
中国文化典籍计算机整理与开发技术研究系列
丛书主编◇侯汉清

NONGYE LISHI WENXIAN
SHUZIHUA JIANSHE YANJIU

# 农业历史文献数字化建设研究

曹 玲 薛春香◎著

安徽师范大学出版社

国家出版基金项目

中国文化典籍计算机整理与开发技术研究系列
丛书主编◇侯汉清

NONGYE LISHI WENXIAN
SHUZIHUA JIANSHE YANJIU

# 农业历史文献数字化建设研究

曹　玲　薛春香◎著

安徽师范大学出版社

责任编辑：房国贵　　责任校对：潘　安
装帧设计：丁奕奕　　责任印制：郭行洲

**图书在版编目（CIP）数据**

农业历史文献数字化建设研究/曹玲，薛春香著．—芜湖：安徽师范大学出版社，2013.11
（中国文化典籍计算机整理与开发技术研究系列/侯汉清主编）
ISBN 978-7-5676-0997-6
Ⅰ.①农… Ⅱ.①曹… ②薛… Ⅲ.①农业技术—古籍—数字化—研究 Ⅳ.①G257.39

中国版本图书馆 CIP 数据核字（2013）第 238885 号



农业历史文献数字化建设研究

曹　玲　薛春香　著

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）　E-mail：asdcbsfxb@126.com
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：2013 年 11 月第 1 版
印　　次：2013 年 11 月第 1 次印刷
规　　格：700×1000　1/16
印　　张：15
字　　数：205 千
书　　号：ISBN 978-7-5676-0997-6
定　　价：36.00 元

# 出版说明

## 一

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，是中华儿女应尽的义务和职责。

我国古籍资源数字化工作起步于20世纪80年代初期，经过几十年的发展，已取得令人瞩目的成就。第一批《国家珍贵古籍名录》和全国古籍重点保护单位的申报工作早已完成，制定古籍数字化标准列入议程，古籍整理与保护工作进入一个新的历史阶段。

古籍资源数字化最初主要是制作书目数据库，后来发展到古籍全文数据库，直至如今的网络检索系统。信息技术的发展和数字化成果的不断涌现，对古籍数字化提出了更高的要求。专家认为，数字化的古籍资源除了实现文本字符的数字化、具有基于超链接的浏览阅读环境和强大的检索功能外，还需具有“研究支持功能”。所谓“研究支持功能”，是指能够提供有关古籍内容本身科学、准确的统计与计量信息，提供与古籍内容相关的参考数据、辅助工具。这些信息、数据或工具都是古籍内容的增值或补充。北京大学计算语言研究所和古文献研究所合作开发了“古诗研究计算机支持系

统”，并取得了阶段性成果。

时值古籍数字化研究日新月异、如火如荼之际，安徽师范大学出版社于2011年精心策划、2012年成功申报、2013年落实出版国家出版基金项目“中国文化典籍计算机整理与开发技术研究”（编号：2013G2－011），在数字化古籍诸项功能特别是“研究支持功能”上给予探索。

改革开放30多年来，国泰民安、政通人和，中国传统文化日益受到政府重视，有关科研机构加大了对古籍整理研究的力度。安徽师范大学出版社能够有机会申请到国家出版基金项目的资助，本项目丛书能顺利进行，实在与国家关注出版事业、关注中国传统文化、关注文化典籍计算机整理工作密切相关。

## 二

“中国文化典籍计算机整理与开发技术研究”项目主要内容如下：

第一，探索与试验古籍知识库、模式库，将之改造为规则库。

本项目利用命名实体识别、词汇同义词关系的识别、文本主题概念的提取等技术，从各类古籍数据库抽取人名、地名、文献名、职官名、物品名、年号等；并将人名表、地名表、书名表、年代年号表等，与引书模式、异名别称模式、断句模式、分类模式等模式库整合成一个古籍整理与开发专用的知识库，以方便中文古籍整理与开发。

本项目构建的各类知识库，具体有：古代官名、人名和地名表；避讳字、异体字和繁简字对照表；常用古籍名称库；专业术语词典，按专业分为历史、天文、农业、医学、宗教等多个专业词典；主题术语词典，按主题分为动物、植物、矿物等若干主题词

典；古代关联词语表，用语义相似度计算和基于词典释义的同义词识别算法，开发古代关联词语表；禁用词典。

本项目构建的各类模式库有：异名别称模式库，包括别称词、避忌特称、地域特称、文献特称等；断句标点模式库，包括句法特征词法、同义语标志词法、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等多种模式识别库；古籍分词模式库。大多数古籍文本无标点，分词的长度及方法需要单独构建。

这些知识库与模式库，采用拿来主义，并经过计算机检验与筛选，最终形成适用于计算机处理古籍的规则，合成为一个综合的规则库，从而为计算机处理古籍提供有力的规则支撑。

第二，重点探索与试验下列古籍智能整理与开发的关键技术。

自动校勘技术：采用对校法，借鉴中文文本自动校对和模式匹配技术，通过比对程序校勘古籍。

自动断句标点：对现有部分标点本古籍进行数理统计，归纳、总结其断句和标点模式。同时结合语言学方法，进一步优化断句和标点模式，从而实现计算机辅助断句与标点。

自动分词和标引：利用汉语现代文本的分词理论和方法，探索古籍文本的自动分词技术，并利用统计学方法（N-gram 等），从古籍数据库中筛选出有一定表达意义的实词词汇。同时利用异名别称模式，创建并完善古籍用词同义词典。在此基础上，引入文本数据挖掘、主题提取和自动分类技术，探索基于知识库的古籍文本的自动标引与分类。

自动编纂：让计算机模拟人脑从大量古籍文本中判断、选择出与编纂主题相关的资料，实现古籍专题资料的自动编纂工作。

自动注释：收集已有古籍专业词汇及其注解，构建古籍语词注解知识库。

第三，在上述基础上，将它们整合为计算机整理与开发古籍的

“一条龙”服务，即构建出古籍整理与开发的专家系统或智能处理系统。

将以上各种词汇、知识、模式整合起来，构建成一个内容丰富、功能多样的古籍规则库，再与自动校勘、自动断句标点、自动分词标引、自动编纂、自动注释等各项技术结合，从而实现文化典籍整理与开发的“一条龙”服务，提出并设计一种集成各种古籍整理与开发智能技术的原型系统。该系统集知识与模式于一身，集规则与技术于一体，具有合成性，既适用于古籍数据库的建设，又适用于古籍数据库的开发使用。

第四，在上述基础上，本研究进行四项个案研究，在实践中探索上述集成的古籍整理与开发智能技术原型系统的可行性与应用性。

农业历史文献数字化：构建农史文献资源库，对农史文献进行自动标引和自动分类，提供农史文献的浏览与检索服务。

建立农史文献门户：构建农史门户网页智能搜索引擎和农史网页自动标引与自动分类实验系统，构建农史门户实验网站。

探索民国农业文献自动索引：在民国农业文献数字化整理中的具体应用，研究索引自动编纂、电子图书编纂、电子索引编纂、数据库建设和主题网关构建等技术方法。

地方志中农业资料的挖掘：从《方志物产·广东》中选取比较实用的全文数据库、物产索引、引书索引、物产分析和引书分析等几个方面进行研究。

总之，本项目充分利用目前在现代汉语文本已经取得成功的中文信息处理技术成果，并根据此成果中的模式识别技术、聚类技术、信息自动提取、信息检索及其他自然语言处理技术等，对照现已建成的大量数字化文化典籍数据库，归纳并修订各类知识库与模式库，研究古籍的自动校勘、自动断句标点、自动分词标引、自动

编纂、自动注释等技术，合成古籍整理与开发的专家系统或智能处理系统，从而为大规模建设新的更多古籍数据库作准备。

## 三

本项目成果的推广和运用，不但对于探索数字时代古籍文本自然语言处理的理论和方法具有一定意义，而且对推动古籍整理和研究的自动化和智能化、促进我国文化典籍资源的建设和开发以及弘扬传统文化等方面，均具有重大的现实意义和很高的应用价值，可以为继承与发扬中华古籍文化、为建设中国特色社会主义文化服务。

本项目丛书主编由南京农业大学信息科技学院博士生导师侯汉清教授担任。侯先生是中国古籍整理专业第一个硕士研究生，早年在北京大学任教，现执教于南京农业大学，系中国古籍整理专家、中国索引学会副理事长。中图分类法就是侯先生主创起来的。2008年，侯先生主持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），本套丛书即此项目的纸质成果。

本丛书分为六册，各册的内容及其撰写者简要介绍如下：

《古籍计算机自动断句标点与自动分词标引研究》，侧重于自动断句标点、自动分词标引研究，兼顾古籍计算机整理与开发系统的构建与集成。作者黄建年，博士，研究馆员，现就职于南京财经大学。

《古籍计算机自动校勘、自动编纂与自动注释研究》，侧重于自动校勘、自动编纂与自动注释研究，兼顾古籍计算机整理与开发系统的构建与集成。作者常娥，博士，现就职于东南大学，硕士生导师。

《古籍计算机自动索引研究——以民国农业文献自动索引为例》，侧重于自动索引研究，并以民国农业文献自动索引为样本。作者王雅戈，博士、博士后，中国索引学会理事，现就职于常熟理工学院。

《古籍计算机全文数据库及内容挖掘研究——以〈方志物产·广东〉为例》，侧重于数据库内容挖掘研究，并以《方志物产·广东》之物产、引书等内容挖掘研究为样本。作者衡中青，博士，中国索引学会理事，现就职于佛山科学技术学院。

《古籍计算机信息门户自动构建与应用——以农史学科为例》，侧重于信息门户自动构建与应用，并以农史学科信息门户构建与应用为样本。作者刘竟，博士，现就职于江苏大学。

《农业历史文献数字化建设研究》，侧重于农史文献数字化实践——中国农业遗产信息平台建设，并介绍其实际应用。作者曹玲、薛春香，均为博士，分别就职于南京信息工程大学、南京理工大学。

本项目丛书的出版发行，可为正在有志于从事本领域研究和工作的人员提供一个可资借鉴的文本。我们期待本丛书能为中国从文化古国向文化大国、文化强国迈进尽绵薄之力。

# 目　录

# 1 引 言

我国农业历史源远流长，在上万年的农业历史长河中，劳动人民在长期的农业生产实践中积累了丰富的经验和教训，留给我们后人宝贵而丰富的农业遗产。今天研究我国的农业历史，一方面为了认识我国丰富的农业遗产，继承和发扬我国农业的优良传统，另一方面为了鉴古知今，即为目前的“三农”建设和整个社会经济发展提供决策支持，走出一条具有中国特色、符合中国国情的农业现代化道路。

## 1.1 农史文献概述

农业史（下文简称“农史”）是一个横跨多种学科的专题研究领域，是一个农业科学与历史科学、自然科学与社会科学交叉的学科，内容涉及农业科学、经济、历史、地理、政治、资源、生态、工程等多个相关学科。国内学界对于农业史的界定是：农史是以政治、经济、文化等综合的观点研究农业生产与技术、农业经济、农村社会、农业思想历史演进及其规律的一门交叉性学科，运用自然科学与社会科学相互交叉、农业科学与历史学相互结合的方法来探讨农业产生和发展的动因、动力、影响及规律[1]。

文献是记录有知识的一切载体，是人类知识和经验的记录。所谓农史文献，既包括历朝历代流传下来的古农书，其中有善本、珍

本、孤本、拓本、稿本等，也包括对这些农书校释、校评、译注、今释、标点而产生的新版本，还包括大量的农史论文以及史话、普及读物等[2]。农史文献随着农业的发生发展而产生，为农史研究提供依据，并随着农史研究的深入不断累积。

史料的采集与挖掘历来是史学研究的前提与基础，通过对古代典籍和考古资料的研究，很多农史研究学者撰写了大量的专著和论文。《中国农业古籍目录》[3]一书共收录农书3 705种。据中国农业博物馆资料室编辑的《中国农史论文目录索引》[4]统计，仅从1878年到1991年的农史研究论文就有19 255篇。自20世纪80年代以来，几种农史专业期刊（如《农史研究》《中国农史》《农业考古》《古今农业》等）上就刊载了14 943篇论文（统计来源为中国知网“中国期刊全文据库”，以期刊名为检索入口，检索年限范围为“1980—2012”）。另外，综合性期刊、相关学术期刊、会议论文集及报纸也刊载了大量的农史研究文章，文献总量可能有40 000篇。

数以万计的农史文献是今人研究我国农业历史丰硕成果的总结与展示，是研究我国农业历史的又一笔崭新的信息资源。这些信息资源内容丰富，涉及中国农业科技、农业经济、农村社会等诸多领域，资料翔实而利用价值极高。但这些资源多为纸质文献，收藏地点分散，保存、查找、获取、利用均极为不便。有鉴于此，对农史文献信息资源进行有效的整理和组织十分必要。

## 1.2 农史文献数字化建设的必要性

信息资源数字化建设包括两个方面的含义：其一是对原有传统文献信息资源进行数字化加工，即把纸质文献信息转化为用计算机存储设备中电磁、光电信号存储的信息，从而使文献信息的载体由过去单一的纸质载体向多种载体形态发展；其二是开发建设新的数

字化信息资源，直接以电子出版物或网页的形式出版或发布的文献信息，并对二者进行科学的整合、重组、分类、组织，形成新的数字化资源体系[5]。

数字化是信息时代的潮流和趋势，它极大地影响和改变了人们学习、工作、生活的方式和理念。我们从事农史研究，也不能再如从前一般钻在故纸堆中皓首穷经，把大部分的宝贵光阴耗费在资料的查检和抄写上，而要充分利用现代信息技术带来的资料获取的快捷便利，把自己从艰苦而繁琐的爬梳、翻检工作中解放出来，并投入创造性学术研究之中。农史文献数字化的必要性，表现在以下三方面：

第一，农史文献信息资源分布散乱，使用不便。农史文献信息资源形式多样，数量庞大，长期以来分散在多个农史研究机构、公共图书馆、高校图书馆、档案馆、博物馆中，还有部分古农书收藏在美国、日本、韩国等国家。这些机构各有所藏，但都收藏不全。这种收藏上的分散，给研究人员查阅资料带来了许多不便。甚至有些机构因为缺乏专人管理，导致很多重要的信息资源尘灰覆身，被束之高阁，而研究人员却遍寻难得。

第二，农史文献保存和使用的客观要求。农史文献资源绝大多数为纸质文献，特别是农业古籍，不乏众多“珍善本”资源。随着时光荏苒，它越“古老”就越珍贵，加上纸张的老化、残损问题严重，已经不起多次翻阅。因此，如何有效地保护好这些珍贵的文献，又不影响正常使用，就成了各文献收藏机构迫切需要解决的问题。而数字化正是解决文献信息机构信息资源“藏”与“用”这对矛盾最有效的方式。

第三，数字化对农史研究的影响。数字化能“原汁原味”地反映纸本文献、实物资料的内容，有利于实现农史文献资源的共享，能让更多的人坐在家中通过网络就能快捷地获取所需的信息

资源。

数字化是研究人员利用现代化信息技术辅助挖掘农业历史文献信息资源中所蕴藏的知识的前提。计算机在资料的查检、统计、汇集方面远远甚于人力。例如，查检古农书中所有出现“稻”的条目，人力花费数月都未必能全部找到，但是通过计算机不用1秒钟就可以返回古农书中所有“稻”的条目。此外，通过数据挖掘、知识发现工具，还能够从海量信息中发现很多鲜为人知或无人知晓的规律或知识。这一切都将大大促进农史研究工作的顺利开展。可以说，农史文献数字化是农史研究工作现代化的前提和重要保障。

## 1.3 农史文献数字化建设的重要意义

农史文献是进行农史研究的重要基础。依托先进的数字化技术和网络化手段，进行综合性的农史文献数字化建设工作具有重要意义，这主要表现在以下三个方面：

第一，有利于保护和保存优秀的文化遗产。农史文献的主体为农业古籍，由于年代久远，纸张变得脆弱易碎，许多收藏单位为了保护古籍，采取了严格限制阅览、流通的手段，这与图书馆“藏以致用，读者至上”的原则是相违背的。在科技迅猛发展的信息时代，用计算机存取的方法将农业古籍这块瑰宝存贮于易传输、易检索、易复制、可永久保存的现代化管理系统中，有利于文献的传承和保护。

第二，宣传农史文化，推动农史研究。近年来，农史研究得到很大发展，农史文献量增加速度也相当快，仅期刊论文数量就有几万篇，但由于研究成果散见于各地或各类报刊上，不便于研究者了解。在这种情况下，建设农史文献数据库和网站将会在农史研究和