中国古籍数字化的现状与展望

陈　力 （中国国家图书馆）

一、中国古籍数字化之现状

1、简单的历史回顾

利用计算机技术对文献进行加工处理，已经有很长的历史了。但对文献内容本身进行数字化，仅仅只有二十多年的历史。

中国古籍的数字化最早是从计算机事业最发达的国外开始的。七十年代末期，国外的 OCLC 和 RLIN 首先建立了《朱熹大学章句索引》、《朱熹中庸章句索引》、《王阳明大学问索引》、《王阳明传习录索引》、《戴震原善索引》、《戴震孟子字义疏证索引》等数据库，用计算机对中国古籍进行处理。

国内何时用计算机进行古籍整理，现在没有十分确切的资料，比较早是八十年代初彭昆仑先生完成的“《红楼梦》检索系统”(1983 年 11 月初步建成，但发布是在 1987 年)。1984 年 8 月 20 日第 127 期《古籍整理出版情况简报》刊登了《微电脑与古文献研究》，提出了关于古籍数字化的设想：

> 随着微型机数量的增加、功能发展以及分布的扩大，其信息的贮存量会愈来愈多，并在一定范围，从一个地区到全国以及世界各地组成网络，形成一个巨大的资料库，所有信息资源便可共享。实现了这个目标，我国几千年来汗牛充栋而又星罗棋布的古文典籍，可尽行收入方寸之地，召之即来。使用微型机对这些古籍进行版本研究、文句校勘、文字订正、字义诠释、篇章会注、作品编年、古语今译，乃至标点、分段等等都将成为现实。

二十多年来，中国古籍数字化的道路基本上是通过两个方面来进行的，一是利用计算机对古籍进行揭示，建立古籍的书目型数据库，方便读者检索使用；二是利用计算机对古籍的内容进行数字化，使读者不仅能通过计算机来阅读古籍，并且能够通过磁盘、光盘和网络进行传播。

2、古籍书目数字化

古籍书目数字化即古籍书目数据库的建设从八十年代就已经开始了，它经历了自主开发到统一标准、统一规范、联合开发的历程。

目前，中国古籍书目的计算机机读目录格式已有通用标准，在大陆地区主要采用 CNMARC，在台湾和香港地区则主要采用 CMARC，并且著录规则也有一些差异，一般而言，大陆的著录规则要详细一些，而台湾则稍微简单。

大地地区已有比较完备的相关标准、规范：

《中国文献编目规则》(现正修订)

《汉语文古籍机读目录格式使用手册》

《中文拓片机读目录格式使用手册》

台湾地区也有相应的标准规范。

从 2000 年起，两岸五地中文文献资源共建共享确立了大陆、台湾等凡收藏有中文古籍的机构开展古籍的联合编目项目，由台湾汉学研究中心负责，由于各种原因，目前进展缓慢。不过，两岸的古籍编目工作都在近年受到了高度重视，并已有很大的进展。台湾方面，已经建立了“台湾地区善本古籍联合目录” (116034 笔)。

大陆方面，国家图书馆 2003 年已经完成了全部 27 万册善本古籍和 160 余万册普通古籍的编目

工作，所有的数据都已经上网供读者使用并通过中国国家图书馆联合编目中心为图书馆界提供下载服务，目前正在进行已建数据的维护以及特种文献如金石拓片、舆图等的编目。其他一些大的公共图书馆如上海图书馆、南京图书馆等也都正在进行古籍的编目工作。高校部分，最近 CALIS 也在组织进行古籍的联合编目。

目前，中国国家图书馆还在进行古籍人名、地名等名称规范（Authority）数据库的制作。

目前，大陆地区进行的古籍编目工作大部分都仍然采用 MARC 格式，一些特种文献则开始尝试用 DC 格式进行编目。中国国家图书馆、北京大学图书馆、中国科学院图书馆等单位正在联合进行“ 中文元数据标准规范 ”的研究。

3、《中国古籍总目》的编纂情况

下面介绍一下与古籍数字化有密切关系的现存古籍的调查与编目工作。

1994 年，由国务院古籍整理出版领导小组负责组织的中国古籍总目编纂工作开始，到 1997 年由于各种原因暂停，从 2004 年 1 月起该项目又重新启动。

《中国古籍总目》以国家图书馆、上海图书馆、南京图书馆、天津图书馆、辽宁省图书馆、山东图书馆、浙江图书馆、湖北省图书馆、北京大学图书馆、复旦大学图书馆、中科院图书馆等十一家所藏古籍为基本馆藏，十一家已有收藏者，其他馆藏就不再著录，十一家均未著录者，则都予补入。

预计此项目将于 2005 年完成。

《中国古籍总目》是一个品种目录，此项目完成後，必将对今后古籍的数字化提供非常有用的参考。

4、古籍内容的数字化

古籍内容的数字化与书目数据的建设几乎同时起步，也已经历了二十多年的发展，目前已经成为中国古籍数字化的主流。下面，就简要介绍一下主要的情况：

八十年代，古籍内容的数字化刚刚开始，大部分的工作主要还是在学者的书斋中进行的，并没有对社会产生大的影响。进入九十年代以后，随着计算机的普及及网络技术的发展，古籍作为一种重要的民族文化遗产，受到了高度重视，因此在最早出现的一些读书网站中，如“ 黄金书屋 ”等，就已经有了数字化的古籍，这些数字化的古籍主要的内容包括古典小说、历代正史、儒家经典和诸子等等，形式主要是手工输入的一般电子文本。

台湾地区的古籍数字化

在台湾地区，从八十年代末，一些重要的研究机构就开始研发以古籍为主的大型的资源库，这里面最成功的要数台湾中央研究院开发的“ 翰典全文检索系统 ”，收录了不少重要的典籍。

除中央研究院外，台湾还有一个较为庞大的“ 数位典藏计划 ”，包括：

善本古籍典藏数字化

金石拓片典藏数字化

古籍附图典藏数字化

以及“ 台湾地区地方文献典藏数字化 ”和“ 期刊报纸典藏数位化 ”，具体的数字化数量此从略。

九十年代中期以后，在大陆地区一些大的出版机构、学术单位和商业公司介入了古籍的数字化工作，古籍数字化的规模迅速扩大，下面重点介绍一些影响较大的古籍数字化项目。

大陆地区的古籍数字化

书同文公司

书同文公司与台湾迪志公司合作，其开发的主要产品是《四库全书》和《四部丛刊》。

《四库全书》在汉字处理上颇具特色，采用 UNICODE，很好地处理了繁简字、异体字、避讳字等等。在内容处理上，以 DC 元数据和 XML 相结合。由于使用 XML 技术，使得对古籍内容的处理与交换符合目前数字图书馆的通行标准，为实现不同数据库之间的跨库检索提供了有利的条件。该数据库实现了版面还原、全文检索、字（词）频统计，并配有一些知识工具，如字典、干支换算等。

《四部丛刊》情况与《四库全书》相似，但由于《四部丛刊》所收各书字体、版式不似《四库全书》那样整齐划一，因此在汉字识别及版面还原方面难度更大。

国学公司

《国学宝典》v8.0 版数据工作正在处理，计划收书总数达到 3600 种，总字数约 7.5 亿汉字

《中国历代基本典籍库系列》，全套光盘分为“ 先秦两汉魏晋南北朝卷 ”、“ 隋唐五代卷 ”、“ 宋元辽金卷 ”、“ 明清卷 ” 四种，共收入三千多部（六亿多汉字）中国古代重要的典籍文献。

还有其他一些产品。

主要功能：全文检索、统计、摘录、打印输出、生成卡片、浏览

北京大学《中国基本古籍光盘库》

根据媒体介绍，该古籍库光盘将收录古籍万余种，每种典籍有 1 个通行版本的全文信息，另附 1 至 2 个珍贵版本的图像信息。预计全文 20 亿字，版本图像 2000 万页。该光盘采用了书同文的技术，因此其数字化的方法与书同文的产品基本相似，增加了版本对照。

北京大学图书馆

北京大学图书馆的古籍数字化内容计划包括馆藏敦煌文献、宋元版书、明代嘉靖、古代舆图、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）、 手绘本、家谱、古代戏曲、地方志等等，目前已有部分成果可以通过网上阅览。

关于其数字化的相关标准方面，该馆有如下考虑：根据对不同类型的资源，如印本、写本、舆图、拓片、敦煌卷子等制订相关的扫描加工标准，包括加工用途、加工级别、精度及色彩要求、存储格式等，同时根据不同类型的资源设计相应如拓片元数据标准、古籍元数据标准等。（以上参见肖珑、冯英：《基于古文献特藏的数字图书馆系统的设计与实现》，《文津流觞》第八期）

北京大学图书馆在古籍数字化方面的成就主要体现在相关的数字化标准上面，在数字化时也考虑到了相关工具的使用，如中西历转换工具、康熙字典、古今地名对照、人名规范等等。另外，在拓片的数字化方面，也考虑采用地理信息系统来进行检索，不过只是用于简单指示碑石的出土地，而如果拓片本身不能明确其地理方位而采用地理信息检索系统的话，甚至有可能出现漏查的问题。

从北京大学图书馆的古籍数字化流程图来看，主要是对原始资料进行数字化扫描，然后通过元数据的形式对其进行描述和管理，相关的参考工具也是在系统之外附加，即图像+描述的方式。

清华大学和浙江大学

由清华大学图书馆，计算机科学与技术系，清华大学建筑学院三方合作共同研制开发的“ 建筑数字图书馆 ” 现在仍在进行中，虽然就内容而言在中国的古籍数字化方面并不具有特殊的意义，但其采用的方法值得注意。在该项目中，他们根据《营造法式》所记述的建筑结构，用数字化的动画进行模拟演示，使枯燥无味的古籍内容变得有声有色、形象直观。可以说这是用数字化的方式对古籍内容进行知识重组。

浙江大学承担的《中美百万册数字图书馆》

中华书局的中华古籍语料库

由于该库尚未正式向外公布，具体情况不详。但据说主要是以中华书局标点整理本为基础进行古籍的数字化，其最大长处在于古籍的底本选择上较好。

中国国家图书馆的古籍数字化工作（详后）

宗教文献的数字化

宗教文献数字化的代表有 CBETA 中华电子佛典协会的“ 线上藏经阁 ”，该数据库采用 XML 对佛教文献进行数字化。

网上主要中文古籍数据库调查统计表

二、中国古籍数字化工作之检讨

1、古籍数字化的格局

目前，中国大陆的古籍数字化的格局基本上由三大部分构成：一部分为教学和研究机构，一部分为图书馆，还有一部分则是商业机构。

2、不同机构古籍数字化的特点

上述三部分在进行古籍数字化时是各有其特点的：

教学和研究机构对数字化对象选择目的性强，数字化的目标及方法主要是根据教学和研究工作需要来决定，例如中国社会科学院的数字化项目包括：《全唐诗》《先秦魏晋南北朝诗》《全上古三代秦汉三国六朝文》《十三经》《全唐文》《诸子集成》等等，北京大学的《全宋诗》、深圳大学的《红楼梦》皆是如此。

图书馆所进行的古籍数字化，则主要是根据其馆藏特色来进行，如北京大学图书馆、中国国家图书馆的古籍数字化项目基本上是按这个原则来规划的。

至于商业机构对古籍的数字化主要是根据市场来决定的，哪一类文献有市场，就进行哪一类文献的数字化，考虑到市场的动作，常常选择大型古籍丛书如《四库全书》、《四部丛刊》等等。

目前古籍数字化工作中存在的问题

古籍的数字化是一项文化遗产的保护和弘扬工作，具有浓厚的公益性色彩，需要各方面加强协调，有一个整体的规划。整体规划不仅包括数字化对象的内容确定和合作分工，同时包括相关标准、规范的统一。

1、协作方面的问题

关注焦点过于集中，重复建设。

中国古籍的数字化目前是各自为阵，虽然数量已经不少，但关注的焦点过于集中，并且多数都带有商业性或者追求规模与宣传效应，致使古籍的数字化集中于“ 少数 ”常用特别是丛书类的古籍，而大多数学术界需要的古籍的数字化无人顾及。例如文渊阁《四库全书》先后已有四家进行影像的数字化（上海、山东、武汉、浙江大学等），一家进行了影像、全文文本的数字化（书同文）。

利益不同，各自为阵，封闭建库。

由于制作单位不同，各自的利益不同，所制作的古籍数据库常常是封闭的，在技术上很难与其他数据库融为一体，造成知识体系的割裂。

标准规范不统一。

出于各自利益的考虑，不能协同作战，特别是统一标准与规范，实非易事。

属于数字化本身的问题

应该说，上面所提到的几类问题，都是属于表面层次的问题，是比较容易发现的。在我看来，目前的古籍数字化还存在另外两方面的问题：

第一、对古籍数字化工作的定位不够明确。古籍数字化与其他文献的数字化是什么关系？它在整个数字图书馆建设中处于什么样的地位？

第二、对古籍数字化工作的特点认识不够。古籍较之其他类型文献有什么特点？如何在数字化时体现这些特点？

三、古籍数字化之展望

1、古籍数字化与数字图书馆建设

古籍数字化是数字图书馆建设重要的组成部分

古籍数字化的成果，就目前的情况而言，无论其制作单位是什么，但读者大多是通过图书馆或其他一些文献收藏机构来利用这些成果的。由于各大系统各有其考虑、各有其利益，因此在标准与规范方面难以统一，并且独立成库，互不开放，不仅难于与其他古籍数字化项目共享资源，也很难纳入各图书馆整个的文献资源体系之中。

我们认为，古籍是文献的一种，古籍的数字化不应该与现代普通文献的数字化割裂开来，也不应该与目前各图书馆使用的书目型数字库截然分开，我们希望在一个通用的平台上，读者既可以进行一般性的书目包括现代图书与古籍的查询，同时根据需要可以直接切换到古籍甚至相关的现代研究性著作的全文上。为此，应该注意：

一、古籍数字化是数字图书馆重要的组成部分，换句话说，在数字图书馆中，不仅应该包含古籍方面的内容，还应该包括现代不同类型、不同内容的数字化信息，甚至，没有数字化的各种载体形式的文献也将通过某种形式（如书目的联接）与数字化的信息融为一体，从而构成一个完成的知识体系，因此，古籍数字化应该按数字图书馆的模式去组织、加工、发布。

二、古籍数字化应该以是开放式的、分层次的、结构化的数据库来组织与揭示，在进行数字化加工时应与现代图书遵循统一的标准规范，古籍的特殊性应该在统一标准规范的框架下进行细化，可以通过某种形式的共享协议或技术，使所有的资源能够在同一平台上使用并互相调用，以节省加工成本；

三、古籍数字化应该建立一些公用的知识库作为支撑，在大多数情况下，古籍与现代图书的知识库应该是可以共用的，如字典、历史年表、纪年换算、历法换算、各类规范文档、地理信息系统等等。

四、基于上述三点认识，各古籍数字化的制作单位应该密切协作，否则必将事倍功半。

2、古籍的特点与古籍数字化

如何进行古籍的数字化？国内已有学者进行过专题研究，北京大学李国新教授认为：第一是必须实现文本字符的数字化，第二是具有基于超链接的浏览阅读环境，第三是具有强大的检索功能，第四是具有研究支持功能。

李国新教授所列前三项是一般文献数字化都应该具有的，也就是说，并非古籍的特性。关于第四项，李国新教授提出的具体内容是：“ 所谓‘ 研究支持功能 ’是指能够提供有关古籍内容本身科学、准确的统计与计量信息，提供与古籍内容相关的参考数据、辅助工具。这些信息、数据或工具都是

古籍内容的增值或补充。比如古籍字数、字频、词频的统计数据，异体字的汇聚显示，读音的自动标注和朗读，行文风格特点的概率统计，必要的背景知识、参考数据的汇聚，在线标点断句工具的配备，不同版本比勘校对接口的设置，字典词典、历史年表、历史地图等研究工具的加载，等等。有了这些研究支持功能，不仅可以极大地改善研究者的研究条件，而且还会带来研究思路、研究方法的变革。”（〈中国古籍资源数字化的进展与任务〉，《大学图书馆学报》，2002 年第 1 期）实际上，如前面所介绍的，许多机构在进行数字化时也都考虑到了相关的工具的开发。除此之外，我们认为，古籍的数字化还有许多工作需要做。

我们认为，**古籍数字化应该根据古籍的特点来进行，数字化的过程是一个信息重组并上升为知识的过程。**

较之其他类型的文献，古籍有什么特点？

第一是版本的问题。

1、古籍的版本选择问题

古籍的整理都有一个版本的选择问题，在传统的古籍整理工作中，也是学者们普遍遵循的原则之一。就古籍的数字化而言，目前的古籍数字化由于许多要追求商业利益或规模效益，因此在许多公司的图书馆进行古籍数字化工作时，通常喜欢选择丛书或一类特色文献来进行古籍数字化，而不是根据版本来安排数字化文献。因此，今后的古籍数字化应该引入专家都古籍的版本进行筛选，尽量提供好的版本进行数字化。

2、与古籍版本选择直接相关的就是古籍版本的比较问题。许多古籍都不止一个版本，虽然我们可以勉强说某种古籍的某个版本比较好，但这并非绝对的，因为不同版本之间的异同也许各有长短，同时，根据对不同版本异同的分析我们也许能从中了解更多、更重要的信息，因此，版本的比较就非常重要了。在传统的纸质文献下，我们常常会蒐集不同版本的古籍进行比较研究，在数字化时代，直接采用扫描的数字影像文献由于阅读不便，因此即使我们有了不同版本的数字影像文献，但使用起来会非常困难，远不如纸质文献。如何利用现代信息处理技术来处理不同版本的比较问题将是我们今后必须考虑的，这也是提高古籍数字化水平最重要的途径之一。目前中华书局正在进行的“ 中华古籍语料库 ”，根据已经由学者进行过校勘整理的古籍来数字化，这也许是其中一个比较好的解决方案。当然，“ 中华古籍语料库 ”由于主要是根据现代整理本来进行数字化，它可能较好地解决不同版本的校勘问题，但没有各种版本的版式、图像，美中不足。

3、已有古籍研究成果的利用问题

在以往的古籍数字化过程中，业界通常采取原版图像和全文检索本配合使用的办法，这当然符合古籍本身的特点，但这种做法也带来了一些问题，其中最大的问题就是前人在古籍研究和整理方面（包括校勘、批注等等）的成果难于被利用。如果按现行的做法，全文检索本只是原版的 OCR 产品，而原版绝大多数都不能反映前人对其进行研究的成果。如何既要向读者提供反映古籍原貌的影像，同时在全文检索和文本阅读时又能享受前人的研究、整理成果，这也是需要注意的。解决这个问题的方法，恐怕需要建立一定的知识支撑系统或者对文本本身再进行处理。

4、全文检索与规范控制

全文检索是古籍数字化最先受到重视的技术。简单的全文检索在几乎所有的文本编辑和对象数据库软件中都能实现，但为了防止过多的“ 噪音 ”出现，因此人们非常重视汉字的标引特别是词典切分标引。词典切分标引对于现代文献可能相对较易，但对于古籍，由于古籍及古代汉语的复杂性，在实践中更为困难，它不仅要解决防止“ 噪音 ”过多的问题，更要解决与名称规范相关的问题，如同书异名、同名异书、同一作者有不同的称谓，其他如职官、地名、事件名等都与现代很不相同，例如李世民=唐太宗、南京=天京（太平天国）、太平天国=洪杨之乱等，非常复杂，这是一个尚待研究的课题。

上述问题有些在传统的文献整序时已经有了解决的办法，也就是我们在文献编目时经常要提到的“ 名称及主题规范 ”、“ 权威档 ”（Authority），通过对文献进行规范控制，我们可以基本上解决一

般性的异名问题。但是，由于古籍的数字化同文献编目不同，它主要是对文献内容的数字化处理，而不是对文献某些特征的抽象性描述，有些问题可能需要建立一些知识性的支撑数据库（或工具库）来解决，如对古籍中地名、职官名的处理。以地名为例，古籍中的地名与今天的地名很不相同，一地有数名，一地的辖区在不同时代各不相同等等，这使得古籍中的地名规范处理起来非常困难，因此，可能有必要考虑建立一个以现代地理信息系统（GIS）的方式构建的古代地理信息系统，作为全国乃至全世界同行在进行古籍数字化时通用的知识支撑系统，这个系统并不简单是附上一个现代的电子地图，而应该正确地反映不同时代政治地理的变迁情况，同时辅以古代地名名称规范数据库。这将是一项极为浩大的工程，需要各方协作。

中国国家图书馆过去进行了大量的规范控制研究与规范数据库制作，目前已有中文名称及主题规范数据库 40 余万条，其中包括古籍题名规范、人名规范等等，下一步可能要考虑一些更深层次的问题。

3、中国国家图书馆的古籍数字化

数字化建设项目

自九十年代以来，中国国家图书馆开始了古籍内容的数字化工作，目前，国家图书馆数字资源的建设也是以古籍的数字化为主，在刚刚通过的国家图书馆数字资源建设(2003-2005 年规划)中，确定了今后几年数字化的重点项目有：

数字方志
碑帖菁华
敦煌遗珍
西夏碎金
甲骨世界
《永乐大典》
中国研究资源库
馆藏各类文献书目数据库

前六个资源库都属于古籍类，中国研究资源库中也有一部分涉及古籍，如中国历代法律文献库。另外，西夏文献的数字化已经完成，敦煌文献正与英国国家图书馆合作进行数字化。结合古籍的修复，我们已在逐步建立古籍修复档案数据库，在建库的过程中已有一些新的发现，如宋元本图书的用纸问题等等，相信这个数字库都全世界研究汉籍的学者和图书馆古籍修复人员都是很有参考意义的。

这里重点要介绍一下“ 数字方志 ”项目。

数字方志

国家图书馆所藏中国各地方志居世界之首。方志作为地方性的百科全书，对于传统文化的研究具有重大的意义。同样，对于古籍数字化，对地方志的数字化在技术上也有其特殊意义。

地方志从文献内容和结构来说，具备几乎所有类型古籍的特征，它所涉及到的不仅有一般古籍的版本、版式、文字识别等等，还涉及到舆图、地理沿革等等一般古籍所不具备的问题。另外，地方志较之一般古籍更具有可扩充性，我们不仅可以将现在一般古籍数字化时所用到的一些工具库如各类词典库等引入，同时我们还可以引入地理信息系统解决地理沿革变迁的问题；引入现代多媒体资料扩充某些方志的内涵，如谈到泰山，我们可以链入相关的图像、音频、视频和其他多媒体资料，使方志本身的信息得到扩充。可以说，“ 数字方志 ”不仅是中国国家图书馆一个古籍数字化的项目，而且，我们是将作为未来数字图书馆的一个模型来看待的。

数字方志的结构我们初步设想是：

1、对方志进行图像扫描，为读者提供完整的原始信息；

2、对方志扫描图像进行识别，以达到全文检索和其他现在一般古籍数字化所能实现的功能；

3、为更好地解决全文检索所遇到的因人名、地名、事件名古今称谓不同而导致的漏检问题，我们考虑引进传统图书馆在解决这类问题时所采用的规范控制概念，解决人名、地名、事件名等涉及名称规范的问题，在结合我馆在建的古籍名称规范、人名规范、地名规范数据库，再拟建若干专题数据库，如更完整的地名库、更完整的人物库、景观库、插图库、作品库以及研究文献库等等，不仅解决名称规范的问题，同时使这个数据库具有知识扩充的功能。

4、考虑到方志在地理方面的特殊性，打算引入地理信息系统的概念，以解决古今地名及地域的变迁。

5、数字方志只是国家图书馆众多资源库中的一种，并且它与其他资源库甚至与馆藏纸本文献从知识层面上看都有着或多或少的联系，因此数字方志资源库它不应该是一个封闭的数据库，而应该具有开放性，既能与其他开放型数据库共享知识工具包括各种书目型数据库、规范文档，还能与其他文献整合起来，为读者提供更为全面的知识性的服务。

6、对读者而言，除了实现基于数据库底层数据库管理的全文检索以外，这个数据库还应该具有：原版再现、不同版本的比勘、与方志库以外的其他数据库的跨库链接、个性化编辑、繁简转换以及其他功能。

展望未来的古籍数字化，我们不仅依赖于技术的进步，更依赖于同行间的合作与用户的帮助。我希望读者、同行既是古籍数字化的使用者，更应该是古籍数字化的参与者。