



中华农业文明研究院文库
文化典籍计算机整理与知识组织丛书

黄建年◎著

古籍计算机断句标点与分词标引研究

丛书主编 侯汉清



中华农业文明研究院文库
文化典籍计算机整理与知识组织丛书

丛书主编 侯汉清

古籍计算机断句标点与分词标引研究

黄建年◎著

安徽师范大学出版社

本书研究和出版获得如下基金项目的资助：①2008 年国家社会科学基金重点项目“文化典籍整理与开发智能技术研究”（编号：08ATQ002）；②2008 年教育部人文社会科学基金项目“农业古籍自动分词及索引编制研究”（编号：08JA870006）

图书在版编目（CIP）数据

古籍计算机断句标点与分词标引研究/黄建年著. —芜湖：安徽师范大学出版社，2011. 12

（中华农业文明研究院文库·文化典籍计算机整理与知识组织丛书）

ISBN 978 - 7 - 81141 - 674 - 9

I. ①古… II. ①黄… III. ①计算机应用—古汉语—句读—研究②古籍—自动标引—研究
IV. ①H141 - 39②G356. 6

中国版本图书馆 CIP 数据核字（2011）第 259242 号

古籍计算机断句标点与分词标引研究

黄建年 著

出 版 人：张传开

责任编辑：潘 安

装帧设计：丁奕奕

出版发行：安徽师范大学出版社

芜湖市九华南路 189 号安徽师范大学花津校区 邮政编码：241002

网 址：<http://www.ahnupress.com/>

发 行 部：0553 - 3883578 5910327 5910310（传真）

经 销：全国新华书店

印 刷：安徽芜湖新华印务有限责任公司

版 次：2012 年 2 月第 1 版

印 次：2012 年 2 月第 1 次印刷

规 格：787 × 1092 1/16

印 张：10

字 数：200 千

书 号：ISBN 978 - 7 - 81141 - 674 - 9

定 价：20.00 元

凡安徽师范大学出版社版图书有缺漏页、残破等质量问题，本社负责调换

序 一

中国农业的发展历史有上万年，但对农业历史进行有组织的整理和研究时间却不长，大致始于 20 世纪 20 年代。1920 年金陵大学建立农业图书研究部，启动中国古代农业资料的收集、整理和研究。同年，中国农史事业的开拓者之一万国鼎先生（1897 年—1963 年）从金大毕业留校工作，发表了第一篇农史学术论文《中国蚕业史》。1924 年，万国鼎先生就任金陵大学农业图书研究部主任，亲自主持了《先农集成》等农业历史资料的整理与研究工作。1932 年金陵大学改农业图书研究部为金陵大学农经系农业历史组，农史工作从单纯资料整理和研究向科学普及和人才培养拓展，万国鼎先生亲自主讲了“中国农业史”和“中国田制史”等课程，农业历史的研究受到了更为广泛的关注。1955 年，在周恩来总理的亲自关心和支持下，农业部批准建立由中国农业科学院和南京农学院双重领导的中国农业遗产研究室，万国鼎先生被任命为主任。在万先生的带领下，南京农业大学中国农业历史的研究工作发展迅速，硕果累累，中国农业遗产研究室成为国内公认、享誉国际的中国农业历史研究中心。2001 年，南京农业大学在对相关学科力量进一步整合的基础上组建了中华农业文明研究院。中华农业文明研究院继承了自金陵大学农业图书研究部创建以来的学术资源和学术传统。80 余年风雨征程，80 春秋耕耘不辍，中华农业文明研究院已发展成为一个特色鲜明、实力雄厚的以农业历史文化为优势的文科研究机构，取得了骄人的业绩：中华农业文明研究院出版了《中国农史》；创建中国高校第一个中华农业文明博物馆；先后投入 300 多万元开展中国农业遗产数字化的研究工作，建成“中国农业遗产信息平台”和“中华农业文明网”；承担中国科学技术史学会农学史专业委员会、中国农业历史学会畜牧兽医史专业委员会等学术机构的组织和管理的工作；形成农业历史科学研究、人才培养、学



术交流、信息收集和传播展示“五位一体”的发展格局。万国鼎先生毕生倡导和为之奋斗的事业正在进一步发扬光大。

中华农业文明研究院有着整理和编辑学术著作的优良传统。早在金陵大学时期，农业历史研究组就搜集和整理了《先农集成》456册。1956年—1959年，在万国鼎先生的组织领导下，遗产室整理出《中国农史资料续编》157册。60年代初，辑成《地方志综合资料》、《地方志分类资料》及《地方志物产》共689册。在这些宝贵资料的基础上，遗产室陆续出版了《中国农学遗产选集》稻、麦、粮食作物、棉、麻、豆类、油料作物、柑橘等八大专辑，出版《农业遗产研究集刊》、《农史研究集刊》等，撰写了《中国农学史》等重要学术著作，为学术研究工作提供了极大的便利，受到国内外农史学人的广泛赞誉。

为了进一步提升科学研究工作的水平，加强农史专门人才的培养，2005年85周年院庆之际，研究院启动了《中华农业文明研究院文库》（以下简称《文库》）。《文库》主要以中华农业文明研究院科学研究工作为依托，以学术专著为主，也包括部分经过整理的、有重要参考价值的学术资料。侯汉清先生主编的《文化典籍计算机整理与知识组织丛书》现纳入《文库》。我们希望，对前辈的工作有所继承又有所发展，希望更多地关注经济与社会发展，而不是就历史谈历史、就技术言技术。万国鼎先生倡导，做学术研究时要将“学理之研究、现实之调查、历史之探讨”结合起来。研究农业历史，眼光不能仅仅局限于农业内部，而要关注农业发展与社会变迁的关系、农业发展与经济变迁的关系、农业发展与环境变迁的关系、农业发展与文化变迁的关系，为今天中国农业与农村的健康发展提供历史借鉴。

中华农业文明研究院院长
教授、博士生导师

二〇一一年十二月六日

序 二

南京农业大学侯汉清教授主编的《文化典籍计算机整理与知识组织丛书》，作者均为侯老师指导的博士生。我曾有幸参加过这一研究团队的博士论文答辩，与各位作者相熟经年，对丛书内容也稍熟稔。丛书源于博士论文，又高于博士论文，或修改章节，或增加附录，或补充最新科研进展，或各取精华合二为一。

丛书以古籍中的农业文献及农史信息资源为研究对象，利用计算机技术及现代情报技术进行整理与开发，但各册研究重点并非完全一致，或标点与分词，或编纂与校勘，或知识组织，或内容挖掘，或索引编制，或关注信息门户，或研究古籍数字化技术。虽各有分工、各有侧重，但却互相补充、紧密联系。丛书的各位作者不尚空谈，唯务实证。他们分别构建了实现不同功能的原型系统，相关的结论亦基于翔实的数据和近乎实用的计算机系统之上，因而言之有据，令人信服。侯教授埋头苦干、踏踏实实的治学态度在学界有着很好的口碑，其博士生群体能够继承这一作风，不仅使我看到了严谨、求实治学精神的传承与弘扬，也使我看到了亲密无间的师生情谊。

这些成果是这一博士生团队将计算机技术及现代图书情报知识全面应用于中国古籍整理的一次尝试。自2003年9月伊始，这一团队就一直沉浸于中国古籍整理的研究与开发，取得了可喜的成果。他们的研究由浅入深，由表及里，逐步深入古籍的内容，对古籍中所载的知识进行分析、组织和挖掘，探讨了基于内容的知识组织方法与技术。看惯了以现代文本为处理对象的图书情报学论著，也阅读过一些古籍整理研究的专著，但甚少见及利用现代计算机及图书情报技术研究文化典籍的系列著作，侯老师与其弟子合作编著的这套丛书无疑令人有耳目一新之感。



学术研究，尤其是古籍整理研究，须有“板凳坐得十年冷”的耐心与定力，需要付出长久的努力。侯老师竟多年之功以成其事，实属不易。而其学生一门深入、坚持不懈，亦属难能。据我了解，侯老师尚有几位弟子于毕业后一直从事古籍整理研究，近年来新作不断涌现，令人欣慰。

近十年来，国泰民安、政通人和，传统文化日益受到重视，有关科研机构加大了对古籍整理研究的资助力度。本套丛书的研究与出版得到了2008年国家社会科学基金重点项目“文化典籍整理与开发智能技术研究”（编号：08ATQ002）和2008年教育部人文社会科学基金项目“农业古籍自动分词及索引编制研究”（编号：08JA870006）、2011年教育部人文社会科学基金项目“民国文献数字化信息组织和智能处理——以民国报纸为例的研究”（编号：11YJA870023）等基金项目的资助，实在与国家关注文化典籍整理紧密相关。能够获得诸多项目的赞助，不仅说明了侯老师及其团队选题的精准，也说明这样的研究达到了一定的深度。

当然，这套丛书仍然有未尽完善之处：部分内容出现一定程度的重复，仍可以作进一步的精简优化；计算机古籍整理软件，比如断句标点、编纂、校勘、索引编辑功能等方面的实用性，亦存在较大的提升空间。瑕不掩瑜，侯汉清教授长期致力于现代文本的智能信息处理技术研究，但能及时而又敏锐地捕捉到面向古籍文本、基于内容的古籍整理学研究热点开拓新的研究领域，并迅速带领团队取得令人敬佩的成绩，的确难能可贵。

愿侯老师的团队继续前行，为古籍整理和知识管理做出更大贡献！

南京大学教授
博士生导师

二〇一一年十二月六日

序 三

中国文化典籍是中华民族在数千年历史发展过程中创造的重要文明成果，蕴含着中华民族特有的精神价值、思维方式和想象力、创造力，是中华文明绵延不绝的历史见证，也是人类文明的瑰宝。对古籍的整理、保护与开发是炎黄子孙应尽的义务和职责。

中国的古籍整理，有着悠久的历史 and 传统。从孔子删定《六经》、刘向父子编撰《七略》，到清人编定《四库全书》、《古今图书集成》，大规模的古籍整理持续不断，影响深远。新中国成立后古籍整理领域取得的巨大成就举世瞩目。尽管如此，农业古籍的整理与开发仍然不足，已经整理出版的农业古籍只占全部存世农业古籍的 15% 左右，大量的农业古籍亟待整理。

中国古籍的断句标点东汉时期已经开始了，其后各代这一工作连绵不绝，只是盛衰易势、治乱更迭而已。明代《永乐大典》所收各书无一不加圈点，而综观清代《四库全书》全书却无一圈一点，虽同为官方编撰类书，然差距之大、观念之异令人惊奇。民国后新式标点兴起，古籍断句标点之风方始流行，新中国成立后政府更是极力推动，新式标点整理古籍因而数量渐增。1989 年我国制订《信息处理现代汉语分词规范》，然以现代文本为处理对象，而应用于古籍整理领域的专用古籍文本的分词规范尚未引起注意。

正是基于这一现状，本书以农业古籍为研究对象，研究古籍断句标点、分词的历史与现状，重点探讨计算机技术在农业古籍断句标点、分词标引中的应用，构建农业古籍断句标点、分词标引的原型系统。主要研究内容如下：

第一，借鉴文本模式匹配、句法分析等技术，研究并设计出农业古



籍自动断句标点的算法,设计出农业古籍断句标点的原型系统。

通过对约 2 000 万古籍文本汉字的统计与分析,总结出断句标点常用的 11 种方法。首先采用句法特征词、同义语标志词进行初步断句;其次利用反义复合词、引书标志、时序、数量词、重叠字词、动名结构及比较句法进一步对子句进行断句、标点;最后使用农业用语和禁用模式进一步提高断句、标点 after 农业古籍的可读性和准确性。

根据这些断句标点方法与规则,采用自动构建与人工优化相结合的办法,构建断句标点模式库与断句标点禁用模式库两类断句标点知识库。两者共同保证了断句标点功能的正常开展,目前已经构建的标点模式库共有 1 166 条规则,断句标点禁用模式库共 184 条规则。

依据这些断句标点规则,利用本系统对 6 种农业古籍的断句标点测试,取得了 60.5% 的断句正确率与 40.5% 的标点正确率。

第二,借助 N 元分词、词典分词等技术,设计出农业古籍自动分词专用算法,设计出农业古籍分词的原型系统。

考虑到目前尚无现成的古籍分词词典可用,因此构建古籍分词词典势在必行。而构建一部全面而权威的古籍分词词典又非短时间所能够完成,因此采用基于词典分词与 N 元语法分词的综合分词方法成为目前古籍分词较理想的方法。

本实验构建了基本词典群与禁用词典群等两个分词词典群共 10 多个数据库,其中基本词典群包括人名、地名、书名、职官名、物产名等数据库,而禁用词典群则包括成语、年号、虚词、数量词、时序词等数据库。分词词典群目前共收录各类词汇 20 万条,基本上满足了古籍分词的需要。

以分词词典词汇为切分标志的 N 元语法综合方法进行古籍分词,并采用子串比较过滤、相邻词过滤、高频词过滤、低频词过滤等方法对分词结果进行过滤,分别以 12 种农业古籍和 379 种《广东方志物产》为语料进行了古籍分词测试。从 12 种农业古籍中共识别出旧词 1 164 个,约



占总词汇量的 31%；新词 2 530 个，占总词汇的 69%。从 379 种广东方志物产资料中共识别出旧词 6 314 个，占总词汇的 8%；新词 75 438 个，则占总词汇的 92%。其中：出现 10 次以上的词汇为 8 044 个，占总词汇的 10%；出现 20 次以上的词汇共 3 760 个，占总词汇的 5% 左右。

通过对 379 种《广东方志物产》分词结果的分析，我们发现：当词频等级位于区间（2 000 ~ 8 000）时，词频等级与频次乘积基本为常数 23 000 000，这一结果说明齐夫第一定律在古籍文本中同样适用。

利用计算机实现农业古籍文本的断句标点与分词标引，并开发相应的系统，是农业史、情报学、中文信息处理技术等结合的产物。正因为如此，本项研究仍然稍显稚嫩，尚有进一步深入的必要。

第一，目前采用的断句标点模式库共有各种规则 1 100 余条，数量有限，而且各个规则之间仍然有待于进一步梳理与优化。另外，目前断句标点所用的方法为模式识别方法，主要基于语词的应用，而对于句法特征的采用仍然有限。这主要因为目前缺少农业古籍分析的熟语料库，特别是缺少农业古籍词汇属性库，使本实验难以进行有效的句法分析。随着农业古籍词汇属性库的构建，古籍句法规则的分析将会逐步深入，立足于农业古籍词汇属性库与古籍句法规则库的断句标点将会取得更佳的效果。

第二，分词系统采用词典分词与 N 元语法结合的综合分词方法，通过词典识别出的词汇占总词汇的比例仍然不高，在农业古籍中占 31%，而在方志物产中其比例却低于 10%，这一结果很显然乃分词词典收录的词汇在各子学科中分布并不均匀所致。所以，优化分词词典将是下一步需要继续研究的问题之一。

本课题得到了国家社科基金与教育部人文社会科学基金的支持。因为课题涉及的范围太过广泛，且因为时间因素，很难对涉及的方方面面作全面而深入的探讨，所以有些问题只能留待于将来作更进一步的探讨与研究。



科学技术的发展永无止境，对古籍整理学的追求与探讨也同样如此。前人的研究与探索为我们后人的研究开创了一条大道，我们现在正在从事的研究与探索或许同样将对后人有所寸进之功。但愿这种孜孜以求、上下求索的精神继续激励一代一代学人继续前行。

是为序。

草于 2011 年坤月吉日

目 录

序 一.....	王思明
序 二.....	叶继元
序 三.....	黄建年
1 绪 论.....	1
1.1 课题依据及意义.....	1
1.2 国内外研究综述.....	2
1.3 研究的主要理论与技术路线.....	6
1.4 研究的主要内容、结构与创新之处	8
2 古籍断句标点技术研究	12
2.1 断句标点概述	12
2.2 古籍自动断句标点技术进展	28
2.3 古籍自动断句标点算法、流程与功能设计.....	33
2.4 实验结果评价与展望	55
2.5 本章小结	57
3 古籍分词标引技术研究	61
3.1 古籍文本分词标引研究进展	61
3.2 分词标引理论与关键技术研究	66
3.3 自动分词主要算法与流程	68
3.4 分词效果测试	74
3.5 分词结果分析	77
3.6 分词结果应用	83
3.7 本章小结	85



4 古籍整理与开发系统的构建与集成	91
4.1 系统开发背景	91
4.2 古籍断句标点子系统	92
4.3 古籍分词子系统	96
4.4 古籍系统设置子系统	100
4.5 本章小结	101
5 结 语	102
6 附 录	104
附录一 计算机断句样例	104
附录二 计算机标点样例	106
附录三 标点规则库样例	110
附录四 计算机分词样例	112
附录五 常用农业史资料、索引一览表	116
附录六 新中国农业古籍整理出版简目	121
附录七 《广州府志》等四种古籍索引样例	129
附录八 全书索引	144
后 记	147

1 绪 论

1.1 课题依据及意义

中国文化典籍是中华民族在数千年历史发展过程中创造的重要文明成果，蕴含着中华民族特有的精神价值、思维方式和想象力、创造力，是中华文明绵延数千年的历史见证，也是人类文明的瑰宝。对古籍的整理、保护与开发是炎黄子孙应尽的义务和职责。

古籍整理，有广义与狭义之分。现在一般人把有关古籍各方面的学术工作统称为古籍整理，这是广义的古籍整理。从这个意义上来说，它研究的范围包括古文献的种类、源流、形制、版本、功用、断代、辨伪、编目、整理、收藏、传播等。其中很多方面又可以形成独立的学问或分支学科，例如研究古书形制发展的“书史”，研究版本的古籍版本学，研究古书编目的古籍目录学，研究古书断代与辨伪的文献断代辨伪学，等等。这样的定义显然容易与古文献学相混。但实际上，古籍整理乃是着重研究古文献学中的“整理”这个方面，它是古文献学的一个重要组成部分，二者还是有着相对明确的泾渭分野。

当然，严格意义上的古籍整理，刘琳、吴洪泽在其专著《古籍整理学》中认为“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。比如校勘以订正文字，标点以分清句读，注释以阐明文义，翻译以通连古今，辑佚以摭拾遗文，抄纂以采其菁华等等”。^[1]²换句话说，古籍校勘、标点、注释、翻译、辑佚、编纂才是古籍整理的主要内容。

黄永年认为，古籍整理“是对原有的古籍做种种加工，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”。这样的定义显然较为宽泛，但是“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”^[2]⁵，明确将古籍研究与古籍整理分开。为了进一步说明古籍整理的领域，黄氏还列举了10种基本工作内容，即选择底本、影印、校勘、辑佚、标点、注释、今译、索引、序跋、附录，内容要比刘琳、吴洪泽二氏所列的广泛一些。

来新夏在《古籍整理散论》和《古籍整理讲义》中，共论及分类、目录、版本、句读、工具书、校勘、考据、传注等八个方面的内容。时永乐先生的《古籍整



理教程》分为古籍版本、校勘、标点、注释、辨伪与辑佚、古籍整理的其他工作（撰写序跋、编辑附录、编制索引）等 8 个部分进行论述。

程毅中则认为，古籍整理“除校勘文字之外，还包括了编纂（如新编的《全宋词》）、标点、注释、今译和辑佚（如果原书有佚文的话），还有写序跋、编索引、选择影印底本等任务”。^{[3]2}

表 1-1 古籍整理主要内容对照表

专著名称	主要工作内容
《古籍整理概论》	选择底本、影印、校勘、辑佚、标点、注释、索引
《古籍整理散论》	分类、目录、版本、句读、工具书、校勘、考据、传注
《古籍整理教程》	古籍版本、校勘、标点、注释、辨伪与辑佚、古籍整理的其他工作（撰写序跋、编辑附录、编制索引）
《古籍整理学》	古籍校勘、标点、注释、翻译、辑佚、编纂
《古籍整理浅谈》	编纂、标点、注释、今译、辑佚、写序跋、编索引、选择影印底本

综上所述，有关古籍整理包括的具体内容，专家的观点也并不完全一致，但是，对于核心部分还是有着共同的看法。比如校勘、标点（其中 1 家标为句读）、辑佚、注译为上述 5 家共同主张，索引为 4 家共同主张（《古籍整理散论》工具书部分包括索引^{[4]110}），可见校勘、标点、辑佚、注译以及索引才是古籍整理的核心部分。在本文中，古籍整理即采用这样的定义。

我国是历史悠久的农业大国，历代留下来的农业古籍十分丰富，数量庞大，它们是我国文化遗产的重要组成部分，史学价值极高。因此，农业古籍的数字化整理对传统文化的保存和研究具有深远的影响。随着现代信息技术（主要是中文信息处理和计算机技术）的发展，已经为我们从更广的领域、更深的层次进行农业古籍数字化整理加工，从而向读者和研究人员提供全新的服务创造了条件。如何借鉴已有的中文信息处理技术（如自动分词、模式识别等）和农业古籍数字化成果，并结合农业古籍自身特点，建立一个断句标点、分词标引等功能的农业古籍整理与开发系统，是本书研究的重点。利用这些功能，不仅可以极大地改善农史研究者的研究条件，而且还会带来研究思路、研究方法的变革。

1.2 国内外研究综述

在现代社会，技术与系统已经密不可分。技术通常通过系统来体现，一个先进的技术必须以系统达到最优的运行效率为前提；同样，一个完善的系统需要集成不同的技术，以保证良好的性能。古籍整理领域也不例外。

古籍智能处理技术，包括两层含义：第一，智能处理理论、技术在古籍整理中



的应用;第二,古籍智能处理的各种系统、各种平台。基于此,我们将具体的技术与系统分开论述。

智能技术在古籍整理手段的应用有一个从外面形态加工到内容加工的渐进过程。将古籍拍成缩微胶片保存,将古籍影印后校点出版,比传统的书库存版、手刻雕印效率提高许多。古籍排版、古籍扫描、古籍光盘、古籍网络数据库使古籍研究插上了腾飞的翅膀。利用计算机实现古籍自动校勘、标点、翻译、注释、编纂,才能真正被称做是古籍整理的智能化。

古籍智能处理技术与古籍数字化技术有着明显的差异。所谓古籍数字化,“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”^{[5]2}。古籍智能化则是在古籍实现了文本字符的数字化,具有基于超链接的浏览阅读环境、强大的检索、研究支持功能外,建立古籍整理的专家系统,实现古籍版本的自动校勘、自动编纂、自动断句标点、自动分词标引、自动注释、自动翻译为白话等。当然,古籍数字化是古籍智能化处理的基础,智能信息处理方法的发展是古籍智能化的技术保障,古籍智能化是古籍数字化的目标之一。

早期古籍研究受传统训诂方法影响较深,以及研究工具的有限,语法分析以意义为主,忽视了语法形式、语法功能的分析应以“举例式”论说为主。一方面随着西方的语言学理论,如索绪尔的符号系统学说、布龙菲尔德的结构主义、乔姆斯基的转换生成语法等的影响;另一方面智能处理技术在古籍中广泛应用,促进了古籍研究方法的变革。

中国古籍与现代数学,这两个无论从形态、本质到文化都相距极大的事物,新加坡的华裔学者林大芽却令人匪夷所思地将二者结合起来。这是一种运用辩证逻辑与数理逻辑的思维方式,从一种新的角度对中国古籍进行开创性探索,科学揭示优秀古籍的文学艺术美。此种方法既依赖于古籍智能处理技术的发展,又进一步丰富了古籍智能处理技术。^[6]

关于古籍整理技术的发展情况,刘琳、吴洪泽在其专著《古籍整理学》中列有专门章节加以叙述,不仅论及了古籍整理所用的各种技术,同时亦介绍了各种支撑系统。^{[1]330-345}与刘、吴二氏集中介绍相关技术的做法不同,黄永年在《古籍整理概论》在索引编制技术中介绍了计算机技术的应用。^{[2]156-162}

除上述两本专著外,对古籍整理技术有所论述尚有《古籍整理出版十讲》(岳麓书社,2002年版)、《古籍整理浅谈》(燕山出版社,2001年版)等。其中相关内容所占篇幅虽不及刘、吴二氏,但文中叙述亦多有可取之处。

至于论文则更多,在CNKI中,仅题名中带有“计算机”和“古籍”的论文就



有 31 篇^[7]，其中最早的论文为 1987 年施德庆撰写的《陕西省开发中医古籍整理工作计算机系统》^[8]。1988 年曹书杰提出计算机可能应用于古籍整理的具体领域，他把整理研究分成标点、校勘、注释、训诂、辑佚、今译、汇编等 7 种类型，并归纳为 3 种模式：一是内容添加式的，包括标点注释、训诂、校勘中的勘误；二是内容结构式的，包括辑佚、汇编、校勘中的校异；三是对应转换式的，即今译。根据当时的计算机功能与水平，曹书杰认为后两种模式，在人工配合和语言程序、软件知识库的支持下，计算机可以完成。而第一种模式中的各种方式，则存在一定难度。^[9]在其后相当一段时间内，计算机应用于古籍整理领域，基本沿用这样的思路，取得了不少成就。

新中国成立后，党和政府就已经注意到古籍整理这一项工作的开展。1958 年国务院科学规划委员会成立了古籍整理出版规划小组，主持制定了《三至八年整理和出版古籍的重点规划（1960 年—1967 年）》（草案），从此古籍整理出版工作有了全面的安排和统一的部署，计划性、目的性大大增强。内容是多方面的，既有人才的培养，也有大型古籍整理工程的实施，更有现代网络技术、计算机技术的大量引进与应用。参加古籍整理的机构众多，既有高校，也有图书馆，更有专门的古籍整理研究机构。整理古籍的语种，不仅有汉语，也同样包括藏文、回文等少数民族语言在内。

1992 年国家古籍整理出版规划小组制定《中国古籍整理出版十年规划和“八五”计划（1991 年—1995 年—2000 年）》，明确提出要“加快古籍整理出版手段现代化的步伐”，并认为“当前基本典籍经过整理，输入微机等工作应扩大试点……以推进古籍整理与出版现代化”。^[10]这是我国官方有关古籍整理出版手段现代化的最早记载。

2001 年颁布的《国家古籍整理出版“十五”重点规划（2001 年—2005 年）》“对于古籍整理出版的电子化工程给予了特别的关注”，并“提倡出版古籍的电子版，并大力促进古籍信息在网络上流通。”^[11]这一系列措施的出台，使电子化、网络化成为古籍整理的助推器。

作为国家古籍整理与出版的领导机构，制定的规划具有重要的指导意义，古籍整理与出版取得的成果已经得到了社会广泛的认可，也使计算机技术在古籍整理领域更加普及。

1987 年施德庆提出开发中医古籍整理工作计算机系统^[8]，1991 年彭立提出构建“微机古籍辅助整理系统”，^[12]1995 年姚松认为“古籍整理专家系统”应该可以实现不同古籍版本的自动校勘、自动查重查错、自动为白文断句加标点、自动注释、自动翻译成白话、自动生成输出格式等功能^[13]。于亭提出应构建“古籍整理辅助研究