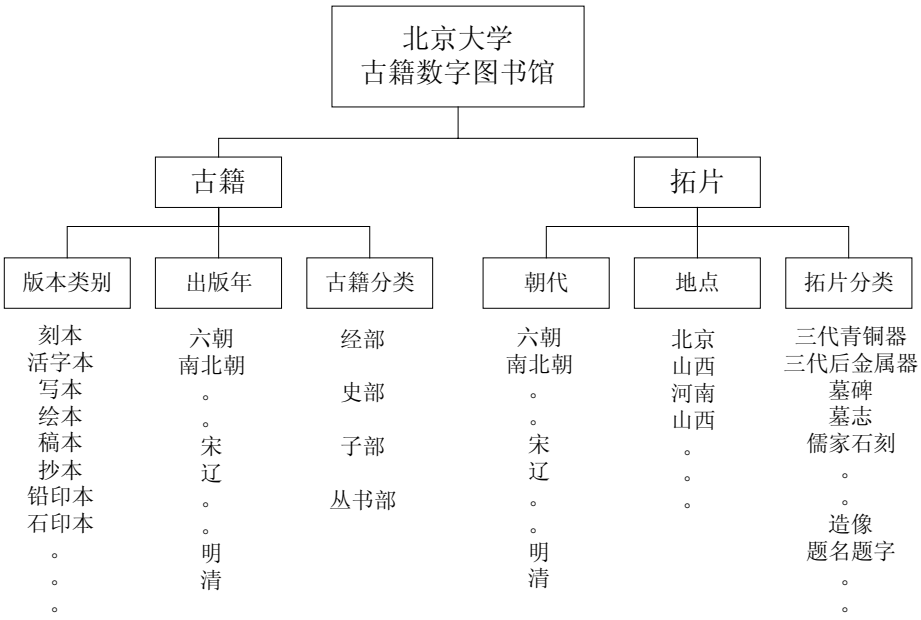


图 3：古籍数字图书馆知识组织系统

该系统是一个树状结构体系，用户可以从古籍数字图书馆入口进入，按照古籍的知识组织体系层层深入，在叶子节点找到相关信息。例如：古籍数字图书馆→拓片→拓片分类→三代青铜器→拓片名称→大盂鼎，即是这样一个典型应用。

4. 结果集的整合排序规则

由于出版和印刷等特殊因素所致，古籍和拓片都存在这样的现象，一种书或一种拓片有不同的版刻、版本或版印，并作为不同的对象记录在不同的元数据记录里。例如，来自一块石碑的几份拓片，因为拓印时间或技法的不同，被分别作为几个不同的记录来处理，并通过元数据来描述他们之间的关系。这种方式与现代印刷型书籍的著录方式有所不同。

因此，当用户检索的时候，就存在不同级别的结果集整合问题，即如何把内容相同，但记录不同的对象整合在一起显示给用户，整合、排序的规则又是什么。在北京大学古籍数字图书馆，古籍采取了版刻、版印、复本的三级整合方式，拓片采取的是版刻、版本、复本的三级整合方式，即在系统上搭建了三层结构，再通过不同的记录标识解决了这个问题。

## 三、系统结构与相关技术的应用

1. 系统结构与相关技术应用

古籍数字图书馆系统结构首先基于北京大学数字图书馆系统的总体框架，该框架采用多层结构，每层的功能相对独立，每层之间留有标准接口，不但可以保证系统的可扩展性、灵活性与开放性，同时也能方便地进行系统接入与管理。


图 4：北京大学数字图书馆总体框架

其中，数据服务层负责管理数据的物理存储与读取、以及内部的数据维护。存取层则负责检索数据库、存取对象、资源调度等用于支撑上层应用的服务。事务处理层（即应用服务层）则处理在数字图书馆环境下的应用服务，如在线参考咨询、用户认证、信息推送、电子商务等。最上层的表示层负责与用户界面以及与用户的交互，在该层用户可以通过数字图书馆门户网站接入数字图书馆系统，也可以由专用客户端软件进行接入。

根据北京大学数字图书馆的总体框架，并结合北京大学古文献特藏的特点，我们设计了以下的系统模式。

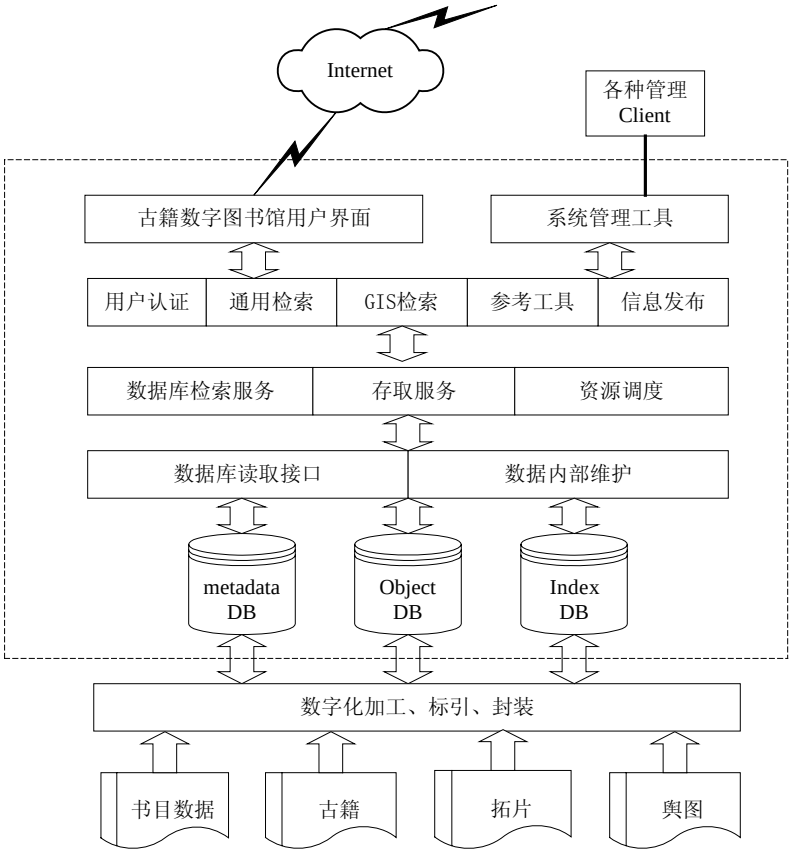


图 5：北京大学古籍数字图书馆系统结构

对于古籍拓片资源，除去采用简单检索、复杂检索等通用检索功能外，还提供了以下可以充分揭示资源特性的检索技术：

- 结合 GIS 检索技术的辅助检索工具，用户可以通过地理信息系统检索古籍拓片，突破了传统的文字检索模式，也使历史文化资源的时空特性得以充分揭示。

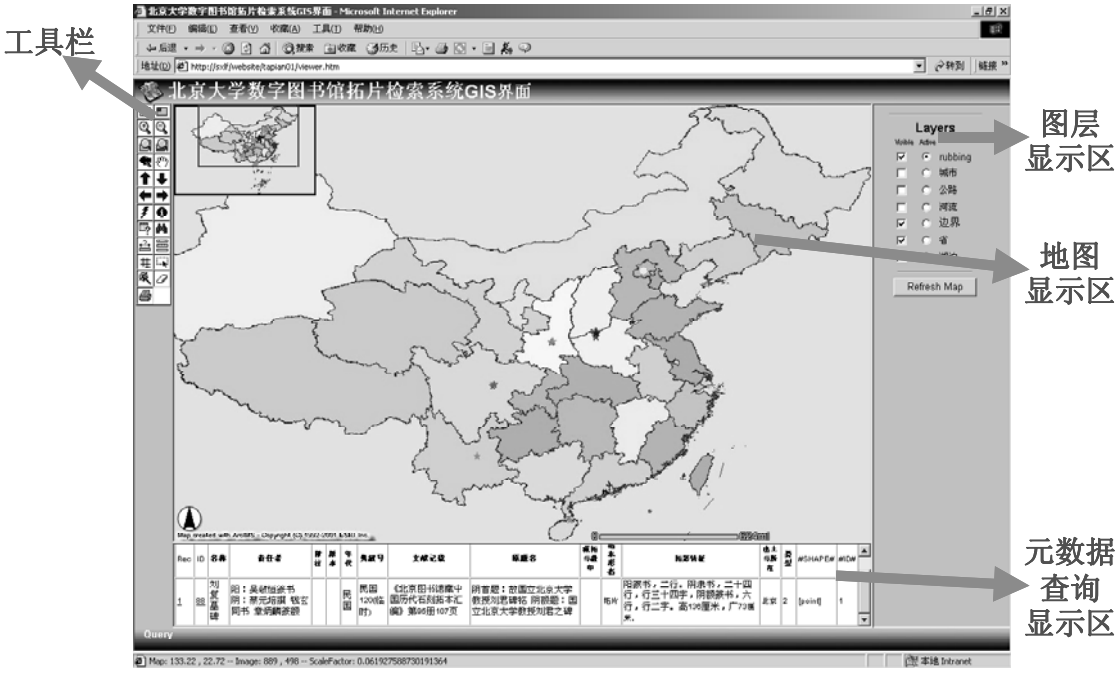


图 6：北京大学古籍数字图书馆 GIS 检索界面

- 按照知识组织系统搭建起来的树状浏览结构，如用户可以按照拓片类型分类浏览。
- 古代人名、地名索引，用户可以输入任意一词，定位在索引的某一位置上，上下浏览相近词汇。
- 同时选择支持简繁体互检的搜索引擎。

此外，系统还将整合多种与古文化相关的参考工具，如中西历转换工具、康熙字典、古今地名对照、人名规范等等。这些参考工具大多来自于其他出版商和数据库商。

2. 系统的开放性与互操作性

首先，在资源数字化的环节，依据制定的相关元数据标准以及数字化加工规范，对数字对象资源进行标引与封装，使之成为符合标准的数字对象。凡是符合封装标准的数字对象，都可进入北京大学数字图书馆系统。通过数据管理工具，将对象、元数据装载到相应的数据库中进行存储与管理。

在进行资源发现时，系统采用支持互操作的检索协议，如 OAI，可以方便地进行不同资源数据库的互检。通过统一的资源命名规则，系统可唯一定位到所要请求的资源。同时对不同资源的存取，采用统一的数据/对象存取协议进行，然后根据对象发送的封装标准将对象结果打包成规范格式发送给代理服务器发送到客户端。

根据这样的系统模式，数据提供单位或资源服务提供者，只需要在系统的某一层上符合规定的接口标准，就可以方便地整合到数字图书馆系统中。

## 四、服务系统建设

服务是资源建设和系统建设的最终体现，也是最能反映数字图书馆价值的所在。北京大学古籍数字图书馆为用户提供了多种服务功能，这些功能构成一个综合性服务系统，并在数字图书馆门户上统一整合并应用。

服务功能包括：

用户认证（user authentication）：这里主要指的是用户一次性权限认证，即用户通过北京大学数字图书馆门户网站进行一次认证后，系统不再要求用户反复提供用户名和密码，而是自动将用户信息传递给各个不同的数据库。

检索（retrieving）：如上文所提，提供包括通用检索、地理信息检索、浏览、索引等多种检索方式在内的检索功能，帮助用户尽快找到所需信息。

数据下载：为用户提供限制性批量下载元数据的功能，如一次可以下载 30－50 条记录等。

检索辅助服务：在检索的同时，允许用户保存检索史，以便随时查询自己的检索记录，并在各个页面设立检索帮助（help）。

在线参考咨询（virtual reference）：包括三种方式，一是为用户提供一个提问接口，用户可以随时发送与利用资源相关的各类提问，并得到参考咨询馆员的答复和帮助；二是总结归纳用户提问，设立 FAQ（常见问题解答）栏目；三是提供各类参考咨询工具，如中西历对照表，古今地名、人名对照，康熙字典，中国大百科全书等。

在线培训（web training）：为用户使用该数字图书馆提供培训课件，用户可以随时学习

如何使用系统查询自己所需信息。

推送服务（push service）：根据用户的要求以及保留在系统中的需求（如检索词、发送频率等），向用户定时、主动报导资源与服务的更新情况。

文献传递（document delivery）：当用户索取文献时，如资源的原件（如影印古籍、影印拓片等）允许提供复制品，则向用户提供文献传递服务。

信息发布：通过某个窗口随时向用户公开报道资源与服务的变化情况。

## 结语

北京大学古籍数字图书馆试图通过上述资源、标准规范、系统和服务等方面的建设，为用户提供一个综合传统文化资源和现代化服务的数字图书馆，使这些珍贵的古文献特藏得到更进一步的保存和更广泛的应用。

**参考文献：**


1. 肖珑，陈凌等. 中文元数据标准框架及其应用. *大学图书馆学报*，2001，19（5）
2. 北京大学图书馆. 北京大学古籍数字图书馆项目建议书
3. 北京大学数字图书馆研究所，http://www.idl.pku.edu.cn/
4. 北京大学图书馆. 非数字型资源的数字加工标准参考方案
5. 李峰，王爱华. GIS 技术在古籍数字图书馆中的应用（会议报告）
6. The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.htm
7. Reference Model for an Open Archival Information System (OAIS), http://www.ccsds.org/
8. William Y. Arms, Christophe Blanchi, Edward A. Overly, An Architecture for Information in Digital Libraries, http://www.dlib.org/