



古籍文本数据 格式比较研究

肖禹 | 著

 上海遠東出版社



古籍文本数据 格式比较研究

肖禹 | 著

上海遠東出版社

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

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》《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.....	1
一、引言.....	1
(一) 古籍数字化.....	1
1. 概念.....	3
2. 层级.....	6
3. 问题与对策.....	7
4. 标准规范.....	9
(二) 古籍数字化与学术研究.....	13
1. 数字人文.....	14
2. 知识遮蔽.....	15
二、古籍文本化.....	17
(一) 数据.....	17
(二) 加工过程.....	17
三、古籍文本化理念.....	18
(一) 面向应用.....	18
(二) 服务学术.....	18
(三) 利用技术.....	19
(四) 工程项目.....	19
(五) 保存信息.....	21
(六) 标准规范.....	21

第二章 古籍文本模型	23
第一节 简单对象	23
一、文字	23
(一) 文字类型	24
1. 字符集	24
(1) Unicode	25
(2) 中华字库	32
2. 集外字	33
(1) 集外字问题	34
(2) 集外字处理方法	35
(二) 文字属性	36
1. 字体	36
2. 字号	37
3. 文字位置	38
4. 文字颜色	38
5. 文字变形	38
6. 文字旋转	38
二、符号	39
(一) 符号类型	42
1. 标点符号	43
2. 校对符号	43
3. 版式符号	44
4. 专类符号	46
(二) 符号属性	47
三、图形	47
(一) 图形类型	47
1. 线段	48
2. 圆弧	48
3. 圆形	49
4. 矩形	49
(二) 图形属性	49

四、图像	50
(一) 图像类型	50
1. 版框内插图	50
2. 书叶内插图	52
3. 其他插图	52
(二) 图像属性	54
1. 图像尺寸	54
2. 分辨率	54
3. 颜色模式	54
第二节 复杂对象	54
一、大小字	54
二、墨围	56
三、墨盖子	57
四、表格	58
五、图形组合	60
六、特殊图像	60
(一) 牌记	60
(二) 印章	61
七、版式	61
(一) 普通版式	62
(二) 特殊版式	63
1. 无版式	63
2. 不规则版框	64
3. 格抄本	64
4. 多截板	64
5. 图文混排	65
第三节 结构对象	66
一、古籍的物理结构	66
(一) 古籍装帧形式	66
(二) 古籍图像	67
二、古籍的逻辑结构	67

第三章 纯文本	69
第一节 纯文本格式	69
一、源起	69
二、现状	70
(一) 汉籍电子文献资料库	70
(二) CBETA 电子佛典集成	72
(三) 中国基本古籍库	74
(四) 古籍电子定本工程	75
(五) 《汉籍全文数字化工作流程指南》	76
三、数据模型	78
(一) 结构对象	78
(二) 简单对象	81
1. 文字	81
2. 符号	84
3. 图像	85
4. 图形	85
(三) 复杂对象	86
1. 大小字	86
2. 墨围	87
3. 墨盖子	87
4. 表格	88
5. 图形组合	89
6. 特殊图像	89
第二节 纯文本格式描述	89
一、纯文本 XML 结构	89
(一) 文件头	89
(二) 书目元数据	90
(三) 文本数据	91
(四) 集外字数据	92
二、纯文本 XML Schema	93

第三节 纯文本 XML 示例	105
一、示例 1	105
二、示例 2	107

第四章 位置文本	110
----------	-----

第一节 位置文本格式	110
一、源起	111
二、现状	112
三、数据模型	112
(一) 结构对象	112
(二) 简单对象	115
1. 文字	115
2. 符号	117
3. 图像	117
4. 图形	119
(三) 复杂对象	119
1. 大小字	119
2. 墨围	121
3. 墨盖子	122
4. 表格	123
5. 图形组合	125
6. 特殊图像	125

第二节 位置文本格式描述	125
一、位置文本 XML 结构	125
(一) 文件头	126
(二) 书目元数据	126
(三) 卷目数据	126
(四) 文本数据	127
(五) 集外字数据	128
二、位置文本 XML Schema	128

第三节 位置文本 XML 示例 140

一、示例 1 140

二、示例 2 146

第五章 版式文本 153

第一节 版式文本格式 153

一、源起 153

二、现状 154

(一) 文渊阁四库全书电子版 155

(二) 爱如生大型古代数据库 161

(三) 数字方志 166

(四) 《中文文献全文版式还原与全文输入 XML 规范》 173

三、数据模型 175

(一) 结构对象 175

(二) 简单对象 180

1. 文字 180

2. 符号 183

3. 图形 184

4. 图像 186

(三) 复杂对象 187

1. 大小字 187

2. 墨围 189

3. 墨盖子 191

4. 表格 192

5. 图形组合 199

6. 特殊图像 201

7. 版式 202

第二节 版式文本格式描述 202

一、头文件 XML 结构 202

(一) 文件头 203

(二) 书目元数据 203

(三) 卷目数据	203
(四) 默认版式数据	203
(五) 集外字数据	204
二、叶文件 XML 结构	204
(一) 文件头	205
(二) 叶文本	206
(三) 集外字数据	206
三、版式文本 XML Schema	206
(一) 头文件 XML Schema	206
(二) 叶文件 XML Schema	220
第三节 版式文本 XML 示例	232
一、示例 1	232
二、示例 2	237

第六章 语义文本 245

第一节 语义文本格式 245

一、源起	245
(一) 语料库	245
(二) 内容标注	247
(三) 数据抽取	248
二、现状	249
(一) 台湾地区“中研院古汉语语料库”	249
(二) 北大 CCL 古代汉语语料库	251
(三) 国家语委古籍语料库	252
(四) 中华古籍语料库	252
(五) “汉语史语料库建设研究”项目	252
三、数据模型	253
(一) 结构对象	254
(二) 内容对象	256
1. 图像	256
2. 图形	257

3. 表格	257
(三) 标注对象	261
1. 文本碎片属性	261
2. 句型	262
3. 词类	263
第二节 语义文本格式描述	266
一、语义文本 XML 结构	266
(一) 文件头	266
(二) 书目元数据	267
(三) 来源文本属性	267
(四) 卷目数据	268
(五) 标注集合	268
(六) 文本数据	269
(七) 集外字数据	269
二、语义文本 XML Schema	269
第三节 语义文本 XML 示例	289

第七章 部分文本 304

第一节 谱系文本格式 304

一、源起	304
二、现状	305
(一) GEDCOM	305
(二) 浙江图书馆家谱全文数据库	308
(三) 中华寻根网	309
(四) 家谱世系数据规范	311
(五) GEDCOMX	314
(六) “家谱谱系数字化模型研究”项目	318
三、数据模型	318
(一) 实体	319
(二) 实体间关系	321

第二节 谱系文本格式描述	325
一、谱系文本 XML 结构	325
(一) 文件头	325
(二) 书目元数据	326
(三) 卷目数据	326
(四) 实体间关系数据	326
(五) 实体数据	327
(六) 集外字数据	328
二、谱系 XML Schema	328
第三节 谱系文本 XML 示例	346
一、宗族模式示例	346
二、家庭模式示例	359
 第八章 文本格式比较	 379
第一节 文本格式分析	380
一、全文文本	380
(一) 格式比较	380
(二) 格式简化	382
1. 数据模型简化	382
2. 数据描述简化	383
(三) 格式转换	384
1. 版式文本转换为纯文本	384
2. 纯文本转换为版式文本	385
(四) 语义文本	387
二、部分文本	387
第二节 复合文本格式	387
一、复合文本	388
二、复合文本示例	388
(一) XML Schema	389
(二) XML	415
1. 示例 1	415

2. 示例 2.....	421
--------------	-----

参考文献.....	432
-----------	-----

一、专著.....	432
二、标准.....	433
三、论文.....	434
四、电子和网络文献	443

附 录.....	450
----------	-----

一、古籍元数据规范 (CDLS-S05-013)	450
二、中文文献全文版式还原规范	453
三、中文文献全文版式还原规范 XML Schema.....	464
(一) 头文件 XML Schema.....	464
(二) 叶文件 XML Schema.....	480
四、家谱谱系数据规范	495
(一) 结构说明	495
(二) 标签及属性说明.....	495
五、家谱世系数据规范 XML Schema.....	498
六、“中研院”上古汉语语料库词类与特征标记表	500
(一) 词类标记表	500
(二) 词类标记说明表.....	501
(三) 特征标记表	502
七、GEDCOM 5.5 标签与 GEDCOM XML 对应关系	503

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.