古籍数字化简述

黎知谨

（国家图书馆善本特藏部敦煌吐鲁番资料中心）

一、数字图书馆与古籍数字化

随着信息化时代的到来，作为公共信息和教育服务体系重要组成部分的数字图书馆受到越来越多国家的重视。美国最早开始数字图书馆理论研究和建设。1991年俄亥俄州政府投资建立州内图书馆网络中心，开始了数字图书馆的尝试。1994年6月，在德克萨斯召开了以“数字图书馆理论与实践”为主题的第一次数字图书馆的理论研究会议。同年9月，美国国家科学基金会（NSF）、国家宇航局（NASA）和国防部高级研究署（AKPA）联合发布《数字图书馆启动计划》，领导、组织和资助美国的数字图书馆研究和开发。继美国之后，英、法、德、日等国也先后提出各自的数字图书馆计划。1995年2月25～26日，在比利时布鲁塞尔召开了全球信息社会研讨会，大会确立了全球数字图书馆计划与数字博物馆计划是全球信息社会化的两个重要组成部分。

在中国，数字图书馆建设也已纳入国家的发展战略。1999年6月中国数字图书馆发展战略组、中科院计算所等单位联合主办了“99数字图书馆论坛”。2002年5月中华人民共和国信息产业部、中国数字图书馆等8家单位联合主办“2002年数字图书馆国际论坛”。2002年9月，江泽民同志《在庆祝北京师范大学建校一百周年大会上的讲话》中指出“加快数字图书馆等教育公共服务体系建设”。除了理论研究和思想认识，中国目前已启动了一些工程项目，进入了数字图书馆的建设阶段，例如中国高等教育文献保障体系CALIS、中国国家科学数字图书馆工程及中国国家数字图书馆工程等。

数字图书馆的工程建设已经全面展开，但对数字图书馆的内涵还在讨论之中，尽管有各种说法，但基本认识还是一致的，即把图书馆的各种文献转换成数字信息并通过网络发布和传输，同时采集、加工各种公共信息为全社会提供优质的信息服务和决策咨询，使数字图书馆成为信息社会的核心部分。在我国，数字图书馆建设目前的主要工作还是对现有文献的数字化，其中一项基本内容就是传统古籍的数字化。

传统古籍的数字化有着重要的意义。传统古籍是中国五千年文化积淀的瑰宝，维系中华传统文明的进步与传承，也是我国各图书馆馆藏的重要组成部分。古籍的数字化和上网，是中华优秀文明由纸张等媒质流传方式转为数字信息等现代方式传播的重要步骤，是对传统的中华文化传播和继承方式的革命。古籍的数字化和上网，是互联网上中文信息完整性的重要保障，对确立中华文化在互联网上的整体优势地位和树立文化大国形象具有不可替代的地位。传统古籍的数字化和上网还具有现实的价值，可以有效地解决古籍保存与使用之间的矛盾。古籍通常是1911年以前抄写、出版的图书，包括民国时期出版的古人所著的线装图书，往往具有重要的史料价值和很高的文化价值。许多古籍保存单位严格限制古籍的使用，以达到古籍保护的目的，但同时也使古籍的研究利用受到影响。古籍的数字化和网上发布，使研

究者可以在网络终端上浏览古籍，还能避免直接接触对古籍造成的损坏，能有效地解决古籍保存和使用的矛盾，为中外学者方便地研究古籍提供便利，对古籍研究工作必将产生巨大的推动作用。

二、古籍数字化中的几个问题

传统古籍不同于普通文献，这使古籍数字化的进程面临许多问题。在图书馆业务中，古籍整理工作相对独立，采、编、阅、藏自成体系，古籍整理保留的旧有模式也最多，采用旧的分类法、沿用传统的著录方式，馆际之间也没有国家统一标准，仅分类法就有四库法、中图法、科图法、人大法、刘国钧“十五大类”等等。古籍整理的这种现状制约了古籍数字化的进程。

作为图书馆数字化的重要部分，人们在热烈讨论图书馆数字化的同时也开始研究古籍数字化的理论和技术问题。1992年，李致忠先生在《北京图书馆馆刊》（下简称《北图馆刊》）发表了《略谈建立中国古籍书目数据库》一文，文章针对古籍整理中的各种问题提出建立古籍数据库的前提条件。1995年，李针对古籍数字化面临的问题和解决方案，在《北图馆刊》发表《再论建立中国古籍书目数据库》，讨论了中国古籍书目数据库的建库规范，包括中国古籍分类法，标准著录，古籍书目的机读格式，使用的软件及接口，使用的字库等。尽管李文讨论的古籍书目数据库还远不是古籍本身的数字化，但内容已经涉及到古籍数字化各方面的重要问题，勾勒了古籍数字化整体规范的基本框架。

在李以后，学者们从不同方面思考古籍数字化遇到的问题以及解决方案。1999年，《国家图书馆学刊》（下简称《馆刊》）第2期发表了朱岩的《中国古籍书目数据分析》。朱文从信息处理角度对古籍书目数据做出分析，对《中国古籍善本书目》在机读格式中的使用作出示范。制订统一的机读目录是古籍数据库建库规范之一，机读目录通过对书目数据信息进行标识，完成书目的信息统计、整理和检索。充分利用机读目录提供的字段标识数据信息，能够提高数据库的检索功能，方便读者的检索查阅。《中国古籍善本书目》是由国家古籍整理出版规划小组领导编纂的大型书目，历时十五年完成，分经、史、子、集、丛五类，共九册，收录56000种善本古籍。此书的编写仍然采用传统的古籍编目规则，不利于编制机读目录。论文从检索点的切分入手，具体讨论了书名信息、责任者信息、版本信息、附注文字、分类信息、层次关系等方面信息的标引及其机读目录的实现等问题。

《馆刊》同一期刊登了史睿的《论中国古籍的数字化与人文学术精神》。史文从人文研究角度出发讨论古籍数字化意义及解决方案，强调了在数字化时代传统古籍整理工作的重要性。史文认为古籍数字化能为人文研究提供便捷、准确的查询工具，但要实现这一目标，必须对传统古籍整理工作进行变革，要求建立数据库统一的规范，包括分类法、著录格式都要有一定的修改，以建立国家标准，并使古籍数据库与数字图书馆的其他数据库保持整体的统一性。文章对计算机技术也给予了同样的重视，分析了人文研究对计算机技术的要求，计算机技术为古籍数字化准备的条件，并讨论了两者的有效结合。

《馆刊》1999年第3期《首届“中文古籍开发利用研讨会”纪要》一文，记录了1999年5月12－14日国家图书馆主办的“中文古籍开发利用研讨会”的内容，反映了当时古籍数字化进

程。会议着重对《古籍机读目录格式字段表（试用稿）》作出讨论，探讨了制定国家标准的机读目录格式和统一的古籍分类法的可能性，以及图书馆界在古籍数据库方面所作的尝试及成果。这次会议反映了图书馆界已经开始着手文献资源的网上服务和资源的共建共享。

《馆刊》2002年第2期刊发了鲍国强的《古籍机读目录的文献连接技术及其应用》。鲍曾参与编写《汉语文古籍机读目录格式使用手册》。鲍文结合实际工作，从具体问题入手，讨论古籍机读目录的文献连接技术。文章分析实现文献连接的前提条件、文献连接的类型以及连接技术的应用，针对机读目录文献连接技术的要点，展示应用机读目录中的文献连接技术，以充分发挥古籍书目数据库的文献检索的功能。

由于李、朱、史、鲍都是图书馆工作人员，因而对数据的著录格式以及机读目录都给予了更多的关注，更重视机读目录在标引、检索中的应用，以提高数据库的信息检索能力，提高信息的查全率和查准率，但他们对古籍数字化本身以及带来的相关后果考虑还嫌较少，例如：如何通过计算机与网络技术展示出与古籍原本质地和观感一致的数字化形象，古籍数字化后的人文和学术价值，以及对纸张等实物介质留存的古籍的影响，古籍数字化过程中，工程建设的技术、管理、运营和法律问题，如何在与国外先进技术交流中既达到吸取先进的成果和经验，同时确保古籍的国家信息主权和版权的独立和完整。

三、古籍数字化的主要成果

古籍的数字化是一项庞大的系统工程，除了理论研究，还需要各方面的技术专家特别是古籍研究人员、图书馆工作人员、计算机人员以及其他相关领域的人员通力合作。目前，通过国家基金资助，公司参与以及国际合作等方式，我国的古籍数字化工作已经取得了相当的成绩。目前古籍数字化的工作取得了一定的成绩，已经完成或正在进行的有关古籍数字化的大型项目有：

电子版《四库全书》，由上海人民出版社、香港迪志公司、北京书同文公司合作开发，选用国际标准 ISO/IEC 10646（GB 13000/Unicode）作为数字化的字符集，采用 XML 语言作为文献内容的标识语言，使用 OCR 技术实现图文数码转换，使用数据库加知识工具库多种信息关联的全文检索引擎。书同文公司是大陆最大的致力于古籍数字化的公司，现拥有《四库全书》、《四部丛刊》、《康熙字典》的电子版。此外还有《中华文化通志》、《汉语大词典》、《中华古汉语词典》等产品。目前在制作《永乐大典》和《历代石刻史料汇编》的全文检索版。该公司亦将地方志的数字化列入了规划。

北京大学中文系《全唐诗》网上电子检索系统，由211工程资助、北京大学中文系李铎博士主持开发，历时一年完成。该项目主体部分由《全唐诗》及《全唐诗补编》组成，辅助项由《乐府诗集》、《玉台新咏》、《文选》等组成。参考类则由重要唐代史料《新唐书》、《旧唐书》、《唐才子传》、《历代诗话》、《唐诗纪事》等资料组成，共1700万字。全部文献错误率控制在三万分之一以下（共五校），《全唐诗》文本控制在五万分之一以下（共六校）。所有文献均使用Unicode内码，在Windows2000平台上，不需要任何转码工具，全球任何语言的操作系统均可在网上直接检索《全唐诗》及相关资料，并且兼容Windows9x，WindowsNT，Unix，Linux等平台。检索系统由两个版面组成，一是浏览界面，它提供以原书为序浏览，

浏览内容只限于《全唐诗》。另一界面是检索界面，此界面是本系统的核心，可以检索全部资料。主体部分除全文检索功能外，另有诗题检索、作者检索、体裁检索、音韵检索等功能，检索结果显示诗歌全文（以首为单位）、作者小传、诗文校注、诗歌体裁、原书页码、册、卷等。

“中国基本古籍库”光盘工程，由北京大学刘俊文先生主持，是北京大学的重点项目，1998年经全国高校古委会的批准立项，正式启动。著名学者季羡林、国家图书馆馆长任继愈担任编纂委员会主任，两院院士罗哲霖、工程院院士李国杰担任技术委员会主任，由北大方正技术研究院提供技术支持。全套光盘库共500张，根据中国古籍自身的特点，参照国际通行的图书分类法分为哲科、史地、艺文、综合4个子库，20个大类，近百个细目。范围涉及先秦至民国的重要典籍1万余种，每种典籍有1个通行版本的全文信息，另附1－2个珍贵版本的图像数据，预计全文20亿字，版本图像2千万页。基本可以满足文史和其他方面研究者的研究需求。适用于中、英、日、韩多语种操作平台，并提供多重检索功能。用户只需懂得一些基本的电脑操作方法，就可在极短的时间内，查找所需的资料，每次检索均可在5秒内完成。

台湾中央研究院《汉籍电子文献》，始于1984年7月，前身是为开发二十五史全文数据库而成立的“史籍自动化计划”，现已完成的数据库，共约一亿两千万字，其中较大型的是二十五史、医药文献、明实录、历代史料笔记丛刊和十三经，这些数据库已包括中国唐代以前的大部份重要文献（道教资料除外）；正在建设中的数据库多达一亿八千万字，准备逐步将宋代以下的重要文献数字化。所有文献通过人工与机器共进行3次校对。在制作技术上得到中央研究院计算中心的支持。使用者可以在一秒之内，查到二十五史数据库中四千万字的任何字词。

“国际敦煌学项目”（The International Dunhuang Project，简称 IDP），旨在通过国际合作促进敦煌写卷的研究与保护。由英国图书馆开发，开始于 1993 年。项目计划逐步将全世界各国各收藏单位的敦煌文献数字化。目前可在线查看英国图书馆收藏的 3 万余件中亚写本和印本文件，以及 15000 余件残卷的高质量彩色图片。2001 年 3 月，中国国家图书馆与英国国家图书馆签署五年合作项目，加入此项目。中国国家图书馆国际敦煌学项目的数字化内容主要包括：1、馆藏敦煌文献数字化。使用扫描图像展示写卷的全部内容——正面、背面，甚至没有文字的地方，图像的清晰度与看原卷没有区别。同时使用国际敦煌学项目提供的专门设计的 4D 数据库详细描述写卷的物理性质。2、研究论著目录数据。包含四个专题书目数据库：敦煌吐鲁番学日文论著目录数据库；敦煌吐鲁番学西文论著目录数据库；敦煌文献研究论著目录数据库和敦煌吐鲁番学学者档案数据库。3、中国国内散藏敦煌文献联合目录。

以上介绍的是目前古籍数字化的重要工程项目，随着数字图书馆建设的进行，国家资金的投入和各方面专家的努力，古籍数字化中的各种各样的问题必将得到妥善的解决,我国五千年的优秀文化必将得到更好的传承。