



地方志文献特性 与数据抽取研究

张毅 | 著

 上海远东出版社



《说文解字》与 《说文解字》研究

《说文解字》

《说文解字》



地方志文献特性 与数据抽取研究

张毅 | 著

上海遠東出版社

图书在版编目 (CIP) 数据

地方志文献特性与数据抽取研究 / 张毅著 . —上海 :

上海远东出版社, 2017

ISBN 978-7-5476-1313-9

I . ①地… II . ①张… III . ①地方志—古籍整理—数据处理—研究—中国 IV . ① K290-39

中国版本图书馆 CIP 数据核字 (2017) 第 262192 号

国家科技支撑计划“中国地方志数字化关键技术与演示平台设计”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

地方志文献特性与数据抽取研究

著者 张 毅

责任编辑 / 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: 11.25 字数: 200 千字

2018 年 1 月第 1 版 2018 年 1 月第 1 次印刷

ISBN 978-7-5476-1313-9/G · 829

定价: 45.00 元

版权所有 盗版必究 (举报电话: 62347733)

如发生质量问题, 读者可向工厂调换。

零售、邮购电话: 021-62347733-8538

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》《地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

出版说明.....	1
第 1 章 绪 论.....	1
1.1 地方志.....	2
1.2 地方志数字化	7
1.3 研究对象和思路.....	12
第 2 章 地方志的内容特征.....	13
2.1 基本特征	13
2.2 时代特征	19
2.3 类型特征	23
第 3 章 地方志的结构特征.....	29
3.1 平列体.....	30
3.2 纲目体.....	36
3.3 纪传体.....	39
3.4 三宝体.....	41
3.5 章节体.....	43
3.6 其他.....	45
第 4 章 地方志的记述方式与版式特征.....	48
4.1 序	48
4.2 凡例	51

4.3 修纂者信息	52
4.4 目录	55
4.5 图	58
4.6 表	61
4.7 志	66
4.8 传	72
4.9 文	73
4.10 其他	75
第5章 地方志门目整理	77
5.1 门目数据的来源	79
5.2 门目数据的识别与提取	84
5.3 门目数据的整理	95
第6章 地方志数据抽取	108
6.1 地方志数据抽取的意义与局限	109
6.2 碎片化与本体建设	112
6.3 地方志数据抽取示例	137
附 录	157
1. 明永乐十年颁降凡例	157
2. 永乐十六年颁降纂修志书凡例	159
3. 乡土志例目	161
4. 康熙二十九年河南巡抚颁布修志凡例	164
5. 旧志资料类编目录索引	165
参考文献	167

第1章 绪 论

地方志，又称方志、志乘，“是记载一定地区（或行政区划）自然和社会各个方面的历史与现状的综合性著述”。^①其历史悠久，起源甚早，被称“最古之史”^②，有学者将之追溯至先秦，有学者认为始于两汉，还有的断于唐宋。但即使是最晚的唐宋，距今也有千余年的历史。更可贵的是，在上千年的历史长河中，虽然经历了无数的灾害、战火以及多次改朝换代，地方志仍在被不间断地编修，一直延续发展至今。为继承和发扬这一中华民族的优秀文化传统，全面、客观、系统地编纂新方志，科学、合理地开发利用地方志资源，充分发挥地方志在经济社会发展中的作用，2006年5月，国务院公布了《地方志工作条例》。条例着眼于当下的地方志事业，将地方志析为地方志书与地方综合年鉴，并定义了地方志书的内涵：全面系统地记述本行政区域自然、政治、经济、文化和社会的历史与现状的资料性文献。要求地方志书每20年左右编修一次。

面对自古及今、卷帙浩繁、凝聚了无数人心血的珍贵方志文献，如何更妥善地保存、管理，便捷地揭示其知识脉络，深入挖掘其价值内涵，使之在当今的文化建设、国民教育、学术研究等领域发挥出其应有的作用，是摆在我们眼前的重要课题。

随着社会的发展，科技的进步，互联网和新媒体广泛应用于各领域，地方志的数字化、信息化日成趋势，成为当下解决藏与用、编辑与出版、开发与利用等问题的最行之有效的路径。地方志数字化、信息化的方法也从最初的图像、全文、图文对照，扩展延伸至如今的数据挖掘、知识抽取、自动编纂等。而在进行方志资源的深度开发与数字化时，首要的任务便是对地方志文献特性进行深入研究。可以说，没有对文献特性的正确理解和把握，理论和技术便只能停留在设想阶段。本书的出发点即是基于此，希冀通过对地方志文献特性进行系统、全面的梳理，推动地方志数字化进展，促使数据抽取等技术能够更好地服务于地方志的研究和利用。

^① 来新夏主编：《方志学概论》，第1页。

^② 梁启超：《中国近三百年学术史》，第333页。

由于我国现存地方志较多,不可胜数,且颇具个性差异,而笔者所见又十分有限,因此本书对于地方志文献特性的归纳或多有疏漏与差误,谨以此抛砖引玉,期待能够促进相关的探索与研究。

1.1 地方志

1.1.1 概述

地方志是我国灿烂的文化遗产。其涵盖内容丰富,包罗万象,无所不载,堪称地方的百科全书,举凡一地的建置、沿革、疆域、山川、名胜、物产、气候、人物、文化、教育、民族、风俗、事件等,无所不包,所以又被誉为一方之全史。北宋著名史学家司马光在《河南志序》中称:“《周官》有职方、土训、诵训之职,掌道四方九州之事物,以诏王知其利害,后世学者,为书以述地理,亦其遗法也。唐丽正殿直学士韦述为《两京记》,近故龙图阁直学士宋君敏求字次道,演之为《河南》《长安》志。凡其兴废、迁徙及宫室、城郭、坊市、第舍、县镇、乡里、山川、津梁、亭驿、庙寺、陵墓之名数,与古先之遗迹,人物之俊秀,守令之良能,花卉之殊尤,无不备载。考诸书记,其详不啻十余倍,开编粲然,如指诸掌,真博物之书也。”^①指出方志在演进过程中记述范围越来越广,已成“博物之书”。

地方志数量众多。一是历代编修的数量多。宋代即已达到了“郡县必有志”的程度,据《宋朝方志考》考辨统计,“两宋路、州(府、军、监)、镇(乡)志诸类志书合计存佚达一千零三十一一种之多”。^②至明代,方志修纂的数量更是可观,约有3470余种。^③二是现今存世的数量多。耗时七年时间之久,成书于20世纪80年代的《中国地方志联合目录》著录了“全国三十个省、市、自治区的一百九十个公共、科研、大专院校图书馆、博物馆、文史馆、档案馆等”度藏的1949年以前编纂的地方志,共收录“历代地方志八千二百余种”,包括通志、府志、州志、厅志、县志、乡土志、里镇志、卫志、所志、关志、岛屿志等。总志及山、水、寺庙、名胜等专志未计入内。成书于20世纪90年代、收录范围相同的《中国地方志总目提要》将这一数字更新至八千五百七十七种。随着古籍普查的不断深入,现存旧方志的数量还在攀升,据统计已经“超过1万种”^④。

地方志品类多样。就综合志而言,有记述全国范围的总志、一统志,有记述一省范围的通志,有记述一府、厅、州、县范围的府、厅、州、县志,有记述一个城市的

① 王晓岩:《分类选注历代名人论方志》,第43、44页。

② 顾宏义:《宋朝方志考》,第4页。

③ 巴兆祥:《论明代方志的数量与修志制度——兼答张升〈明代地方志质疑〉》,《中国地方志》2004年第4期。

④ 陈红彦:《关于进一步加强古旧方志资源利用的思考》,《图书馆工作与研究》2013年第12期。

都邑志，有记述一乡一镇范围的乡镇志，有记述一里一巷范围的里巷志，还有记述特殊区域范围的边关志、卫所志、盐井志等。此外，也有将两省或数县合并于一起记述的合志。仍据《中国地方志联合目录》统计，现存通志 122 种，府志、直隶州志、直隶厅志 875 种，县志、散州志、散厅志 5728 种，乡镇志 326 种，其他 1213 种。^①

地方志涉及地域广泛，近乎覆盖了我国九百六十万平方公里的广博土地，“大至万州，小逮村镇，禹域之广，几无无志者”^②。明清方志常言“郡邑莫不有志”，迄今，全国各省、市、自治区基本都有相应的方志。据《中国地方志联合目录》统计，北京市有 55 种，上海市 139 种，天津市 26 种，重庆市 49 种，河北省 567 种，山西省 431 种，内蒙古自治区 48 种，辽宁省 130 种，吉林省 93 种，黑龙江省 65 种，陕西省 401 种，甘肃省 198 种，宁夏回族自治区 32 种，青海省 39 种，新疆维吾尔自治区 81 种，山东省 541 种，江苏省 540 种，浙江省 590 种，安徽省 379 种，江西省 477 种，福建省 315 种，台湾省 49 种，河南省 528 种，湖北省 332 种，湖南省 401 种，广东省 395 种，海南省 50 种，广西壮族自治区 221 种，四川省 622 种，贵州省 139 种，云南省 282 种，西藏自治区 44 种。^③ 这些方志中大者到省，细者到乡村里镇。

地方志纵跨古今。一是自宋代方志定型以后，历代皆有方志修纂。据《中国地方志联合目录》统计，现存宋以前方志 3 种，宋代方志 28 种，元代 9 种，明代 942 种，清代 4889 种，民国 1187 种^④。二是方志记载往往自古及今，志书中的部分内容，如人物、沿革等，常有上溯至夏商周三代者。傅振伦先生在论及《北平志》之断限时称：“为史之道，贵于通古今，而尤要在乎侧重现代。”^⑤

地方志中既有古籍又有民国文献，古籍即中国古代书籍的简称，指“书写或印刷于 1912 年以前具有中国古典装帧形式的书籍”^⑥；民国文献“指 1911 年辛亥革命以后至 1949 年 9 月中国出版的中文文献，包括图书、期刊、报纸等出版物和部分手稿”^⑦。不同历史时期、不同地域的刻本、活字本、石印本、铅印本以及抄稿本方志具有较高的历史文物价值，弥足珍贵。

1.1.2 功用

与史上流传下来的很多被誉为精神食粮的珍贵典籍不同，地方志自诞生之初，便

① 刘纬毅等：《中国方志史》，第 22 页。

② 瞿宣颖：《方志考稿甲集》，序。

③ 刘纬毅等：《中国方志史》，第 20-21 页。

④ 刘纬毅等：《中国方志史》，第 20 页。

⑤ 傅振伦：《编辑北平志蠡测》，《傅振伦方志文存》第 43 页。

⑥ 《古籍修复技术规范与质量要求》。

⑦ 郑春汛：《民国文献的价值与保护对策研究》，《图书馆理论与实践》，2008 年第 4 期。

是讲求实用之作。清代方志学家章学诚曾称：“夫修志者，非示观美，将求其实用也。”^①无论是彰显盛世的一统志，还是着力乡土教育的乡土志，无论是地方政府主导下的方志修纂，还是私人撰著的外志，都蕴含、寄托了倡修者的主张与情怀。而修志活动之所以世代传承，绵延不断，其中很重要的原因之一便在于地方志具有资治、存史、教化等实际功用。

古代守土之吏有下车伊始即查阅地方志的传统。如南宋理学家朱熹“官车所莅，必先搜辑志乘”^②；明永乐间滁州知州陈琏上任“尝求志书，久而未得，深以一郡事迹湮没无闻为可慨”^③；清顺治间泗水县知县刘桓“筮仕兹土即搜县志一书”^④；清康熙间王绶奉命知平遥，“至则按图志考其传记，欲以周知一邑置山川、田赋，自上世以迄今兹文物之盛衰，风俗之醇疵”，等等。^⑤通过稽考方志，封疆大吏与地方官员们“不下堂阶”即可对一地之“山川险易、户口登耗、土田荒辟、学校举废，与夫屯粮盈绌、风俗美恶”等情况“了如指掌”。^⑥所以有学者总结方志是“执政者的一种工具书，具有工具性、实用性”^⑦。清末民国时期，美国、日本等国家不遗余力地征集、搜购地方志也是其为实用之书的旁证。

作为特定区域历史发展的最直接记录，地方志中保存了不少正史未曾记录的资料，如地方的庙宇桥梁、乡贤才俊、风土民情、疫病灾荒等等，可补史之漏，参史之误，具有不可替代的文献价值。民国文史名家瞿宣颖曾总结了方志“往往足供”^⑧然不为正史所具备的六类资料，其后，史学家、方志学家李泰棻增加“古迹金石”，扩之为七类：“各地社会制度之隐微递嬗，不见于正史及各史者，往往于方志中见之，其一也；历朝人物，应登正史而未列，或在当日无入正史之资格，而以今日眼光视之，其入靡重者，往往见于方志，其二也；遗文佚事，赖方志以存者甚多，其三也；地方经济状况，如工商各业物价、物产等，其变迁多见于方志中，其四也；建置兴废，可以窥见文化升降之迹，其五也；古迹金石，可以补正史及文字之遗缺者，其六也；氏族之分合，门第之隆衰，可与他史互证，其七也。”^⑨自清末以来，地方志得到国内外历史学、自然科学史、社会学、文学、民族学等多领域研究者的重视，并形成了一门独立的学科——方志学。利用方志进行的学术研究也取得了丰硕的成果。最早且具代表性者如顾炎武利用方志资料编纂出传世之作《肇域志》与《天下郡国利病书》；中华人民共和国成立后，著名

① 章学诚：《文史通义》卷八，外篇三，《记与戴东原论修志》。

② 《[康熙]永和县志》，《永和县志序》（刘愈深撰）。

③ 《[康熙]滁州志》，《滁州旧志序》（陈琏撰）。

④ 《[康熙]泗水县志》，《重修泗水县志序》（刘桓撰）。

⑤ 《[康熙]重修平遥县志》，《平遥县志序》（王绶撰）。

⑥ 《[嘉庆]延安府志》，《重修延安府志叙》。

⑦ 毛佩琦：《方志资治，一个伟大的传统》，《首届中国地方志学术年会方志文献国际学术研讨会论文集》。

⑧ 瞿宣颖：《方志考稿甲集》，《方志考稿序》。

⑨ 李泰棻：《方志学》，第16页。

气象学家、地理学家竺可桢通过整理研究大量方志，编写出极具科学价值的《中国近五千年来气候变迁的初步研究》，等等。

文以载道，地方志还承载了敦风尚、正人心、乡土教育等功能。梁启超有论：“以中国之大，一地方有一地方之特点，其受之于遗传及环境者盖深且远，而爱乡土之观念，实亦人群团结进展之一要素。利用其恭敬桑梓的心理，示之以乡邦先辈之人格及其学艺，其鼓舞浚发，往往视邈远者为更有力。”^①体现这一功用最明显的莫过于乡土志，其编纂初衷即作为教材，对童蒙进行爱国敬乡教育。

除上述外，地方志还有一些功用有待开发。如旅行向导。相较于虚实参半、真伪难辨的风景、名胜、古迹等旅游资源的讲解与介绍，地方志更具参考性；开发新的旅游产品、重建与再现历史景观、提升自然景观的人文价值等，地方志均可提供借鉴；提高旅行的文化品位、拓宽眼界、增长知识，地方志可以备览。民国时期著名学者黎锦熙曾提出方志之“四用”，其中之一即为“旅行指导”^②，中华人民共和国成立后，陈国生、巴兆祥、林衍经等教授也曾先后撰文论述地方志的旅游价值。

1.1.3 特征

由于时代不同，地域有别，纂修者的初衷、见识各异，地方志大多具有各自的特点。但同时，地方志又具有鲜明的共性，使其区别于其他文献，成为一种特殊的著作形式。早在明代的官、私书目中，方志就已成为相对独立的门类被著录。如编修于正统六年（1441）的《文渊阁书目》，该目以千字文排序，著录文渊阁藏书7297种，其中第十八至二十卷为古今志（杂志附）、旧志、新志；成书于万历三十三年（1605）的《内阁藏书目录》则专设“志乘部”，收录志书748种，885部。私人藏书目录中，陈第的《世善堂藏书目录》设有“方州各志”，著录府州县志、边关志、山水志等；朱睦㮮的《万卷堂书目》则分“地志”“杂志”两部分集中著录志书；董其昌的《玄赏斋书目》卷三分8个子目集中著录志书，分别为：地理志、山川、朝聘行役、名胜、游览、别志、属夷、人物；祁承燾的《澹生堂藏书目》按照统志、约志、省会通志、郡邑志、边镇、山川、祠宇、梵院、胜游、题咏、园林的顺序分11类进行著录，其中“省会通志”又以省志、府志、州志、县志、边关卫所志的顺序排列；此外，赵琦美的《脉望馆书目》与徐渤的《徐氏家藏书目》也将方志集中著录，并按地域排列。^③

关于地方志的特征，学界有诸多论述。笔者所见最早者为朱士嘉先生总结之四个特征：区域性、连续性、广泛性、可靠性。^④其后，来新夏先生在此基础上增加资料性，

① 梁启超：《中国近三百年学术史》，第348页。

② 黎锦熙：《方志今议》，第6页。

③ 张毅：《从明代书目看时人之方志收藏》，《方志文献国际学术研讨会论文集》，中华书局，2012年。

④ 朱士嘉：《中国地方志的起源、特征及其史料价值》，《史学史资料》1979年第2期。

扩为：地方性、连续性、广泛性、资料性、可靠性五个特征。^①刘光禄进一步概括为：记述范围的区域性；内容上的广泛性、资料性和可靠性；编纂工作上的连续性；笔法上的扬善隐恶。^②林衍经认为应当全方位对地方志进行整体考察，从记述的空间范围看，区域性是最显著的一个特征；从编纂的意义上看，史鉴性是又一个特征；从编纂要求看，方志具有纪实性特征；内容的广泛性也是一个灼然可见的特征；此外还具有体系性的特征。^③

20世纪90年代以后，涉及地方志特征研究的论著相当多。著作方面。如仓修良著《方志学通论》认为方志的特征为：地方性、连续性、广泛性、多样性和时代性。^④黄苇等著《方志学》将地方志（包括新旧志书）的特征归纳为：地域性、连续性、普遍性、广泛性、资料性、可靠性、思想性、时代性、实用性、系统性^⑤。刘纬毅等著《中国方志史》主张方志最主要的特征为：空间的地域性、资料的可靠性、内容的广泛性、体裁的类编性以及修纂的连绵性。^⑥沈松平著《方志发展史》介绍方志有地方性、时限性、资料性、叙述性、官修性、时代性、广泛性、多样性、普遍性、可靠性（科学性）等特点。^⑦

论文方面。甄人《试论地方志的本质特征、性质与功能》^⑧一文认为地方志的本质特征有四个方面：地域性、连续性、综合性与资料性。王晖在《方志性质、特征和文献属性》^⑨一文中由方志定义出发，认为方志特征为：地方性、时代性、系统性、资料性与科学性；随后又在《四论方志性质与特征》^⑩一文中从方志空间、时间、结构、内容、质量、编纂六个方面概括方志特征为：地方性、时代性、系统性、资料性、科学性、文献性。赵春满在《地方志的基本特征刍议》^⑪中概括地方志的六个基本特征为：地域性（也可称为地区性、区域性、地方性）、广博性（也可称为广泛性、全面性、百科性）、资料性（也可称为文献性、史料性）、连续性（也可称为延续性、继承性）、规范性、权威性。詹跃华《论地方志书的性质形成与基本特征》一文中又提出志书编纂体例的独特性^⑫。

其他相关论著对地方志特征的概括多未脱离以上表述。综合上述众多观点，学者总结的方志特征主要有：地方性（区域性）、连续性、广泛性、全面性、综合性、资料性、

① 来新夏主编：《方志学概论》，第22页。

② 刘光禄：《中国方志学概要》，第8-10页。

③ 林衍经：《方志学综论》，第21-25页。

④ 仓修良：《方志学通论》（增订本），第9页。

⑤ 黄苇等：《方志学》，第280-285页。

⑥ 刘纬毅等：《中国方志史》，第13-14页。

⑦ 沈松平：《方志发展史》，第11页。

⑧ 载《中国地方志》1996年第5期。

⑨ 载《广西地方志》2004年第1期。

⑩ 载《中国地方志》2005年第1期。

⑪ 载《黑龙江史志》2015年第6期。

⑫ 载《新疆地方志》2012年第3期。

系统性、体系性、可靠性(科学性)、多样性、时代性、详今性、真实性、社会性、普遍性、思想性、实用性、多用性、文献性、类编性、叙述性与官修性等。这些对方志特征的概括与描述有的指向方志修纂活动,有的指向志书内容,包括志书的收载范围、体例结构、体裁文风,还有的指向方志的价值功用,面面俱到。但是我们在方志数字化实践中发现,仅有这些立足于方志本身的特征描述还远远不够,需要更为详细、具体、专门的面向数字化的关于地方志文献特性的分析与阐释。

1.2 地方志数字化

地方志数字化指“利用计算机技术,将纸质地方志或者胶片地方志转换为影像、全文文本数据,并在此基础上制作书目、索引、资源库等数据的过程”。^①我国的方志数字化建设始于20世纪90年代,迄今,无论是数字化实践还是技术、方法的研究,均取得了不俗的成绩。

1.2.1 地方志数字化实践

目前许多图书馆、博物馆、高校、研究所、方志办以及商业机构都对地方志进行了数字化,开发了书目数据库、全文图像数据库和全文文本数据库,取得了卓著的成绩。就地方志数字化实践的建设与未来方向,已有一些学者进行了分析、总结和展望,如蓝凌云、莫争春合撰《我国旧志数字化的探索与思考》(载《中国地方志》2005年第4期),毛建军撰《国内数字方志资源的开发与建设探析》(载《广西地方志》2006年第3期)与《中国地方志数字资源建设的进展与前瞻》(载《广西地方志》2013年第1期)等。

下面就地方志数字化实践中之要者进行简要介绍。

1.2.1.1 书目数据库

地方志书目数据库主要用于方志资源检索,与纸本地方志目录功用相似,但使用更为方便、快捷。

台湾地区“中央研究院”的“中国大陆各省地方志书目查询系统”^②著录了国内外藏各代地方志,起自宋元迄于公元2003年,“堪称当今世上最完整者”。该系统以《中国地方志联合目录》为基础,参照《中国地方志总目提要》等,增补1949年以后各地新编如省、县、市志、自治州志、乡志、街志等二千余种。针对各地原有孤本地方志新刊行者,亦加以著录。著录项目包括:书名、卷数、纂修者、丛书名、出版地、修纂时间、出版时间、版本、备考、馆藏地与稽核项。

^① 王芸、肖禹:《地方志数字化模式与案例分析》,第1页。

^② <http://webgis.sinica.edu.tw/place/>

一些文献公藏机构立足馆藏或所属地区,建设了专门的地方志目录、提要数据库。如北京师范大学的“北京师范大学图书馆线装方志书目数据库”^①、湖南图书馆的“湖南地方志”^②等。有的则在地方文献数据库中提供地方志的书目数据检索,如甘肃省图书馆的“西北地方文献书目提要数据库”^③、厦门市图书馆的“闽南地方文献联合目录库”^④等。还有的在古籍书目数据库中提供地方志的书目数据检索,如由北京大学牵头,联合国内23家高等院校图书馆合力建设的“学苑汲古——高校古文献资源库”^⑤、约30余家中美图书馆参与的“中华古籍善本国际联合书目系统”^⑥、上海图书馆的“古籍书目查询”^⑦等。随着图书馆编目系统的集成化发展,许多图书馆在统一的书目检索平台上即可检索出地方志目录。如国家图书馆^⑧、天津图书馆^⑨等。

海外地区许多文献典藏机构也将其所藏中文古籍建设目录数据库,并于其中提供地方志的检索,如“日本所藏中文古籍数据库”^⑩等。

1.2.1.2 图像数据库

图像数据库(或称影像库)是将各种载体的文献转换为图像数据,并以图像方式存储和呈现的数据库。图像库是当前条件下最能反映文献原貌的数字化类型,可以代替纸本使用。其技术难度、开发成本相对较低,缺点是不能检索。

许多图书馆、博物馆等文献公藏机构都将其所藏地方志图像化,提供互联网或局域网浏览。如国家图书馆的“数字方志”库^⑪,提供国家图书馆藏六千余种地方文献图像的在线浏览。“数字图书馆推广工程·地方志”库^⑫包含了陕西、山东、湖北、云南、贵州、新疆、江西、青海等地的地方志,种类包括通志(存目)、府县志、乡镇志、山水志等,既有旧志,又有新志。

万方数据知识服务平台的“中国地方志数据库”^⑬于2014年开始在新方志数据库基础上,系统集成旧方志文献,预计收录50000册。超星数字图书馆开发的“中国历代方志集(1368-1949)”收录了明清时期及民国时期的各地史志。

一些古籍资源图像数据库同时提供方志浏览,如国家图书馆的“中华古籍资源

① <http://digi2.lib.bnu.edu.cn:8080/digilib/search?channelid=26966>

② <http://www.library.hn.cn/tszy/dfwx/hndfz/>

③ <http://www.gslib.com.cn/xdwx/xdwxjs.asp>

④ <http://www.xmlib.net/sztsg/zjsjk/mndfwx/>

⑤ <http://rbsc.calis.edu.cn:8086/aopac/jsp/indexXyjc.jsp>

⑥ <http://mylib.nlc.cn/web/guest/zhonghuagujishanbenlianhesumuxitong>

⑦ <http://search.library.sh.cn/guji/>

⑧ <http://www.nlc.cn/>

⑨ <http://opacwh.tj.tj.cn:8991/F>

⑩ <http://www.kanji.zinbun.kyoto-u.ac.jp/kanseki>

⑪ <http://mylib.nlc.cn/web/guest/shuzifangzhi>

⑫ <http://mylib.nlc.cn/web/guest/zhengjidifangzhi>

⑬ <http://c.wanfangdata.com.cn/LocalChronicle.aspx>

库”^①、“东京大学东洋文化研究所汉籍影像数据库”^②等。

需要说明的是,随着数字化的推进,很多图像数据库升级为文本或是图文数据库。

1.2.1.3 文本数据库

文本数据库即通过将纸质文献人工录入或将图像数据进行机器识别转换成文本数据,加以存储、利用的数据库。相对书目数据库、图像数据库,文本库更复杂、技术难度更高。但其支持全文检索和内容标引,是信息检索、抽取以及数据分析、挖掘等应用的基础。

现今规模较大、应用较广的地方志全文数据库主要有:北京爱如生数字化技术研究中心研制的“中国方志库”,收录汉魏至民国历代地方志类典籍约1万种,提供图文逐页对照;北京籍古轩图书数字技术有限公司开发的“中国数字方志库”,该数据库先期收录1949年以前地志类文献万余种,提供高清晰影像,新版提供文本检索。

很多古籍文本数据库提供地方志全文检索。如北京爱如生数字化技术研究中心研制的“中国基本古籍库”、天一阁博物馆的“古籍数字资源库”^③等。

现阶段我国地方志数字化实践发展至何种程度呢?我们可以参照数字化层次的划分来进行大致地界定。王立清著《中文古籍数字化研究》按照古籍数字化处理深度的不同,将古籍数字化分为古籍书目的数字化、古籍载体的数字化、古籍图文的数字化、古籍知识及关联的数字化五个层次。其中,书目数字化即将古籍书目转换为计算机可读可检的书目系统;载体数字化即把古籍制作成影像,也即图像化;图文的数字化即将古籍全文文本数字化,实现全文检索,并保留影像;古籍知识及关联的数字化即将古籍全文有关的知识进行重组,挖掘深度的知识。^④王荟、肖禹著《地方志数字化模式与案例分析》将文献数字化分为三个层次,分别为:纸本的替代、数字资源的功能提升、知识提供^⑤。纸本的替代即单纯用以替代纸本文献的图像数据和文本数据;数字资源的功能提升指依托目录、索引、文本等数据,通过检索软件和数据库实现海量资源的全文检索;知识提供则指打破文献原有的组织结构,将大量文献还原为不同粒度的知识和与知识有关的资源。两种描述看似不同,实则相似,其区别在于划分层次的细化程度。由两种描述可以预见,地方志数字化下一步发展的方向当在知识重组与提供。

1.2.2 地方志数字化研究

地方志数字化研究的主要目的应是总结实践中的不足,分析需要突破的困境以及

① <http://mylib.nlc.cn/web/guest/shanbenjiaojuan>

② <http://shanben.ioc.u-tokyo.ac.jp/index.html>

③ <http://www.tianyige.com.cn:8008/Default>

④ 第25页。

⑤ 第11、12页。