

论古籍数字化标准体系的构建

王 昭

迄今为止，古籍数字化已走过了 30 余年历程，经过了书目数据库与全文数据库两个阶段；与书目数据库相比，全文数据库可以说是一个质的飞跃，实现一站式获得；古籍全文数据库的功能可以概括为查找出处，瞬息即得，关键词检索，同类相聚；古籍数字化为学术研究开辟了一条新途径，改变了学者查阅图书的方式，对学术研究所产生的影响与意义是划时代的¹。在古籍数字化的研究与实践中，标准化是不可回避的问题，也是亟待解决的问题。

古籍数字化标准化问题的研究由来已久，观点却各有不同，如王桂平在《我国古籍数字化的现状及展望》中提出，实现古籍数字化的标准化和规范化，可由中国图书馆学会古籍整理专业委员会牵头，以国家图书馆、上海图书馆、南京图书馆为龙头，在深入研究的基础上，确立古籍数字化的统一著录格式和标引方法²；陈立新在《古籍数字化的进展与问题》中提出，古籍数字化应建立统一的古籍机读目录、支持古籍数字化的汉字平台、古籍文献规范文档、古籍影像处理标准等³；陈阳在《中文古籍数字化的成果与存在问题》中提出，如果相关部门能够联合从事古籍数字化的主要单位，制订古籍数字化的统一标准，实现标准化、规范化，在此基础上进行分工合作，建立可共享的资源体系，必定能够促进国内数字化古籍的开发与利用⁴；李筑宁在《关于古籍资源数字化建设中几个问题的探讨》中提出，在古籍数据库建设过程中，各参加馆要坚持执行统一的标准规范，如统一的用户界面、数据格式、数据库建设规则、信息交换协议、馆际互借协议等；数据加工技术采用国际标准、国内标准；服务系统软硬件配置要采用国际通用的开放式操作系统平台技术，网络通讯协议 TCP/IP 技术，面向广域网的数据库技术等⁵；高娟等在《中国大陆地区古籍数字化问题及对策》中提出，在古籍数字化的标准建设上，目前最为迫切的是对国内外已有古籍数字化相关的国际标准、国家标准、行业标准等进行全面分析评价，在此基础上依据我国古籍数字化工程的目标

¹ 陈志伟，盖阔：《中文古籍全文数据库指要》，《图书馆学研究》2014 年第 14 期，第 39-43 页。

² 王桂平：《我国古籍数字化的现状及展望》，《图书情报知识》2000 年第 4 期，第 50-51 页。

³ 陈立新：《古籍数字化的进展与问题》，《上海高校图书情报工作研究》，第 36-38 页。

⁴ 陈阳：《中文古籍数字化的成果与存在问题》，《出版科学》2003 年第 4 期，第 47-48 页。

⁵ 李筑宁：《关于古籍资源数字化建设中几个问题的探讨》，《图书情报工作》2010 年 S1 期，第 312-315 页。

与任务,建立可为各数字化主体认可的标准规范体系框架以及可以共同遵守的统一标准⁶。

从上述观点中不难看出,古籍数字化标准化的重要性毋庸置疑,而如何建立古籍数字化标准、建立什么样的标准等问题,仍有待进一步的研究。究其原因,主要在于“古籍数字化”的概念在不断发展变化中,其外延与内涵仍不明确。

1. 古籍数字化

在20世纪80年代前后,我国台湾地区以及大陆地区陆续开始了将古籍整理与计算机相结合的实践活动⁷。1990年,台湾《国文天地》第九期首次使用了“古籍电脑化”,但是并未对这一术语进行定义。随后出现了“史籍自动化”、“古籍自动化”、“古籍文献资讯化”、“珍藏文献数位化”、“古籍电子化”、“古籍网络化”、“古籍文献数据化”等一系列术语,目前学界和业界普遍使用的术语为“古籍数字化”(台湾地区称为“古籍数位化”),这一术语是1997年刘炜在《上海图书馆古籍数字化的初步尝试》⁸中首次提出的,但文中并未对这一术语进行定义。

“古籍数字化”的概念在不断明确、发展与深化,比较有代表性的说法包括:陈洪澜在《中国古籍电子化发展趋势及其问题》中提出,古籍电子化是利用现代科技手段研究古代社会文明的基础,代表着中国古籍整理与研究的发展趋势⁹;李运富在《谈古籍电子版的保真原则和整理原则》中提出,古籍电子化是指利用现代信息技术,将历来以抄写本、刻铸本、雕版、活字版、套版及铅字印刷等方式所呈现的古代文献,转化为电子媒体的形式¹⁰;彭江岸在《论古籍的数字化》中提出,古籍数字化就是利用数字技术将古籍的有关信息转换成数字信息,存贮在计算机上,从而达到使用和保护古籍的目的¹¹;王桂平在《我国古籍数字化现状及展望》中提出,古籍数字化就是采用计算机技术,对古籍文献进行加工、处理,制成古籍文献书目数据库和古籍全文数据库,用以揭示古籍文献中所蕴涵的极其丰富的信息资源,为古籍的深度开发打下良好的基础¹²;刘琳等在《古籍整理学》中提出,古籍数字化就是将古代典籍中以文字符号记录的信息输入计算机,从而实现了整理、存储、传输、检索等手段的计算机化¹³;陈阳在《中文古籍数字化的成果与存在问题》中提出,古籍数字化是利

⁶ 高娟,刘家真:《中国大陆地区古籍数字化问题及对策》,《中国图书馆学报》2013年第4期,第110-119页。

⁷ 王立清:《中文古籍数字化研究》,北京:国家图书馆出版社,2011,第18页。

⁸ 刘炜:《上海图书馆古籍数字化的初步尝试》,《图书馆杂志》1997年第4期,第33-34页。

⁹ 陈洪澜:《中国古籍电子化发展趋势及其问题》,《中国典籍与文化》1998年第4期,第121-126页。

¹⁰ 李运富:《谈古籍电子版的保真原则和整理原则》,《古籍整理研究学刊》2000年第1期,第1-7页。

¹¹ 彭江岸:《论古籍的数字化》,《河南图书馆学刊》2000年第2期,第63-65页。

¹² 王桂平:《我国古籍数字化的现状及展望》,《图书情报知识》,2000年第4期,第50-51页。

¹³ 刘琳,吴洪泽:《古籍整理学》,成都:四川大学出版社,2003年,第335页。

用现代信息技术将古代文献转化为电子媒体的形式,通过光盘、网络等介质保存和传播¹⁴;李明杰在《中文古籍数字化基本理论问题自议》中提出,古籍数字化是以保存和普及传统文化为基本目的,以知识发现的功能服务学术研究为最高目标,在对传统纸质古籍进行校勘整理的基础上,利用计算机技术将其转换成可读、可检索、及实现了语义关联和知识重组的数字化信息的过程¹⁵;毛建军在《古籍数字化的概念与内涵》中提出,古籍数字化就是从利用和保护古籍的目的出发,采用计算机技术,将常见的语言文字或图形符号转化为能被计算机识别的数字符号,从而制成古籍文献书目数据库和古籍全文数据库,用以揭示古籍文献信息资源的一项系统工作¹⁶;王立清在《中文古籍数字化研究》中提出,古籍数字化就是指利用现代信息技术将传统古籍整理后转化为数字媒体形式的书目数据库和全文数据库,通过光盘、网络等介质保存和传播,以达到保存普及传统文化和服务学术研究之目的¹⁷;葛怀东在《“古籍数字化”课程的建设与实践》中提出,古籍数字化是对古籍或古籍内容的加工与再现,是古籍整理工作在新时期的延伸¹⁸。

上述定义基本都是从古籍的角度来定义古籍数字化,即古籍数字化是古籍的介质转换、古籍的资源开发、古籍整理的延伸等。若以数字化为基础来定义古籍数字化,首先应明确数字化的概念。数字化是把声音、图像、文字、图形等信息全部变为计算机能够识别的二进制数字序列¹⁹。古籍数字化可以认为是对象为古籍的数字化。狭义的古籍数字化是古籍资源转换为计算机可读数据的过程。而广义的古籍数字化是古籍数字资源生产、组织、管理、整合、存储、使用、处置等全过程。

2. 标准化体系构建

古籍数字化标准体系是古籍数字化标准按其内在联系形成的有机整体。贺科伟在《我国古籍数字化标准体系建设刍议》中提出,我国古籍数字化标准体系包括古籍版本标准、古籍分类标准、字符集标准、储存格式标准、古籍影像处理标准、检索标准、元数据标准等²⁰。古籍数字化标准体系应是以技术标准体系为主,包括为实现技术标准的要求而建立的管理标准体系、工作标准体系,以及为有效运行全过程标准化的各项管理要素在内的标准体系,且都应满足目的性、集成性、层次性、动态性和阶段性的要求;技术标准是开展数字化古籍作

¹⁴ 陈阳:《中文古籍数字化的成果与存在问题》,《出版科学》2003年第4期,第47-48页。

¹⁵ 李明杰:《中文古籍数字化基本理论问题自议》,《图书馆论坛》2005年第5期,第97-100页。

¹⁶ 毛建军:《古籍数字化的概念与内涵》,《图书馆理论与实践》2007年第4期,第82-84页。

¹⁷ 王立清:《中文古籍数字化研究》,北京:国家图书馆出版社,2011年,第20-21页。

¹⁸ 葛怀东:《“古籍数字化”课程的建设与实践》,《兰州教育学院学报》2013年第2期,第106-107页。

¹⁹ 沈孟璿:《新中国60年新词新语词典》,成都:四川辞书出版社,2009年,第400页。

²⁰ 贺科伟:《我国古籍数字化标准体系建设刍议》,《科技与出版》2011年第8期,第76-79页。

业的技术条件,包括作业对象、作业条件、作业方式等所作的规定,包括古籍(版本)择取标准、加工标准、设备标准、元数据标准、数据库标准、检索标准、软件标准等;管理标准是对古籍数字化业务流程中各个环节所作的规定,包括项目选题与评估、分类体系、设备管理、作业流程管理、业务外包管理以及标准化管理等;工作标准是对负责项目开发的机构、相关人员的职责、工作要求、考核办法所作的规定,包括职责权利、工作程序、办事细则、考核标准和相互关系准则等²¹。古籍数字化标准体系框架包括技术标准、管理标准和工作标准;技术标准是古籍数字化标准体系的核心,包含古籍版本开放标准、数据加工标准、文档存储标准、著录标准、元数据标准、数据库标准、检索标准、设备标准、软件标准等九类;管理标准是古籍数字化标准化工作的指导,包含选题与评估标准、项目开发管理标准、典藏古籍管理标准、古籍资源分类标准和标准化管理标准五类;工作标准是对古籍数字化具体工作内容的规范,包含作业流程标准、设备管理标准、质量测评标准和业务外包标准等四类²²。

上述观点从不同角度出发,构建古籍数字化标准体系,具有一定的借鉴意义。但是要从古籍数字化的实践出发,建立完备的古籍数字化标准体系,首先要充分认识古籍数字化领域的特殊性。古籍数字化是以用户需求为驱动力,包含文献、工程、学术、技术、成本等诸多因素,以及各种因素间的相互限制与相互作用。古籍数字化是基于文献、面向用户需求的工程,考虑学术含量的同时,要受成本和技术的双重限制。因此,构建古籍数字化标准体系,还要从古籍数字化的基本概念出发,体现其特殊性与复杂性。

基于广义古籍数字化概念,古籍数字化标准体系可划分为引用标准集、专门标准集和参考标准集,如图1所示。

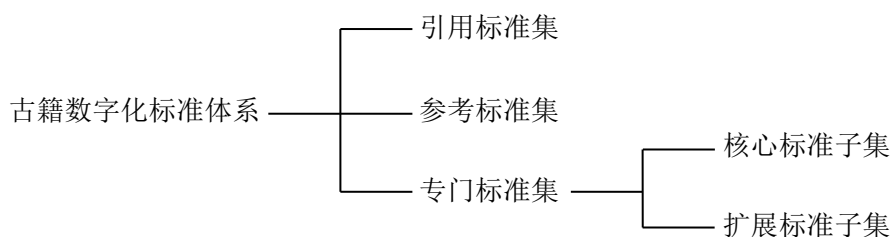


图1 古籍数字化标准体系示意图

引用标准集是古籍数字化标准体系的基础,包含大量其他领域可直接应用于古籍数字化

²¹ 葛怀东:《论古籍数字化标准体系建设》,《图书馆学刊》2013年第1期,第47-49页。

²² 张文亮,尚奋宇:《我国古籍数字化标准体系现状调查及优化策略》,《国家图书馆学刊》2015年第6期,第83-89页。

的标准,例如汉语信息处理词汇 01:部分基本术语(GB/T 12200.1-1990)、汉语信息处理词汇 02 部分汉语和汉字(GB/T 12200.2-1994)、信息技术 中文编码字符集(GB 18030-2005)、中国机读规范格式(WH/T 15-2002)、信息与文献 都柏林核心元数据元素集(GB/T 25100-2010)、图书馆馆藏资源数字化加工规范第 3 部分:图像资源(GB/T 31219.3-2014)、信息技术连续色调静态图像的数字压缩及编码第 1 部分要求和指南(GB/T 17235.1-1998)、信息技术 JPEG 2000 图像编码系统第 1 部分核心编码系统(GB/T 30248.1-2013)、电子成像文件图像压缩方法选择指南(GB/Z 19736-2005)、信息技术 可扩展置标语言(XML) 1.0(GB/T 18793-2002)、文献管理 长期保存的电子文档文件格式 第 1 部分:PDF1.4(PDF/A-1)的使用(GB/T 23286.1-2009)、数字对象唯一标识符规范(WH/T 48-2012)、数字资源长期保存元数据规范(WH/Z1-2012)、管理元数据规范(WH/T 52-2012)等。

参考标准集是古籍数字化标准体系的重要补充,包含大量其他领域可间接或部分应用于古籍数字化的标准,例如汉文古籍特藏藏品定级第 1 部分:古籍(GB/T 31076.1-2014)、古籍修复技术规范与质量要求(GB/T 21712-2008)、缩微摄影技术 在 16mm 卷片上拍摄古籍的规定(GB/T 7517-2004)、缩微摄影技术 在 35mm 卷片上拍摄古籍的规定(GB/T 7518-2005)、索引编制规则(总则)(GB/T 22466-2008)、数码照相机术语(GB/T 20733-2006)、数码照相机曝光指数、ISO 感光度值、标准输出灵敏度和推荐曝光指数的确定(GB/T 20224-2006)、照相镜头第 1 部分:变焦距镜头(GB/T 9917.1-2002)、照相镜头第 2 部分:定焦距镜头(GB/T 9917.2-2008)、平板式扫描仪通用规范(GB/T 18788-2008)、图解技术 输入扫描仪校准用彩色反射指标(ANSI IT8.7/2-1993)等。

专门标准集是古籍数字化标准体系的核心,包含大量专门为古籍数字化制定的标准。专门标准集又可以划分为核心标准子集和扩展标准子集。核心标准子集基于狭义古籍数字化概念,包括数据标准和加工流程标准两类。数据标准可以划分为元数据标准、对象数据标准、知识数据标准和数据加工所需数据标准四个子类。加工流程标准既可以按照数据类型分类,也可以按照加工工序分类。目前,我国大陆地区已发布的古籍数字化专门标准如下:

国家标准:古籍著录规则(GB/T 3792.7-2008)。

行业标准:古籍元数据规范(WH/T 66-2014)。

机构(或项目)标准:国家数字图书馆标准规范,包含古籍元数据标准与著录规范、家谱元数据标准与著录规范、拓片元数据标准与著录规范、舆图元数据标准与著录规范、甲骨

元数据标准与著录规范、中文文献全文版式还原与全文输入 XML 规范²³、古籍用字规范、生僻字避讳字处理规范等；我国数字图书馆标准规范建设项目（2002DEA20018）成果，包含古籍元数据规范、家谱元数据规范、拓片元数据规范、舆图元数据规范、地方志元数据规范等；大学数字图书馆国际合作计划（China Academic Digital Associative Library, CADAL）项目标准，包含古籍元数据著录规范、古籍数字对象制作规范²⁴等。

上述标准都属于数据标准，且大多数为元数据标准，只有“中文文献全文版式还原与全文输入 XML 规范”和“古籍数字对象制作规范”为对象数据标准，“古籍用字规范”和“生僻字避讳字处理规范”为数据加工所需数据标准。

扩展标准子集基于广义古籍数字化概念，包括古籍数字资源生产标准（不包含数据标准和加工流程标准）、古籍数字资源组织标准、古籍数字资源管理标准、古籍数字资源整合标准、古籍数字资源存储标准、古籍数字资源使用标准、古籍数字资源处置标准等多个子类。

此外，还有一类标准在范围中明确指出不适用于古籍，应排除在古籍数字化标准体系之外，例如图书馆馆藏资源数字化加工规范第 2 部分：文本资源（GB/T 31219.2-2014）等。

结论

古籍数字化标准体系构建应以专门标准集为立足点，优先研制一系列古籍数字化核心标准，同时依据实际需要引用已有的标准，并引入相关的标准作为参考。

国家科技支撑计划项目“中国地方志数字化关键技术与演示平台设计”（2015BAK07B00）依据旧志（1949 年以前编撰或出版的地方志）文献特性，建立 5 类 8 个地方志数字化加工规范，包括：文字处理规范 2 个，汉字集外字描述规范和文字认同描述规范；元数据规范 2 个，地方志专门元数据规范和地方志卷目数据标引规范；对象数据规范 2 个，地方志图像数据规范和地方志文本数据规范；语料数据规范 1 个，地方志语料数据规范；知识库数据规范 1 个，中国古今地名数据描述规范。旧志具备古籍一般的文献特性，基于旧志的数字化标准化成果全部或大部分适用于古籍。

地方志元数据规范用于整体描述地方志图像数据和文本数据，内容包括书目信息、数字化参数、版权信息、用户使用提示等；地方志卷目数据标引规范用于描述地方志图像数据的卷次结构，内容包括卷目信息、层次结构、分类、起始页码等；地方志图像数据规范用于描述地方志图像及规范图像采集方法，内容包括图像类型、画幅内必备要素、图像参数、数字

²³ 按：该标准名为“中文文献全文版式还原与全文输入 XML 规范”，依据古籍全文文本化的需求编制，虽然适用范围中包含古籍和普通中文文献，但是在实际应用中，只适用于古籍，因此列为古籍数字化标准。

²⁴ 按：该标准规定了古籍数字对象制作过程中的原则、采集要素、加工标准、存储格式、目录结构、特例处理等。

化设备参数、图像压缩、色彩管理等；汉字集外字描述规范用于描述地方志文本数据中的集外字，内容包括集外字定位、集外字图像（或图形）、IDS 描述等；文字认同描述规范用于描述地方志文本数据中的文字认同情况，内容包括认同字定位、认同方式、认同类型等；地方志文本数据规范描述地方志文本数据的基本结构和全文文本描述方式，内容包括文本数据格式、文字描述、版式描述、图形图像描述等；地方志语料数据规范描述地方志语料数据的基本结构和标引方式，内容包括文本、文字描述、版式描述、图形图像描述等；中国古今地名数据描述规范规定了中国古今地名数据的内容、构成要素及各要素的描述规则，适用于中国古今地名数据库的建立及格式交换。上述规范作为汉王科技股份有限公司的企业标准，在企业标准信息公共服务平台 <http://www.cpbz.gov.cn/> 发布。

目前，全国图书馆标准化技术委员会已完成了汉语文古籍索引数据标准化研究（WH/Y01-2015）、汉语文古籍文本数据标准化研究（WH/Y02-2015）、汉语文古籍语料数据标准化研究（WH/Y03-2015）等标准研究项目，正在进行汉语文古籍文字认同描述规范（WH2016-01）、中国古今地名数据描述规范（WH2016-05）、汉文古籍集外字描述规范（WH2017-03）、汉文古籍文本数据规范（WH2017-04）等标准研制项目。