

数字环境下古籍整理范式的传承与拓新

李明杰

摘 要 中国的古籍整理以其悠久的历史形成了一种相对稳定和固化的模式,其标志就是西汉成帝时刘向等人开创的文献整理的一般程序和方法。这套程序和方法因其广泛的适用性和深远的学术影响,成为中国古代典籍整理的基本范式。近代以来学科的发展和分化,使得两千年来稳定不变的古籍整理范式细化出四种适应于不同学科需要的类型:古籍文本的复原性整理、古籍语义的解释性整理、古籍内容的组织性整理和古籍实体的保存性整理。随着现代信息技术在古籍整理中的深度应用,以上四种范式也从传统的纸本古籍整理延伸到了古籍的数字化整理之中,并衍生出一种新的范式类型——古籍知识的数据化整理,即不仅把数字化古籍看作古籍整理的结果,更是把它作为下一步古籍整理的数据和素材。新范式将对数字古籍整理和人文社会科学研究方法带来深刻的影响。图1。参考文献37。

关键词 古籍整理 古籍数字化 古籍数据化 学科范式

分类号 G256

The Inheritance and Innovation of Ancient Book Collation Paradigm in the Digital Environment

LI Mingjie

ABSTRACT

Collating ancient books has a long history in China and has formed a kind of relatively stable pattern. In Han Dynasty, a famous philologist named Liu Xiang created a general procedure and a set of methods to collate Chinese ancient books, which included collecting various editions of the same book, choosing an appropriate title for the book, reordering papers, correcting words, making a fair copy of the final version, cataloging and classifying books. This procedure and the set of methods had developed into the basic paradigm of collating ancient books because of their wide range of applications and profound academic influence in ancient China.

However, in the process of “Eastward spread of Western learning and cultural awareness” in modern times, the basic structure of Chinese traditional learning which consisted of Confucian classics, history, philosophy and belles-letters was broken. At the same time, some studies which were formerly ignored got more attention and many new subjects appeared. As a result, a split generated correspondingly in Liu Xiang’s generally applicable paradigm of collating ancient books. From the perspective of Historiography, Literature and Library Science, this paper selected some representative monographs on collating ancient books in these three disciplinary fields, introduced and compared their different definitions about the con-

通信作者:李明杰, Email:lmjiewd@163.com, ORCID:0000-0002-1876-9040 (Correspondence should be addressed to LI Mingjie, Email:lmjiewd@163.com, ORCID:0000-0002-1876-9040)

cept of “ancient book collation”. On the base of that, the author argues that Liu Xiang’s paradigm can be subdivided into four branches in the present paper environment such as the paradigm of restoring the text of ancient books, the paradigm of annotating ancient authors’ original ideas, the paradigm of cataloging and compiling ancient books and the paradigm of preserving the physical entity of ancient books. The paper analyzes the different principles for collating ancient books according to these four paradigms and points out their respective function of paradigms in the traditional process of collating ancient books.

Since the 1970s, with the wide application of modern information technologies based on computer in the collation of ancient books, these four paradigms mentioned above have extended themselves into the collation of digital ancient books from that of traditional paper ancient books and a new paradigm of “dataloging” the knowledge in ancient books has been created gradually, which means the digital ancient books are not only the result of the collation but also the data and source materials for the next collation. Through fine-grained data processing of numerous ancient books, the modes to acquire, annotate, compare, interpret and represent the knowledge in ancient books will fundamentally change. This is the biggest innovation of ancient book collation paradigm under the digital environment. Based on a large number of case studies, the paper probes into the development and innovation route of ancient book collation paradigm under the digital environment and argues that the new paradigm of “dataloging” the knowledge will give a revolutionary influence on the collation of digital ancient books and the research of humane-social studies in the future. The value of this study is to reveal the development law of ancient book collation paradigm from the complicated practice of collating ancient books. 1 fig, 37 refs.

KEY WORDS

Collation of ancient books. Digitization of ancient books. Datamation of ancient books. Discipline paradigm.

传统的古籍整理既是中国古代典籍保护和文化遗产的一种手段,也是学术研究的一项基本技能,在其漫长的历史进程中,早已形成了一种相对稳定和固化的模式。用托马斯·库恩(Thomas S. Kuhn)的话讲:“对某一时期某一专业做仔细的历史研究,就能发现一组反复出现而类似标准式的实例,体现各种理论在其概念的、观察的和仪器的应用中。这些实例就是共同体的范式。”^[1]所谓范式(paradigm),指的是某一学科研究者群体共同遵从的世界观和行为方式,包括群体成员均认可的问题、目标、方法、手段、过程 and 标准等。就古籍整理而言,如果从孔子的七世祖正考父校勘《商颂》算起,迄今已有 2 800 多年的历史,期间无数学者孜孜不倦地潜心古籍整理,都是为了实现“经天地,纬阴阳,正纲纪,弘道德”^[2]的社会理想,这就是他们共

同的世界观。汉成帝河平三年(公元前 26 年),谒者陈农求遗书于天下,由光禄大夫刘向领衔的一群优秀学者开始对朝廷所藏先秦以来的典籍进行系统整理,更是开创了古籍整理的一整套程序和方法,包括广罗异本、确定书名、审定篇章、校勘文字、缮写定本、撰写叙录、类分图书等。这套融合了版本、校勘、目录等诸多传统学问在内的程序和方法,对后世产生了极其深远的影响,其中绝大部分被沿用了下来。甚至可以说,汉魏至明清的历代学者,无论他们在古籍整理领域曾取得过怎样的成绩,其基本程序和方法都始终没有突破刘氏的藩篱。

然而近代以来,随着学科的不断发展和分化,两千年来这种稳定不变的状态发生了变化。受西学东渐的影响,中国传统学术“经、史、子、集”的基本格局被打破,而许多在古代中国不受

重视的学科获得了新生,之前没有的“新学”也被引进了中国,其结果就是研究和利用古籍的视角越来越多元化。从不同的学科视角出发,人们对古籍的价值属性的认识和需求是不同的,古籍整理的对象、手段和基本要求也有所差别,这就导致原来普遍适用的刘氏古籍整理范式出现了相应的分化。20世纪70年代以后,以计算机为核心的现代信息技术逐步被应用到古籍整理领域。从最初的文本录入、索引编制和语词统计,发展到今天的数据库、互联网、超文本、知识挖掘、可视化等各项信息的综合应用,人们的古籍整理理念更是发生了超越性的变革,并且在整理路径和方法上有了更多的选择。按照库恩的理论,科学研究方法及观念的更新,意味着旧的科学范式有了突破,而这势必导致科学革命的发生。刘氏古籍整理范式在不同的学科背景下是如何分化的?数字环境下古籍整理范式又是如何传承和拓展的?本文即着重探讨传统古籍整理范式在数字环境下的传承和变革路径,并分析其可能给古籍整理带来的现实影响。

1 传统古籍整理范式的分界与融合

对于古籍整理,前代学者似乎更偏重于实践,而疏于理论归纳。虽然也有像王引之《经义述闻》、俞樾《古书疑义举例》、陈垣《校勘学例》那样对古籍整理规律进行总结的论著,但多偏重训诂和校勘,仅局限于古籍整理的某个方面。甚至对于“什么是古籍整理”这样一个基本问题,时至今日学界仍存在不同认识。从目前的学科背景来看,与古籍整理关系最为密切的是传统的史学、文学和新兴的图书馆学。笔者分别选取了这三个学科领域内古籍整理的代表性著作,将其中关于“古籍整理”概念的表述比照如下。

首先看史学家对古籍整理的理解。已故著名史学家黄永年教授称:“古籍整理,是对原有的古籍作种种加工,而这些加工的目的是使古

籍更便于今人以及后人阅读利用,这就是古籍整理的涵义,或者可以说是古籍整理的领域。超越这个领域,如撰写讲述某种古籍的论文,以及撰写对于某种古籍的研究专著,尽管学术价值很高,也不算古籍整理而只能算古籍研究。”^[3]黄先生所举的古籍整理程序和方法,包括选择底本、影印、校勘、标点、注释、今译、索引、序跋、附录等。四川大学古籍整理研究所刘琳教授认为:“严格意义上的古籍整理,就是对古籍的原文进行某种形式的整理加工,以便于人们阅读与研究。比如校勘以是正文字,标点以分清句读,注释以阐明文义,翻译以通达古今,辑佚以摭拾遗文,抄纂以采其菁华等等。”^[4]

再来看文学界是如何界定古籍整理的。河北大学文学院时永乐教授称:“所谓古籍整理,就是对古籍本身进行校勘、标点、注释及今译等各种加工,使之出现新的本子,以便于今人和后人阅读利用。”^[5]苏州大学文学院曹林娣教授认为:“整理古籍的目的就是要为研究者们提供一本最可靠的本子,尽量恢复古籍的本来面貌。古籍整理的具体手段,主要有校勘、标点、注释、今译、辑佚、索引、序跋和附录等,还有影印珍本善本图书也属于古籍整理的内容。”^[6]通过对比不难发现,史学界与文学界对古籍整理的界定大同小异,都强调古籍整理的目的是为了便于人们阅读和研究古籍,在整理方法上都是沿用传统的文献学方法,以恢复古籍文本原貌或帮助读者理解文本内容,体现了文史不分家的特点。

但在图书馆学界看来,古籍整理的涵义却别有不同。1982年北京大学和武汉大学合编的《图书馆古籍整理》称:“图书馆古籍整理工作,是对图书馆所收藏的古籍进行著录、鉴定版本、分类、典藏,向读者提供所需要的古代文献资料……一方面正确地揭示、反映、宣传图书馆中所藏的古籍,使读者迅速、准确地检索,以得到所需要的资料……另一方面,做好古籍藏书保护工作,使祖国宝贵的文化遗产得以安全、完整的保藏,使其不受损坏,而有利于长期使用。”^[7]廖

延唐、曹之编著的《图书馆古籍整理》亦是将古籍著录、古籍版本、古籍分类作为古籍整理的核心内容。王世伟也说:“首先要区分一下图书馆古籍工作与一般意义上的古籍整理的不同,一般意义上的古籍整理所指的古籍范围比图书馆古籍工作所指的古籍范围要大得多,前者包括标点、注释、今译、校勘、辨伪、辑佚等,而后者一般不涉及以上内容。”^[8]可见,图书馆学界理解的古籍整理通常包括两个方面:一是对古籍进行分类编目,使之易于被读者检索利用;二是对古籍进行典藏,使之得以长久保存。

当然,图书馆学界与文史学界在古籍整理的路径上并不是泾渭分明的。文史学界校勘古籍时首先要广罗同书异本,这就要用到古籍目录和版本鉴定方法,而古籍编目与版本鉴定通常被认为是图书馆学的方法。这种融合的趋势在某些身栖多个学科的学者身上得到了体现。比如,著名史学家、图书馆学家、文献学家来新夏先生认为,古籍整理必须具备八种技能:分类第一、目录第二、版本第三、句读第四、工具第五、校勘第六、考据第七、传注第八^[9]。这里的句读、校勘、考据和传注是传统的文史学方法,而分类、目录、版本和工具书,则主要是图书馆学的技能,两者恰好各占古籍整理的“半壁江山”。

综上所述,虽缺乏理论的提炼,但传统古籍整理范式是实际存在的。它存在于人们对古籍整理这一学术活动所具人文社会价值的共性认识的基础之上,存在于历代学者所遵从的彼此相近的古籍整理理念、程序和方法之中。刘向等人开创的文献整理程序和方法是中国古籍整理最早的范式,但这个范式是总体性的,随着近代学科的发展和分化,逐渐演变成若干基于“问题”和“目的”的小范式,笔者将之归纳为以下四种类型。

范式 T(Text):古籍文本的复原性整理。它以复原古籍文本原貌为目的,表现为:运用版本学的方法,通过版本鉴定和版本源流的考订发现善本,解决文本整体的可靠性问题;通过辨伪方法排除伪书,解决文本的真实性问题;运用校

勘方法,校正古籍的篇章次第及文字错误,解决文本的准确性问题;运用辑佚方法,补齐古籍残缺的内容,解决文本的完整性问题。通过以上的系统整理,获得一个尽可能接近古籍原貌的文本。在此范式下,古籍整理遵从的是孔子提出的“述而不作,信而好古”的基本原则。

范式 A(Annotation):古籍语义的解释性整理。它以解释作者所要表达的真实语义为目的,表现为:运用传统的注释方法(包括传、注、解、集传、集注、集解、训诂、笺注、疏义、章句、音义、直解等),辅之以准确的断句和标点,再用现代通行的语言对局部内容进行解释,或对全文进行通释,以帮助读者准确理解文意,而不至于因为语言文字、地理名称、行政区划、典章制度的发展变迁而曲解作者原意。它要求遵从两条基本原则:一是符合语言文字自身发展的规律,二是符合作者所处的社会历史背景。

范式 C(Catalog & Compilation):古籍内容的组织性整理。它以检索和利用古籍的内容为目的,亦可称之为古籍的整序,包括古籍编目和古籍编纂等。《中国古籍总目》著录现存古籍二十万余种,面对数量如此巨大的古籍,读者在查找和利用时往往无所适从。古籍编目就是将为数众多的古籍的外部特征和内容要点著录成款目,并按需要组织成分类、书名、著者或专题书目,使之成为一种便于检索的工具。为了集中某类古籍或某类内容,还可对古籍的内容结构进行重新组织编排,使之成为一种新的文献类型或知识序列,比如类书、丛书、总集以及政书、年表等各种工具书的编纂等。这类古籍整理以凝练、精准、便捷和实用为基本原则。

范式 P(Preservation):古籍实体的保存性整理。它以修复和长久保存古籍实体为目的,属古籍典藏的范畴。因为古籍生成的年代相对久远,纸张老化破损严重,加上保存环境不善,虫蛀鼠咬、粘连霉烂、糟朽焦脆等现象非常普遍,所以必须对这类古籍进行装补和修复,最大限度延长其物理寿命。它以“整旧如旧”为基本原则,在整理过程中要求做到安全第一、最少干预

和过程可逆。

如图1所示,古籍整理流程包括古籍的入藏、编目、典藏、点校、出版等,而比较核心的环节是编目、典藏和点校,其中编目和典藏主要由图书馆等文献收藏机构来完成,而点校主要由

文史专家来完成。每个环节的古籍整理所遵从的原则是不一样的,编目环节由范式C发挥主导作用,典藏环节由范式P发挥主导作用,而在点校环节,范式T和范式A融合在一起共同发挥作用。

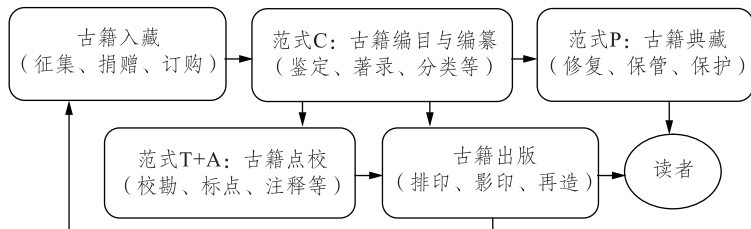


图1 传统的古籍整理流程及范式关系

2 传统古籍整理范式在数字环境下的发展和延伸

以上四种范式都是从纸本古籍整理中总结出来的,其基本原则并不会因为技术条件的发展而改易,故仍适用于数字环境下纸本古籍的整理。本文所探讨的数字环境下古籍整理范式的传承,更多是指在由纸本古籍整理向古籍数字化整理过渡的进程中,这四种范式的自然发展和延伸。

2.1 范式T:数字环境下古籍文本的复原性整理

古籍文本的复原性整理在数字环境下不仅没有削弱,反而得到了加强。这是因为,古籍一旦完成数字化,即可化身无数。版本如果选择不善,校勘不精,其贻害的范围也将被放大。因此,古籍数字化之前必须慎选底本,广校异同,审订谬误,将讹误降至最低。即便在古籍数字化完成之后,校勘工作也并未结束。由于校勘过程中存在很多人为因素,对内容的理解也存在个体差异,因此会出现多个不同校本的局面。从这个意义上讲,古籍文本的复原性整理只是无限接近古籍原貌,校勘永无止境。目前,数字环境下古籍文本的复原性整理主要体现在三个方面。

首先,计算机自动校勘被引入古籍文本的复原性整理。所谓自动校勘,是指利用计算机自动发现并标记出古籍不同版本之间的文字差异,并提供各种校勘辅助工具以帮助专家勘误^[10]。校勘可分成“校异”(即死校)和“勘误”(即活校)两个步骤,前者由电脑完成,后者由专家完成。例如,北京创新力博数码科技有限公司开发的“青典版本校勘系统”^[11],就是将多个版本的古籍经过系统比照之后,突出显示文本之间的差异,再让专家对这些差异进行考证和选择。校异的任务相对繁重,而自动校勘系统完全能够胜任,极大地提高了校勘效率。另外,东南大学的常娥选择《齐民要术》四个不同版本(以缪启愉校释本为底本,参校柯逢时清抄本、明嘉靖马直卿刻本和清光绪浙西村舍刻本)为语料开发的自动校勘实验系统,自动校勘的精准率达到了95.2%^[12]。

其次,古籍在数字化录入或文本转换时,对底本文字原样的处理体现了传统古籍整理范式的要求,缺笔讳字、异体字、俗体字以及假借字通常悉数原样保留。比如,底本中同时有“于”和“於”、“并”和“並”,如果简单地合并为“於”和“並”,就可能造成有价值的古籍文本信息的丢失。有的情况下甚至可能造成理解上的混乱,如《礼记·月令》:“(孟秋之月)修宫室,坏墙垣,补城郭。”此处“坏”读为péi,意为“用泥土

封塞空隙”,如果转换为“壤”,文意则相反。而对于仅仅是由于抄刻习惯不同或因时代、地域差异造成的异体字,则一般应合并为规范的正体字。底本中的特殊标注,如夹注、眉批、印章、画符等,对于理解文本内容和鉴定古籍版本有特定的价值,一般应照录。

再者,以“古籍电子定本工程”为代表的古籍整理软件平台通过图文对照的方式最大程度保留了古籍文本的原貌。由首都师范大学电子文献研究所牵头启动的“古籍电子定本工程”^[13],本着“择优而定,从善为本”的原则,精挑版本,反复校勘,勘定后的每种电子古籍由一套底本原图和三套电子文本(包括原图版式简体、原图版式繁体和标点整理简体)组成。底本图片和文本左右对照,既保留了古籍原貌,又使原书内容纤毫毕现。为便于直接引用,古籍电子定本一律以 Unicode 编码,支持超大字符和多语种 Windows 平台浏览,并可以适时更新和及时纠错,避免了纸本“一朝写定,终生难易”的尴尬。

2.2 范式 A: 数字环境下古籍语义的解释性整理

数字环境下,计算机信息处理技术也被应用于古籍的标点、注释和翻译。20 世纪 90 年代,东吴大学陈郁夫教授主持《古今图书集成》数字化项目时,根据类书中“艺文”和“选句”句法严整、兼有押韵,非常适合自动标点的特点,尝试对这两部分共 1 700 余万字的韵文进行了自动标点^[14]。2008 年,黄建年选取《二十五史》“食货”部分以及《齐民要术》等四部古农书为训练集,以《汜胜之书》等六部古农书为测试集,采用模式识别技术,构建模式识别库,在实施模式优化和转换后,导入到相关模式识别处理程序。测试结果显示,断句、标点的平均准确率分别达到 48% 和 35%^[15]。陈天莹等提出了基于前后文 n-gram 模型的古汉语句子切分算法,能够在数据稀疏的情况下,通过收集上下文信息,对切分位置进行比较准确的预测,从而较好地处理

小规模训练语料的情况,降低数据稀疏对切分准确率的影响。利用该方法对《论语》的切分实验,达到了 52% 的准确率^[16]。张开旭等提出了基于条件随机场(CRF)的古汉语自动断句标点方法,并引入互信息和 t-测试差两个统计量作为模型特征,分别在《史记》和《论语》两个语料库上进行了实验,也取得了较满意的效果^[17]。

对于古籍的自动注释,有学者认为,首先要建成一个庞大的冷僻字词和典故数据库,当电脑扫描古籍文本时,只要发现有与数据库中的字词相同或相似的情况,即可调用该字词的相关解释,并写到程序设定的位置;其次是检索时能适用模糊查询法,遇到自然语言的提问方式,应能自动判断是否给予注释及该调用哪条数据予以解答^[18]。常娥等也认为,通过采集古籍中的名物制度及其注释内容来构建古籍语词注释知识库,是自动注释得以完成的关键。注释内容可从三个来源获得:一是《中国历史大辞典》《中国历史地名大辞典》《故训汇纂》等专业工具书,二是各种校释本(如《十三经注疏》)中的语词及注释内容,三是由专家补充相关词汇注释资料。可考虑利用动态生成文本链接锚点(书签)技术为古籍文本添加语词注释。自动注释原理如下:先以 xml 方式存储古籍文本内容,再将这些文本内容与语词注释知识库中的语词进行匹配。匹配算法采用逆向最大匹配原则,一旦匹配成功就将该语词标记成超链锚点。用户点击这些锚点词后,系统就激活锚点词检索程序,将被点击的锚点词作为检索词对古籍注释知识库进行检索,并将检得的注释内容显示出来^[19]。

自动翻译技术经过多年的发展,已经形成了基于规则和基于语料库的两大方法体系。前者假设翻译的过程就是对源语言的词法、语法、语义和句法进行分析、判断和取舍,然后重新排列组合生成等价目标语言的过程;后者以大规模语料库的分析为基础,又可分为基于统计和基于实例两种方法,通过双语或多语料进行概率运算,依照各语言要素间的相似度来构造语

言模型,从而实现自动翻译。王爽等将基于规则和实例库的方法结合起来,构造了一个古籍自动翻译系统。该系统选用《论语》的50个句子进行测试,一般陈述句翻译的满意度达到了90%,简单疑问句和复杂句的满意度为82%,而反问句的翻译效果不够理想^[20]。郭锐等综合考虑句子长度、汉字字形、标点符号三个因素,设计和实现了古今汉语自动句对齐及相似古文句子检索算法,并指出构建大规模古今汉语平行语料库,从大量古今互译实例中检索与输入句子最相似的源句子,是基于实例的古今汉语机器翻译必须解决的首要问题^[21]。

因为对古籍语义的阐释主观性较强,且受语言特点和时代背景的限制,计算机信息处理技术应用于古籍的自动标点、注释和翻译还处于探索阶段,尚未达到理想效果。

2.3 范式C: 数字环境下古籍内容的组织性整理

计算机应用于古籍索引、书目的编制相对来说比较早,技术也较成熟。20世纪70年代,欧美国家就有学者尝试用计算机编制中文古籍索引。如,1975年德国汉堡大学的吴用彤编制了《诗经索引》,1978年美国人P. J. Ivanhoe编制了《朱熹大学章句索引》等系列经书字词索引^[22]。中国大陆及港台地区运用计算机编制古籍索引始于20世纪80年代,其发展大致经历了三个阶段。

第一,借助计算机软件编制纸质的古籍索引。如,1989年中国社会科学院语言研究所姚兆炜以《寒山子诗》为语料,利用汇编语言编制出《寒山子诗》全书索引。20世纪90年代初栾贵明以《全唐诗》数据库为基础,运用计算机软件编印的33大册《全唐诗索引》在中华书局出版。1992—2002年间,香港中文大学中国文化研究所据“先秦两汉一切传世文献电脑化资料库”编制出版的《先秦两汉古籍逐字索引丛刊》和《魏晋南北朝古籍逐字索引丛刊》^[23]也属此类。

第二,古籍索引自动编制系统。如,1995年湘潭大学研制的RPSYBJ系统,可自动编制古籍的逐字索引、句子索引、人名索引、地名索引及其他专题索引,曾用于编制《宋词别集索引三种》,具有一定的可操作性^[24]。1997年四川大学古籍所完成的全国高等院校古籍整理研究工作委员会(简称“高古委”)项目“中文索引编制系统”,能编制辞书类索引或逐字索引,并按四角号码、汉语拼音、笔画自动排序^[25]。国家古籍保护中心的“中华古籍索引库”以《全国古籍普查登记目录》为基础,系统实现了两大功能:一是索引条目的自动抽取,即从目录正文中抽取题名、著者,同时指明其所在页码或款目编号;二是索引条目的自动编排,即按拼音、部首笔画、四角号码等方式编排^[26]。

第三,以全文检索为基础的多功能古籍检索系统。如,台湾地区在建的“汉籍电子文献资料库”(旧称“瀚典全文检索系统”)可勾选指定检索栏位内的正文和注释进行全文检索,并可扩大异体字和同义词的检索范围^[27]。北京大学与北京爱如生数字化技术研究中心合作研发的中国基本古籍库拥有强大的ASE(automated software engineering)检索系统,可从分类、条目、全文三个方面对所收录的万种古籍进行全方位的检索^[28]。有学者认为,数字化古籍应以全文检索为基础,构建一整套包括关键词检索、条件检索、逻辑检索、模糊检索、组配检索、属性检索等在内的多功能古籍检索系统^[29]。

随着古籍数字化的发展,人们已不满足于古籍全文的获得,而对相同属性的知识内容的聚类有了更多需求。因此,传统的文献编纂方法在古籍数字化整理中有了用武之地,而大量建成的古籍全文数据库则为其提供了丰富的素材。如,国学网研发的《中国历代笔记》《中国历代碑帖精华》《中国古代戏剧专辑》以及各种“备览”等系列国学数字产品^[30],就是文献编纂原理应用于古籍数字化整理的结果。不过,上述成果基本上是以文献完整的“篇”“件”为单位,未能深入更细小的主题单元。为此,常娥构建

了农史专题资料自动编纂系统,选取《齐民要术》等十部重要的农业数字古籍为实验语料,利用计算机自动从中查找并摘录出与“稻”“麦”“豆”“棉”“麻”等五个主题相关的资料,并编纂成册。实验显示,系统自动编纂的结果得分良好率为72.2%^[31]。

2.4 范式 P: 数字环境下古籍实体的保存性整理

传统文献环境下,古籍的修复和保管都是针对古籍物理实体实施的,因其以不改变原件载体为目标,故称为原生性保护。数字环境下,以长期保存古籍物理实体为目的的古籍整理依然存在,但长期保存的对象不再局限于此。笔者曾提出将古籍保护体系划分为三个层级:一是针对古籍物理实体的原生性保护(primeval preservation);二是针对古籍内容的再生性保护(regeneration preservation);三是针对古籍制版工艺的非物质文化遗产保护(intangible cultural heritage preservation)^[32]。原生性保护在保护古籍物理实体的同时,虽也保护了其内容,但这种保护是非常脆弱的。再生性保护中的古籍影印、善本再造虽可以脱离古籍实体来保护其内容,但仍是在纸质文献环境中完成的,遵循的是传统的古籍整理范式。而古籍缩微胶片的制作、古籍全文数据库和影像数据库的建设,则毫无疑问是古籍保存性整理在数字环境下的进一步发展和延伸。

数字环境下古籍保存性整理已不再限于古籍实体,还包括古籍实体所负载的版本特征和制版工艺。构建专门的古籍版本数据库是未来古籍保存性整理的重要方向。赵万里先生曾选辑历代雕版印刷书籍中有代表性的样页,按版刻时代和地区编成《中国版刻图录》^[33],图文并茂地再现了同一时代不同地区的刻本在字体风格上的差异以及同一地区不同时代版刻风格的延续性。该书的编纂对于古籍版本数据库的建设具有参照意义。笔者建议,以全国古籍普查和《国家珍贵古籍名录》申报工作的前期成果为

基础,按朝代顺序、地域范围和版本类型的逻辑结构来构建中国古籍版本数据库。该库不再是简单的书目或文本的数字化,而是以保存古籍版本信息源的影像资料(比如书名页、序跋、目录、卷端、版页、牌记、字体、纸墨、藏印、插图等)为重点,同时著录主要的古籍版本信息(比如书名、作者、著述方式、出版时间、出版地、出版者、版本类型、版式行款、装帧形式等),以备存档和查考。它的建成将实现古籍版本文化遗产保护的数字化和档案化。

3 数字环境下古籍整理范式的拓新:古籍知识的数据化整理(范式 D)

2007年,计算机图灵奖得主吉姆·格雷(Jim Grey)在美国国家研究理事会计算机科学和远程通信委员会(NRC-CSTB)的一次演讲中首次提出科学研究的“第四范式”,即以数据密集型计算为基础的科学研究范式。在这个范式下,自然和人类社会活动被科学家观察、感知、计算、模仿、传播等形成的科学数据,不仅仅是科学研究的结果,更是下一步科学研究的对象和基础,“人们不仅关心数据建模、描述、组织、保存、访问、分析、复用和建立科学数据基础设施,更关心如何利用泛在网络及其内在的交互性、开放性,利用海量数据的可知识对象化、可计算化,构造基于数据的、开放协同的研究与创新模式”^[34]。“第四范式”的本质是方法论的创新,对于数字环境下的古籍整理而言,就是将经过扫描等信息技术处理后的数字化古籍,不仅看作是古籍整理的结果,更是下一步古籍整理的对象和基础——数据。通过对古籍数据的精细化加工和处理,从根本上改变古籍知识的获取、标注、比较、阐释与表现方式。这就是数字环境下古籍整理范式的最大拓新,即古籍知识的数据化处理,本文称之为范式D(Datamation)。

古籍数字化通过图片扫描、字体规范、文本处理等手段将古籍内容转变成计算机可读的数据,只是实现了古籍内容的数字化保存,但在数

据的细粒度、规范性和结构性方面还存在不足,数字古籍的大部分内容还不能被智能检索和识别应用。如果说数字化是对古籍内容简单直观的转述和再现,那么数据化就是对古籍语义的深度解析。在数据的驱动下,从知识元解析、智能检索、语义分析到关联挖掘、趋势预测、知识地图构建,古籍整理本身及其衍生的研究方式都将全面升级。

第一,随着古籍数据量的快速增长,原有针对文献本身的书目索引及文献内字词的全文检索已不能满足用户需求,而针对某个信息点或知识元素的语义检索成为一种新趋势。因此,对于超大规模的古籍数据库而言,除了实现古籍的一般性检索功能外,语义检索功能已成为一种现实需要。可利用语义分析工具将古籍文本解析成碎片化、语义化的知识元,在此基础上进行本体构建和语义标注,通过语义识别建立人物、职官、年号、地点、事件等史籍知识元之间的语义关联。如,肖怀志、李明杰选取《三国志》为例,针对纪传体因人记事的体例易造成同一史实散见于不同篇卷的弊端,构建了三国时期年号纪年的本体库,通过年号纪年本体建立的语义关联来聚集相关历史年代知识元,最终达到聚集同一或相关史实的目的^[35]。也可通过构建历史人物本体库,建立不同人物之间的网络关系和人物与事件的关联信息,在检索相关人物、事件或人物别称、特征词时,就能根据古籍数据库的内容生成以检索词为中心的多维度信息,并可利用可视化技术将其立体地呈现出来。如,中华书局已经对“二十四史”建立了基于本体的知识检索模型,不仅能够实现时间、地点、人物的知识关联检索,还可通过可视化的形式展示人物、时间、地点和事件之间的语义关系。在此基础上建立的“中华基本史籍知识库”,能提供传统纸质图书无法实现的知识检索、类聚、链接以及知识提示等服务功能^[36]。随着语义化、多维度检索功能的实现,超大规模古籍数据库将为人文社会科学研究提供全方位的知识解决方案。

第二,随着古籍数据细粒化程度的加深,除了对古籍文本内容作传统的定量分析之外,还可在数据化的古籍文本中对作者及其作品的相关情况(比如作者的籍贯郡望、家族成员、求学交游、科举仕宦,作品的文体构成、遣词习惯、语言风格、用典情况、后人评价等虚化的信息)进行量化标注,为考证某些佚名作品的作者和创作年代,或评价某个朝代某类体裁作品的创作水平提供全面、准确的数据。比如,古人用典的历史,依靠传统方法虽然可以考辨,但往往只限于一人一时之局部问题。而基于古籍的大数据分析,则可对历朝历代文学作品中的所有用典情况进行宏观而精准的分析。例如,可根据用典数量和用典频率的统计分析,从整体上比较唐宋元明清诗歌的风格倾向的差异。在古代作品的评价方面,武汉大学的王兆鹏教授曾选用现存词作篇数、现存宋词别集版本种数、历代词话中被品评次数、本世纪被研究评论的论著篇(种)数、历代词选中宋代词人入选词作篇数、本世纪词选中两宋词人入选的词作篇数六大指标对两宋词人进行统计排名,总结出了宋词创作的系列规律与质量评级方法^[37]。他还采用类似的方法,统计出了“唐诗名篇排行榜”。此外,古籍中的图像材料及特定文学作品的押韵、平仄、对仗等文体信息也可进行样本标注和智能识别,这些非文本信息转化为文本信息后,再通过信息集成和数据分析可望实现知识发现的功能。

第三,随着读者知识需求的个性化越来越强,利用大数据开展交互式、个性化的知识服务,也是未来数字古籍整理的方向。它既能满足科学研究者的专业需要,也能满足普通读者的一般文化需求。比如,用户在系统中输入作者的姓名、字号、室名等,即可获得人物的生平行状、传记、墓志及后世评价等方面的资料;输入同一时代或多个时代的多个人的姓名、字号,则可通过海量数据的关联分析,发现他们之间横向或纵向的社会关系,并生成人际网络图;输入百家姓中的某个姓氏,就能用地理信息系

统的方式展示该姓氏的起源、世系图谱、历代名人、地理分布、家族迁徙等情况;输入某个朝代年号,即可获知当时的社会风尚、热门话题、审美标准,甚至是当时流传最广的新闻八卦;输入一个地名,就能获知该地的地理名称沿革、行政区划、民俗物产、名胜古迹、地方人物、地方著述等情况;输入某个职官名称,就能知道该职官的历史沿革、职责范围和上下级关系;输入某个主题词和对应的体裁,即可检索历史上已有的文学作品,甚至可由电脑自动生成合乎主题和体裁要求的作品。基于数据分析和处理的古籍知识整理还有很多不可预知的应用领域,有待人们去深入研究和开发。

4 结语

中国古籍整理范式大致经历了三个发展时期:第一时期始于西汉刘向而止于近代。汉成帝时刘向等人开创的文献整理程序和方法,因

其广泛的适用性和深远的影响力,成为这一时期中国古籍整理的基本范式,并在两千多年间呈现出一种稳定的状态。第二阶段始于近代。随着学科的发展和分化,传统的古籍整理范式为了适应不同学科的现实需要,逐渐细化出若干基于不同“问题”和“目的”的范式分支,即古籍文本的复原性整理、古籍语义的解释性整理、古籍内容的组织性整理和古籍实体的保存性整理。这四种范式在各自的适用范围内对古籍整理的具体要求是完全不一样的。第三阶段大致始于20世纪70年代。随着信息技术在古籍整理领域内的应用,以上四种范式逐渐从纸本古籍的整理发展延伸到古籍的数字化整理当中,并逐渐衍生出一种新的古籍整理范式——古籍知识的数据化整理,即不仅把数字化古籍当作古籍整理的结果,更是把它当作下一步古籍整理的对象和数据。由此引发了数字古籍整理方式的深度变革,其对人文社会科学研究方法带来的影响是不可估量的。

参考文献

- [1] 托马斯·库恩. 科学革命的结构[M]. 金吾伦, 胡新和, 译. 北京: 北京大学出版社, 2003: 40. (Kuhn T S. The structure of scientific revolutions[M]. Jin Wulun, Hu Xinhe, trans. Beijing: Peking University Press, 2003: 40.)
- [2] 魏徵. 隋书·经籍志[M]. 北京: 中华书局, 1973: 903. (Wei Zheng. Bibliography in the chronicle of Sui Dynasty [M]. Beijing: Zhonghua Book Company, 1973: 903.)
- [3] 黄永年. 古籍整理概论[M]. 西安: 陕西人民出版社, 1985: 5. (Huang Yongnian. Generality of the collation of ancient books[M]. Xi'an: Shaanxi People's Publishing House, 1985: 5.)
- [4] 刘琳, 吴洪泽. 古籍整理学[M]. 成都: 四川大学出版社, 2003: 2. (Liu Lin, Wu Hongze. Studies of the collation of ancient books[M]. Chengdu: Sichuan University Press, 2003: 2.)
- [5] 时永乐. 古籍整理教程[M]. 保定: 河北大学出版社, 2003: 3. (Shi Yongle. Ancient book collation course[M]. Baoding: Hebei University Press, 2003: 3.)
- [6] 曹林娣. 古籍整理概论[M]. 北京: 北京大学出版社, 2007: 5. (Cao Lindi. Generality of the collation of ancient books[M]. Beijing: Peking University Press, 2007: 5.)
- [7] 北京大学图书馆学系, 武汉大学图书馆学系. 图书馆古籍整理·前言[M]. 油印本. 1982: 4. (Peking University's Department of Library Science, Wuhan University's Department of Library Science. Preface of the collation of ancient books in libraries[M]. Mimeograph. 1982: 4.)
- [8] 王世伟. 图书馆古籍整理工作[M]. 北京: 北京图书馆出版社, 2000: 1. (Wang Shiwei. Collation of ancient books in libraries[M]. Beijing: Beijing Library Press, 2001: 1.)
- [9] 来新夏. 古籍整理散论[M]. 北京: 书目文献出版社, 1994. (Lai Xinxia. On the collation of ancient books[M]. Beijing: Bibliographic Literature Press, 1994.)

- [10] 常娥. 古籍智能处理技术研究[D]. 南京:南京农业大学,2007:71. (Chang E. Research on intelligent processing technologies for Chinese ancient books[D]. Nanjing:Nanjing Agricultural University, 2007:71.)
- [11] 北京创新力博数码科技有限公司. 青典数字化系统[EB/OL]. [2015-04-15]. <http://www.ilibo.com/ProductInfo.aspx?h01PkNywEHgACb5Guz8NRA%3d%3d=Y80K5Ghutw0%3d>. (Beijing Ilibo Digital Technology Co., Ltd. Qingdian Digital System[EB/OL]. [2015-04-15]. <http://www.ilibo.com/ProductInfo.aspx?h01PkNywEHgACb5Guz8NRA%3d%3d=Y80K5Ghutw0%3d>.)
- [12] 常娥,侯汉清,曹玲. 古籍自动校勘的研究和实现[J]. 中文信息学报,2007(2):83-88. (Chang E, Hou Hanqing, Cao Ling. Research on the automatic version comparison and analysis of ancient book and it's realization [J]. Journal of Chinese Information Processing, 2007(2):83-88.)
- [13] 尹小林. 关于“古籍电子定本工程”方案[J]. 中国索引,2009(3):25-29. (Yin Xiaolin. On the project of making the fair copy of the final version of digital ancient books[J]. Journal of the China Society of Indexers, 2009(3):25-29.)
- [14] 陈郁夫. 中文智能处理技术:古文标点自动化[EB/OL]. [2015-04-15]. <http://www.yyxx.sdu.edu.cn/content/guojihuiyi/guojihy-cyf.htm>. (Chen Yufu. Intelligent processing technologies;the automatic punctuation of Chinese ancient books [EB/OL]. [2015-04-15]. <http://www.yyxx.sdu.edu.cn/content/guojihuiyi/guojihy-cyf.htm>.)
- [15] 黄建年,侯汉清. 农业古籍断句标点模式研究[J]. 中文信息学报,2008(4):31-38. (Huang Jiannian, Hou Hanqing. On sentence segmentation and punctuation model for ancient books on agriculture[J]. Journal of Chinese Information Processing, 2008(4):31-38.)
- [16] 陈天堂,陈蓉,潘璐璐,等. 基于前后文 n-gram 模型的古汉语句子切分[J]. 计算机工程,2007(3):192-193. (Chen Tianying, Chen Rong, Pan Lulu, et al. Archaic Chinese punctuating sentences based on context n-gram model[J]. Computer Engineering, 2007(3):192-193.)
- [17] 张开旭,夏云庆,宇航. 基于条件随机场的古汉语自动断句与标点方法[J]. 清华大学学报(自然版),2009(10):163-166. (Zhang Kaixu, Xia Yunqing, Yu Hang. CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose[J]. Journal of Tsinghua University (Science and Technology), 2009(10):163-166.)
- [18] 林尔正,林丹红. 计算机应用于古籍整理研究概况[J]. 情报探索,2007(6):28-29. (Lin Er'zheng, Lin Danhong. General research situation of the utilization of computer in the collation of ancient books[J]. Information Research, 2007(6):28-29.)
- [19] 常娥,黄建年,侯汉清. 古籍智能整理与开发系统构建研究[J]. 情报资料工作,2009(4):43-47. (Chang E, Huang Jiannian, Hou Hanqing. Research on the construction of the intelligent system of collation and utilization of ancient books[J]. Information and Documentation Services, 2009(4):43-47.)
- [20] 王爽,熊德兰,王晓霞. 古文翻译系统的设计与实现[J]. 电脑知识与技术,2009(4):855-857. (Wang Shuang, Xiong Delan, Wang Xiaoxia. The design and implementation of ancient Chinese translation systems[J]. Computer Knowledge and Technology, 2009(4):855-857.)
- [21] 郭锐,宋继华,廖敏. 基于自动句对齐的相似古文句子检索[J]. 中文信息学报,2008(2):87-91. (Guo Rui, Song Jihua, Liao Min. Ancient sentence search based on sentence auto-alignment in parallel corpus of ancient and modern Chinese[J]. Journal of Chinese Information Processing, 2008(2):87-91.)
- [22] 毛建军. 古籍索引的电子化实践[J]. 中国索引,2006(4):37-40. (Mao Jianjun. Practice of electronizing ancient book indexes[J]. Journal of the China Society of Indexers, 2006(4):37-40.)
- [23] 黄建年. 中国古籍索引编制软件概述[J]. 图书馆学研究,2011(10):65-68. (Huang Jiannian. The development of indexing systems of Chinese ancient books[J]. Research on Library Science, 2011(10):65-68.)
- [24] 黄建年. 汉文古籍索引自动化实践与研究概述[J]. 佛山科学技术学院学报(社科版),2011(6):50-56.

- (Huang Jiannian. The overview of Chinese ancient books index automation[J]. Journal of Foshan University (social science edition), 2011(6):50-56.)
- [25] 黄建年. 中国古籍索引编制软件概述[J]. 图书馆学研究, 2011(10):65-68. (Huang Jiannian. The development of indexing systems of Chinese ancient books[J]. Research on Library Science, 2011(10):65-68.)
- [26] 包菊香. 古籍目录索引的自动编制——以“中华古籍索引库”为例[J]. 中国索引, 2013(1):25-29. (Bao Juxiang. Automatic establishment of ancient book indexes; a case study on Chinese Ancient Book Index Database [J]. Journal of the China Society of Indexers, 2013(1):25-29.)
- [27] “中央研究院”历史语言研究所. 汉籍电子文献资料库[EB/OL]. [2015-05-16]. <http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>. (Institute of History and Philology Academia Sinica. Scripta Sinica Database[EB/OL]. [2015-05-16]. <http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>.)
- [28] 毛建军. 《中国基本古籍库》的特色与启示——兼谈古籍全文数据库的标准与规范[J]. 管理学报, 2009(1):104-106. (Mao Jianjun. Characteristics and inspirations on Chinese classic ancient book database; on the standards and norms of ancient book full-text databases[J]. Journal of Management, 2009(1):104-106.)
- [29] 常娥. 古籍智能处理技术研究[D]. 南京:南京农业大学, 2007:32. (Chang E. Research on intelligent processing technologies for Chinese ancient books[D]. Nanjing: Nanjing Agricultural University, 2007:32.)
- [30] 国学网. 国学产品[EB/OL]. [2015-05-16]. <http://www.guoxue.com/cp/cpfront.htm>. (Guoxue. Products [EB/OL]. [2015-05-16]. <http://www.guoxue.com/cp/cpfront.htm>.)
- [31] 常娥. 农史专题资料自动编纂系统的构建与测试[J]. 图书馆学研究, 2009(6):10-14. (Chang E. Construction and testing of the automatic compilation system of Chinese agricultural ancient books[J]. Researches in Library Science, 2009(6):10-14.)
- [32] 李明杰. 构建中华古籍层级保护体系的设想[J]. 图书馆杂志, 2009(3):14-19. (Li Mingjie. A conception on construction of a multi-layered preservation system for Chinese ancient books[J]. Library Journal, 2009(3):14-19.)
- [33] 赵万里. 中国版刻图录[M]. 北京:文物出版社, 1960-1961. (Zhao Wanli. The album of woodblock-printing books in ancient China[M]. Beijing: Cultural Relics Press, 1960-1961.)
- [34] 刘量. 科学研究的第四范式[N]. 中国读书报, 2012-11-14(20). (Liu Liang. The fourth paradigm of scientific research[N]. China Reading Weekly, 2012-11-14(20).)
- [35] 肖怀志, 李明杰. 基于本体的历史年代知识元在古籍数字化中的应用——以《三国志》历史年代知识元的抽取、存储和表示为例[J]. 图书情报知识, 2005(3):28-33. (Xiao Huaizhi, Li Mingjie. The application of historical knowledge elements in the digitization of ancient books based on ontology: a case study of *The Chronicle of Three Kingdoms* [J]. Journal of Documentation, Information and Knowledge, 2005(3):28-33.)
- [36] 顾青. 真正的古籍数字化是建立一套综合的知识服务体系[EB/OL]. [2015-05-16]. <http://www.zhbc.com.cn/shownews.asp?id=2099>. (Gu Qing. The real digitization of ancient books is to establish a series of comprehensive knowledge service systems [EB/OL]. [2015-05-16]. <http://www.zhbc.com.cn/shownews.asp?id=2099>.)
- [37] 王兆鹏. 宋代词人历史地位的定量分析[M]//王兆鹏. 唐宋词史论. 北京:人民文学出版社, 2000:81. (Wang Zhaopeng. A quantitative analysis of the poets' historical position of Song Dynasty[M]//Wang Zhaopeng. Historical comments on poetries of Tang Dynasty and Song Dynasty. Beijing: People's Literature Publishing House, 2000:81.)

李明杰 武汉大学信息管理学院教授, 博士生导师, 《图书情报知识》编辑。湖北 武汉 430072。

(收稿日期: 2015-06-11; 修回日期: 2015-07-07)