古籍文本数据格式比较研究

肖 禹 | 著

上海远东出版社

古籍文本数据格式比较研究

肖禹 | 著

上海远东出版社

图书在版编目（CIP）数据

古籍文本数据格式比较研究 / 肖禹著 .– 上海：上海远东出版社，
2016
ISBN 978–7–5476–1244–6

Ⅰ . ①古… Ⅱ . ①肖… Ⅲ . ①古籍 – 数字化 – 研究 – 中国
Ⅳ . ① G256–39

中国版本图书馆 CIP 数据核字（2016）第 319754 号

国家科技支撑计划“中国地方志数字化关键技术研究与演示平台设计”项目（2015BAK07B00）
“地方志资源调查与数字化加工规范研究”课题（2015BAK07B01）研究成果

古籍文本数据格式比较研究
著者　肖　禹
责任编辑 / 徐忠良　杨林成　装帧设计 / 李　廉

出版：上海世纪出版股份有限公司远东出版社
地址：中国上海市钦州南路 81 号
邮编：200235
网址：www.ydbook.com
发行：新华书店　上海远东出版社
　　上海世纪出版股份有限公司发行中心
制版：南京前锦排版服务有限公司
印刷：昆山亭林印刷有限责任公司
装订：昆山亭林印刷有限责任公司

开本：787 × 1092　1/16　印张：32.75　字数：677 千字
2017 年 4 月第 1 版　2017 年 4 月第 1 次印刷

ISBN 978–7–5476–1244–6 / G · 790
定价：138.00 元

出版说明

国有史，地有志，家有谱，家谱、方志、正史从不同层面构成中华民族历史的记忆。

中国自古就有修志的传统。《周礼·春官》载："外史掌四方之志。"东汉郑玄注："方志，四方所识久远之事。"中国地方志作为珍贵的文献资源，其内容不仅包括各地区的疆域、气候、山川、物产等地理资料，也涵盖户口、人物、赋税、艺文等人文历史各方面的记载，是地方的百科全书，一地之全史。地方志所详细记载本地区的政治、经济、社会等发展状况，形成了独特的区域文化，具有鲜明的地方特征；地方志以记述某一段时间当地的情况为主，是一个特定时期文化积淀和历史的产物，反映出了特定时代的经济、政治、文化等方面的烙印；地方志内容广泛，系统性强，从天文地理、名胜古迹、物产资源、民族宗教、方言俗语、金石碑刻到政治经济、科学文化、典章制度、著名人物、重大事件等，分门别类按照内容的要求选择合理的记录方式；资料性是地方志所有特征中最基础的一个特征，是方志生命力之所在。

据不完全统计，汉文古籍超过 20 万种，地方志约占 5%，地方志同时具备的地域性、时代性、系统性、资料性和科学性，既包含丰富的内容信息，又适合与现代技术相结合，建立资源库、知识库和 GIS 系统，进而构建中国传统文化基础平台。以地方志为核心的中国传统文化基础平台将地方志目录、图像、文本、关联数据等不同粒度的数据与地理信息数据相结合，实现时间、空间、文献三个维度的智能检索、数据分析和图形化显示。同时，平台具有高度的容纳性与扩展性，可将各种类型的文献资源、各种格式的数字资源和各种功能的知识工具有机地整合在一起。中国国家图书馆古籍馆陈红彦馆长和肖禹等专家在地方志数字化工作实践中不断积累，研究古籍数字化中遇到的技术问题，进行理性总结。科技部科技支撑计划"中国地方志数字化关键技术研

究与演示平台设计”正是基于地方志这样的特征，希望通过地方志数字化技术、数据抽取技术、可视化技术的统合应用，为古籍数字资源建设利用做出有益的尝试。

实现现代技术与传统文献的紧密结合，打造基础平台，支持数据分析与智能检索，必须以统一的标准规范为先导，因此项目中设计了实现平台相关功能必需的理论研究、加工规范制定等内容，最终以《古籍文本数据格式比较研究》《IDS 与集外字处理方法研究》《国家图书馆藏清康熙时期纂修方志书录》《方志文献特性与数据抽取研究》《地方志数字化加工规范汇编》《地方志数字化加工规范应用指南》六部书的形式呈现。

上海远东出版社

2017 年 2 月

目　录

第一章 绪 论

关于古籍的概念，目前业界比较通行的观点有三种：古籍（ancient books）指书写或印刷于1912年以前具有中国古典装帧形式的书籍，中国古代书籍的简称[①]；古籍（ancient Chinese books）主要指1911年以前（含1911年）在中国书写或印刷的书籍[②]；古籍（Pre-1912 Chinese books），中国古代书籍的简称，主要指书写、印制于1912年以前又具有中国古典装帧形式的书籍[③]。在信息技术中，文本是人类可以识别的词的序列，而且它具有能够被机器可读的形式，如ASCII；文本与bitmaps和程序代码的形式的编码不同，它们通常是“二进制”形式的[④]。古籍文本数据是古籍数字资源类型之一，具有不可替代的作用。

一、引言

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[⑤]。

（一）古籍数字化

葛怀东等在《国内古籍数字化研究文献计量分析》中以CNKI为数据源，通过人

① GB/T 21712-2008, 古籍修复技术规范与质量要求 [S]. 北京：中国标准出版社，2008：1.
② GB/T 3792.7-2008, 古籍著录规则 [S]. 北京：中国标准出版社，2008：1.
③ GB/T 31076.1-2014, 汉文古籍特藏藏品定级 第1部分：古籍 [S]. 北京：中国标准出版社，2015：1.
④ 李进良，倪健中. 信息网络辞典 [M]. 北京：东方出版社，2001：281.
⑤ 陈志伟，盖阔. 中文古籍全文数据库指要 [J]. 图书馆学研究，2014（14）：39—43.

工筛选去除一稿多发、简讯、报纸、评论、通知以及与古籍数字化研究相关性不大的文章，得到1985年至2012年之间的古籍数字化研究论文797篇；通过计量分析可以发现国内对古籍数字化领域的研究取得了长足的进步，已经开始进入高速发展的阶段，新兴信息技术的发展以及古籍数字化领域重要事件的推动这两大因素，对该领域的发展有着非常重要的意义，该领域的文献开始呈现接近指数增长，核心期刊群、核心研究机构和核心作者正在慢慢形成中[①]。

王荟等在《汉语文古籍全文文本化研究》[②]中以CNKI为数据源，分析1985年至2010年之间古籍数字化论文的研究内容，其中有近27%的内容是对古籍数字化的思考，探讨古籍数字化的理论问题和数字化实践中的共性问题，如李国新的《中国古籍资源数字化的进展与任务》[③]、陈力的《中文古籍数字化方法之检讨》[④]、毛建军的《古籍数字化的概念与内涵》[⑤]等；有近9%的内容是介绍现有的古籍资源和数字化项目，如李明杰的《古籍网络资源述略》[⑥]、许红健的《台湾中文古籍数字化成果特色谈》[⑦]、毛建军的《欧美地区中文古籍数字化概述》[⑧]等；有近8%的内容是讨论古籍数字化的现状与对策，从地区、图书馆、档案馆的古籍数字化实践出发，分析当前存在的问题，提出解决问题的思路，如厉莉的《古籍数字化的现状及对策》[⑨]、王立清的《港台地区古籍数字化现状分析及启示》[⑩]、刘春金等的《中文古籍数字化现状分析》[⑪]等；有近18%的内容是研究古籍数字化的技术与方法，如王燕的《浅谈数字图书馆中的图形化查询技术——GIS在北京大学古文献资源库中的应用》[⑫]、林颖等的《一种灵活可扩展的古籍数字对象的设计与实现》[⑬]、赵云的《图书馆古籍数字化前处理工作研究》[⑭]等；此外，还有一些论文探讨了古籍数字化与学术研究、古籍书目数据库、中医古籍数字化、少数民族古籍数字化、农业古籍数字化等问题。如表1-1所示：

---

① 葛怀东，盖阔. 中文古籍全文数据库指要[J]. 图书馆学研究，2014（14）：39—43.
② 王荟，肖禹. 汉语文古籍全文文本化研究[M]. 北京：国家图书馆出版社，2012.
③ 李国新. 中国古籍资源数字化的进展与任务[J]. 大学图书馆学报，2002（1）：21—26.
④ 陈力. 中文古籍数字化方法之检讨[J]. 国家图书馆学刊，2005（3）：11—16.
⑤ 毛建军. 古籍数字化的概念与内涵[J]. 图书馆理论与实践，2007（4）：82—84.
⑥ 李明杰. 古籍网络资源述略[J]. 图书馆建设，2002（3）：84—86.
⑦ 许红健. 台湾中文古籍数字化成果特色谈[J]. 农业图书情报学刊，2009（1）：130—133.
⑧ 毛建军. 欧美地区中文古籍数字化概述[J]. 数字与缩微影像，2008（1）：36—38.
⑨ 厉莉. 古籍数字化的现状及对策[J]. 江西图书馆学刊，2002（1）：57—58.
⑩ 王立清. 港台地区古籍数字化现状分析及启示[J]. 图书情报工作，2006（8）：87—90.
⑪ 刘春金等. 中文古籍数字化现状分析[J]. 江西图书馆学刊，2008（2）：112—113.
⑫ 王燕. 浅谈数字图书馆中的图形化查询技术——GIS在北京大学古文献资源库中的应用[J]. 大学图书馆学报，2006（1）：58—62.
⑬ 林颖，程佳羽. 一种灵活可扩展的古籍数字对象的设计与实现[J]. 图书馆杂志，2014（12）：56—60.
⑭ 赵云. 图书馆古籍数字化前处理工作研究[J]. 农业图书情报学刊，2007（3）：121—123.