

中国古籍特藏资源数字化调研报告

曹菁菁

数字图书馆和大型电子资料库的开发和建设,近年来业已成为世界各国大型图书馆以及学术研究机构的一个重要发展趋势。其中,古籍资源的数字化得到了跨国界的、跨学科的一致重视。并且,伴随着技术的不断改进以及一些重要数字化产品的相继发行,数字化的古籍资源在学术研究和文化教育中的利用不断扩展。在这个新领域的较量当中,质量和速度一样重要,如果不加强古籍资源的数字化建设,在日益激烈的古籍数字资源开发和信息处理水平的国际竞争中,我们将丧失优势。

本报告将在分析当前国际范围内中国古籍资源数字化发展状况及趋势的基础上,对我馆当前的古籍资源数字化现状进行分析,并力求找出“十三五”规划期内,科学发展我馆古籍资源数字化的有效途径,提高我馆古籍资源数字化成果的竞争力,提升我馆古籍数字化服务的水平,全面提高我馆在古籍资源数字化方面的国际声誉。

本文所提出的“中国古籍特藏资源”,包括汉文古籍、少数民族文字古文献、金石拓片、西域和敦煌文献以及近现代名家手稿、新善本、舆图、老照片、年画、宣传画等特藏,同时还包括殷墟甲骨。本文将对这些资源的数字化开发进行相关调研。

一、国际范围内古籍资源数字化的现状

中国古籍的收藏不仅仅限于中国大陆地区。日本、韩国等东亚国家,美国、加拿大为代表的北美地区、欧洲各国以及澳大利亚都有收藏中国古籍的重要单位。近年来,这些收藏单位都开展了规模不一、程度不同的古籍资源数字化工作。

1.美国中国古籍数字化重要项目概述

A.美国国会图书馆的“联机目录”:包括中文古籍 75 万多册,其中有宋元明清善本、珍本资料。该系统提供浏览检索、高级检索以及关键词检索功能,浏

览检索可以按题名、作者、主题、分类号、标准号等进行检索，用户检索中文资料时需要用汉语拼音输入检索条件，数据结果以英文形式展现。

B.加州大学伯克利分校东亚图书馆“拓片古籍数字化计划”：加州大学伯克利分校东亚图书馆藏有丰富的中、日、韩历史文献。截至目前，该项目完成 200 多部古籍的数字化并进行发布，平台具有目录检索和关键词检索功能，可进行在线阅读。

C.中华古籍善本国际联合书目系统（中国国家图书馆与普林斯顿等大学合作）：该系统由中文善本书国际联合目录项目演变而成，于 2010 年 5 月 20 日正式开通。中华古籍善本国际联合书目系统收录清乾隆六十年（1795）以前在中国印刷或抄写的中文古籍，著录内容包含著录编号、题名、责任者、版本类型、版本信息、装帧形式、载体形态、行款版式、存卷及补配情况、题跋钤印、附注、四部分类、收藏单位和典藏号等十四项信息。

D.哈佛大学哈佛燕京图书馆藏中文善本古籍特藏（中国国家图书馆与哈佛大学图书馆合作）：计划自 2010 至 2015 年完成哈佛燕京图书馆的中文善本和齐如山专藏共计 4210 种古籍的数字化工作。该平台首批发布资源 204 种，支持简体中文、繁体中文、汉语拼音检索，系统设置了题名、责任者、出版时间、出版地、出版者 5 个检索字段，可阅览全文影像。

2.欧洲中国古籍数字化重要项目概述

A.英国藏中国古籍联合目录（UK UNION CATALOGUE OF CHINESE BOOKS）

1980 年始，英国国家图书馆采用 MARC 建立“古籍简明标题目录”，主要收录 15 世纪用活版印刷的文献题录，同时收录部分中文古籍书目。后来，英国国家图书馆联合牛津大学、剑桥大学等 6 所大学、研究所建成英国藏中国古籍联合目录，该联合目录包括英国国家图书馆藏有的中国古籍 6 万多种及其他单位所藏的中文古籍，提供按关键词、作者、日期、图书馆等进行检索的功能，用户可输入中文或汉语拼音来查询书目，所得信息也用中文显示。

B.国际敦煌项目（IDP，中国国家图书馆参与）：该项目是由英国国家图书馆发起的一个具有开创性的国际合作项目，目的是使敦煌及丝绸之路东段相关遗址出土的资源能在互联网发布，供用户使用。参与机构有中国国家图书馆、圣彼

得堡东方学研究所、柏林勃兰登堡科学与人文科学院、法国国家图书馆、匈牙利科学院图书馆等。经过多年的努力，该项目在敦煌文献的保存与修复、数字化建设以及资源共享方面取得了突出的成就，并建立了专题数据库对数字化的资源进行发布。目前世界各地的用户可通过该库检索浏览 41.7 万余幅的绘画、文物、历史照片。

3. 日本中国古籍数字化重要项目概述

A. 最大的汉籍联合目录数字化系统——“全国汉籍数据库”检索系统：该系统旨在网罗日本国内的公立和私立图书馆、大学图书馆、科研单位等相关机构所典藏的汉籍书目，从而为读者提供统一、便捷的书目检索服务。它是日本目前参与机构最多、搜索范围最广的汉籍数据库系统。截至 2012 年，多达 35 家收藏单位的 62 万条数据进入该系统，并主要由京都大学人文科学研究所附属东亚人文情报学研究中心承担主要的运营与管理工作。

B. 东京大学东洋文化研究所汉籍影像数据库：该研究所藏有大量中文典籍，包括原“东方文化学院”藏书、大木文库、仓石文库等，总数量有十万册，且有相当多的善本。该库主要收录了该所收藏的 3753 种古籍的全文影像数据，可以全文浏览。其所收录的以明清古籍为主。具有以下检索功能：基本检索、进阶检索、分类检索、浏览检索。基本检索即在文本框中输入关键词在整个资料库中搜索。

C. 日本国立国会图书馆珍稀古籍图像数据库：该平台对中文古籍也做了整合。目前已有 3000 余件中文古籍可在该库检索阅读。

4. 台湾地区重要中国古籍数字化项目概述

台湾中央研究院是古籍数字化的重镇，其资源的影像效果和元数据信息质量较高，但不提供免费服务。其重要古籍藏品数字化项目有：

A. 文物图像研究室：含简帛金石数据库、居延汉简补编图像、汉画论文目录、武氏祠画像数据库、安丘董家庄汉墓画像系统、番社采风图，以上数据库皆可全文检索，部份数据库且可检索图像。

B. 史语所藏内阁大库档案：史语所所收藏的内阁大库档案是宣统元年整修大库时由其中移出的一部分，包括诏令、题奏、移会、贺表、三法司案卷、实录稿

本、各种黄册、簿册等每件档案依序列出事由、时间、职官、册别件号、登录号、文件别、备注等项

C.瀚典全文检索系统：完成二十五史、十三经、诸子、古籍十八种、古籍三十四种、大正新修大藏经（部分）等文献的数字化工作，并以上述名称归类建档。后增加上古汉语语料库摘要、台湾方志与文献、清代经世文编及续编、中华民国史事日志等专业研究资料。该所决定将数据库全部打散、重整，将所有数字化文献依其内容性质归并为“史书”、“经书与子书”、“宗教文献”、“医药文献”、“文学与文集”、“政书、类书与史料汇编”六大类。

D.傅斯年图书馆善本书目全文资料库：该馆藏有 4 万 4 千册善本书、3 万 8 千余件金石拓片、1 万余件汉简。部分馆藏已可以上线浏览影像，但需要授权。

5.国内重要中国古籍数字化项目概述

很多古籍重点收藏单位都建设了自己的数据库，但是其规模较小，所提供的文献服务也非常有限。其中较有规模的大型数据库有：

A. 国家图书馆数字资源：

古籍特藏的所有目录检索（在 OPAC 系统中实现）；宋元善本、甲骨世界、碑帖菁华、方志、家谱等数据库。

善本古籍：目前提供 241 种古籍的在线阅览；

数字方志：提供 6868 种 230 万简子叶的方志文献全文影像浏览；

中华寻根网：提供 2356 种 250 万简子叶的家谱文献全文影像浏览；

碑帖菁华：提供馆藏石刻资源 19000 余条，影像 25000 余幅；

甲骨世界：提供甲骨实物影像 3500 种 15000 余拍；

此外，年画数据库有影像资源 400 张、老照片数据库有影像约 7000 拍等。

B.北京书同文数字化技术有限公司：

《四库全书》、《四部丛刊》全文检索系统。

C.北京爱如生数字化技术研究中心：

主要产品有 001 中国基本古籍库 002 中国近代报刊库 003 域外涉华文献库 004 中国方志库 005 中国谱牒库 006 中国金石库 007 中国丛书库 008 中国类书库 009 中国辞书库 010 中国经典库 011 中国史学库 012 中国俗文库 013 历代别集库 014 敦煌文献库 015 明清档案库等。

书同文及爱如生公司出版的这些古籍数据库，均可以实现全文检索，但是他们的数字化来源都是已经出版的古籍或者古籍目录，而非古籍本身。

二、国际范围内古籍资源数字化现状分析

1. 古籍数字化对象范围（以古籍类型为中心）

A. 汉文古籍：几乎所有的相关收藏单位都已经开始或完成了各自汉文古籍藏品基本信息的数字化工作，但是信息的公布和利用还有待进一步的发展。关于汉文古籍的各种数字化成果最为丰富，数字化程度也最为深入。可以说，汉文古籍的数字化引领着整个古籍数字化领域发展的方向。在这一领域，走在前列的几个代表性成果有：“中国历代典籍总目”（中国国家图书馆）、中华古籍善本国际联合书目系统（中国国家图书馆）、香港大学冯平山图书馆藏善本书录（中国香港大学）等。

B. 少数民族文字古文献：这类古籍的数字化处于起步阶段。由于大多数少数民族语言文字的计算机国际编码尚未开发，多种民族语文还无法显示，更无法在一个操作平台上进行信息处理。这使少数民族语文古文献的数字化受到很大的限制。当前这类资源的数字化成果主要是两种：一是原物影像辅以汉文信息，如满文文献¹；另外一种是用特殊的数据库收纳数字化资源，但兼容性和扩展性都有限，有很多技术问题需要解决，例如蒙文文献的数字化成果²。

C. 金石拓片：这类古籍的数字化尚未全数完成。从已经公布的数字化信息来看，大陆和台湾地区在此领域走在了前沿。而大陆地区已经取得了国际上公认的成果，如中国国家图书馆网络在线的“碑帖菁华”和北京书同文数字化有限公司出版的光盘版《历代石刻史料汇编》。

D. 舆图：各收藏单位舆图藏品的数字化或正在进行中，或尚未启动，也属于开发潜力很大的一种古籍资源。但是这类资源的数字化程度却没有落后于汉文古籍。在大多数收藏单位还停留在基本信息采集阶段的时候，复旦大学历史地理研究中心、美国哈佛燕京学社、澳大利亚格林菲斯大学亚洲空间数据中心等机构合作开发的“中国历史地理信息系统”已经在历史地理信息处理方面占领了制高

¹ 中央民族大学的“满蒙文数据库”现阶段以满文文献居多，它的主体是一个影像数据库，但需要用汉文进行检索。现在该数据库只能在中央民族大学局域网中使用。

² 内蒙古大学图书馆建立了“中国蒙古文书目数据库”，该数据库把蒙古文 WPS OFFICE 输入法成功地衔接在汉文图书馆软件上，以实现终端机的检索功能。该数据库只能在本图书馆局域网上使用。

点。

E.老照片、年画、宣传画：这类资源的数字化出现了明显的地域不平衡性。收藏量巨大的中国地区还处在藏品信息的初步处理阶段，但是欧洲地区不但已经基本完成了藏品信息的数字化，而且已经有了成熟的数字化成果。如英国伦敦亚非学院的“Internet Mission Photography Archive”以及荷兰国际社会史研究所(IISG)的“Soviet, Cuban and Chinese Posters”。

F.西域文献和敦煌文献：敦煌文献资料的数字化项目——“国际敦煌学项目(IDP)”——是目前比较成熟的一个数字化成果，在数字化的技术和质量上都属于上层水平。中国国家图书馆是该项目的主要参与者。

G.甲骨：甲骨的几大主要公立收藏单位都在进行藏品的数字化。其中，只有中国国家图书馆、香港大学等少数单位对外公布了自己的数字化成果。

2.古籍资源数字化产品的几个类型

A.书目型数字化产品：以最基础的藏品信息数字化为代表，辅以影像数据。

B.提要型数字化产品：在编目型数字化成果上加入藏品深度标引信息以及相关研究信息的数字化产品。

C.全文型数字化产品：提供古籍内容的全文内容，并有简单的节点浏览以及关键词搜索功能。

D.智能型的数字化产品(含校勘、辑佚、索引、系年、标点等知识重组功能)工具，例如“全宋诗分析系统”中的智能格律标定工具、“全球寻根网”项目中的五服图自动生成工具、“资治通鉴分析系统”中的人物、职官、地理、事件的聚类工具、“中国文学史研读系统”中的知识自动推送功能。

3.古籍资源数字化的几个层次

A.藏品基础编目信息的数字化(含元数据的处理)。

B.藏品影像数据的采集和处理。

C.藏品全文的数字化。

D.藏品信息的深层次标引。

E.古籍资源数据内部的整理和系联(即知识本体和知识模型的建立)。

F.古籍研究成果的数字化以及与古籍资源的系联。

4.古籍资源数字化具有前瞻性的几个趋势

A.国际上各收藏单位的数字化编目信息有整合的趋势。当前我馆的“中华古籍善本国际联合书目系统”将整合中美 26 家收藏单位、东京大学东洋文化研究所等单位的书目数据就是一个案例。

B.影像信息成为藏品数字化最基础的信息之一。如何在保护知识产权的基础上让影像信息的开放获取更便捷、保存更持久，成为一个重要问题。

C.藏品信息深层次标引成为大势所趋。藏品信息深层次标引及发布是不可扭转的趋势，香港大学冯平山善本书录所提供的例子将成为以后藏品信息数字化的入门要求。古籍普查平台的登录项目同样具有深入标引的功能，期待未来入库的书目数据能够成为最新的最详备的古籍目录标准。在书目数据之上，全文数据的深入标引也必然提上古籍资源数字化的日程。

D.知识模型已被前沿的数字化成果引入。中国历史地理信息系统（CHGIS）等嵌入知识模型的数字化成果代表着当前古籍资源数字化的最新成就，也代表着将来发展的方向。更加全面、详备的古籍资源知识模型首先需要综合既有的各类古籍工具书和检索工具，制定关于全文的深入标引相关标准，以实现统合各类古籍资源文本的功能。

E.普及性的数字化古籍资源成为一个新的发展领域。如何向全世界受众提供普及性的古籍数字化资源是新出现的课题。中国国家图书馆的“华夏记忆”和“世界图书馆”正在进行这方面的探索。

F.适应学界需要的高端数字化项目代表着数字化成果的最终水平。

5.古籍资源数字化发展的不足

A.对于数字化内涵的认识不同和各单位需求与应用的差异。对于图书馆用户而言，文献数字化的服务水平有三个层次，第一层次是数据服务，即将文献的目录、影像或全文按照一定顺序提供给读者；第二层次是知识服务，利用科学的、便于应用的知识管理体系将文献内容中的各类关键词加以统合、关联，把分散于不同媒介的深藏于文献内部知识直接呈现给读者；第三层次是知识服务加方法服务，即在提供知识服务的同时为读者提供快捷、正确的获取某类知识的方法。文献数字化的发展是从第一层次向第三层次演进的过程，不同的单位基于自身发展方向和应用需求的不同，可以确定不同的层次作为主要目标，但是作为国家图书

馆，显然应该以第三层次的服务水平为目标。

B.数字化编目标准不统一。由于古籍编目缺乏统一的国际标准，加之相关工作人员知识背景、学术认知、语言的差异，造成了数字化编目标准不统一的情况。目前，藏品的数字化编目存在如下主要缺陷：①缺少统一标准的编目规则；②缺乏严格的规范控制；③语言转换之间，术语难以划一；④缺少必要的知识管理体系，目录信息如同散沙，难以整合和重组。这些缺陷导致著录信息混乱，相互转化困难，为古籍资源的国际化整合造成了很多障碍。

C.信息提供的不足。与到馆看原书相比，数字资源存在诸多的信息缺失，这主要是古籍信息深层次标引工作没有深入开展造成的。藏品的各种深层次信息如子目、版本类型、刻印类别、批校题跋、钤印、插图、刻工、分类、地理等信息都应该在数字化古籍中得到立体体现。这项任务仍然任重道远。

D.整体来看，古籍资源数字化水平仍然较低。当前古籍资源数字化的成果主要分为书目数据数字化和古籍全文数据库两种，水平都处于初级阶段，有待于项目设计和技术升级的双重扩展。这两种成果的局限如下：①各大收藏单位的古籍书目数据处理一般都停留在卡片格式转化为机读格式的层面，存在很多问题：缺少目录组织，难于按类别查询；缺乏严格的规范控制，不利于书目信息的汇集、区分和引导；缺乏层级划分，各种书目信息混杂于一个层面，不利于准确检索。②现有的古籍全文数据库主要停留在未加整理和标引的全文数据库基础之上，不但通用性差，不方便用户在同一界面检索；而且只能实现字词的简单检索，不能实现高级检索；知识关联和参照功能无法实现。

E.技术应用上存在缺陷和鸿沟。各大收藏单位在实现数字化时采用不同的软件技术支持，这对资源共享很不利。此外，一些先进技术没有很好的应用于古籍资源的数字化，导致数字化的进程受到阻碍。

三、中国古籍资源数字化的发展方向

古籍资源数字化的发展方兴未艾，但是可以肯定的是，古籍资源的数字化是数字时代任何一个古籍收藏单位都不可能回避的历史主潮流。在这场变革的潮流中，谁占领先机，谁才能在这个新兴领域中立于中流砥柱的地位，才能在将来的发展中掌握话语权，才能在世界范围内的图书馆界树立良好的国际声誉。

中国国家图书馆作为世界上最大的中国古籍收藏单位，她应该也必须成为数字化古籍资源的航空母舰、古籍资源数字化标准的制定者以及古籍资源数字化发展方向的引领者。在信息处理方面，我们必须引入计算语言学、人工智能学等方面的最新研究理论与成果，力争突破西方国家对于信息处理技术的垄断，自主研发适合多语言的强大的信息处理工具，更加广泛和深入地统合古籍资源，将图书馆真正建设为信息枢纽。这是关乎国家文化安全和文化战略的头等大事，也是图书馆人必须肩负的责任。

1.中国古籍资源数字化工作的主要问题：

A.古籍资源数字化程度还处于低层次的水平。我们的古籍尚未完全数字化完毕。在已经完成的古籍数字化资源中，低层次的藏品书目信息数据占了绝大多数的份额，影像数据的采集远远没有跟上藏品书目信息的进展。

B.缺乏普遍适应的标准。编目、影像采集、数据整合、项目流程等等各方面都缺乏普遍适应的标准，不利于将来古籍资源数字化工作的开展。

C.藏品深度标引差，数据缺乏深层次处理。要创建高端产品，要树立标杆性典范，你必须对古籍资源进行“深加工”，要在数字化工作中做到“知识挖掘”，要升级古籍数字化资源的知识服务和方法服务。与数字化产品开发的积极劲头相比，已开发产品在后续发展和提升上十分乏力，产品的完善性有可持续发展没有得到应有的重视。

D.技术滞后。到目前为止，我们的数据标引、数据格式转换竟然还处在手工阶段，这在相关技术已经成熟的大环境下是十分不合理的现象。另外，数字化新馆建成后，“碑帖菁华”数据库的功能竟然萎缩，这也是非常不合理的。技术不应该成为限制数字化工作开展的因素。

2.古籍资源数字化建设规划：

2.1 进一步完善藏品的数字化：

A. 已有善本古籍书目信息应当逐步展开深度标引、全文影像采集的工作。

B. 少数民族语文古文献的数字化工作应当展开调研，撰写报告讨论建设民文数字化项目的可行性及相关标准，并加快民文影像数据库的建设。

C. 老照片、年画、宣传画的数字化应当加快步伐。

D. 舆图、金石拓片的影像采集任务应当继续深入。

E. 未数字化的资源应当按部就班地继续开展数字化。

2.2 统一国际适用的古籍资源数字化标准：

数字资源共享的前提是，有可共享的资源 and 可执行的统一标准，这一工作在目前已经开展了，但是其深入程度还远远不够，以下几个方面需要进一步思考：

- A. 制定国际适用的编目标准。
- B. 制定元数据采集、整理标准。
- C. 制定有效可行的影像数据处理标准。
- D. 探讨多语言术语对译的标准化。
- E. 制定项目的建设的标准流程。

2.3 引入新技术，深入挖掘、科学揭示古籍数字化资源：

功能开发是数字资源建设与整合的一项重要工作，功能的优劣直接体现了服务质量的高低。纵观国际上的中华古籍数字资源整合平台，普遍存在对古籍数字资源开发深度不够的问题。要解决这一问题，必须从以下方面重新思考问题：

- A. 如何快速有效地进行海量数据的标引。
- B. 如何在数据转换和处理中引入人工智能技术。
- C. 如何提高检索效率和准确性。
- D. 如何提供智能化的检索服务。
- E. 如何提供智能分析服务，如错误校勘、古籍内容关联等。
- F. 如何提供个性化的信息服务，如偏好分析、内容解析等。