“中国地方志数字化关键技术研究与演示平台设计”

项目述略

肖禹

国家科技支撑计划项目“中国地方志数字化关键技术研究与演示平台设计”（2015BAK07B00）由文化部文化科技司组织，国家图书馆承担，项目期为 2015 年 1 月至 2017 年 12 月。本项目旨在进一步归纳地方志文献特性，建立地方志数字化标准规范体系，研究地方志数字化关键技术和可视化技术，设计地方志资源服务演示平台，为地方志大规模数字化和全国性地方志资源服务平台构建提供技术支撑。

1.项目概述

1.1 项目意义

地方志是记载一地情况的区域史书。地方志的历史非常久远，其起源可追溯至先秦时期的典籍《周官》，以后历代都有发展。特别是明清时期，地方志之书大量涌现。经过上千年的发展演变，地方志形成了其独特的类型：反映全国性的地方志称“一统志”，如《大明一统志》、《大清一统志》；省级的称“通志”，如《河南通志》等；州、府、县、乡、镇也各有州志、府志、县志、乡志和里镇志；此外还有卫志、关志、盐井志、土司志等。除此之外，还有山志、水志、湖志、塘志、河渠志以及书院志、古迹志、寺观志、游览志、路桥志等。明清时期地方志的编撰逐渐由政府主导，成为各级官府的重要职责。建国以来，党和政府非常重视新方志的编纂、旧方志的整理、出版、地方志数据库建设等，并取得了丰硕的成果。各级政府为统筹地方志工作，都成立了专门的地方志编纂办公室。2006 年国务院还颁布了《地方志条例》。该条例指出了地方志编撰的重要意义，明确地方志编撰主体、时限要求、经费保障等等，对于下一步做好地方志工作提供了保障。

2014 年 4 月 19 日，第五次全国地方志工作会议在北京召开。国务院总理李克强作出重要批示，指出：“地方志是传承中华文明、发掘历史智慧的重要载体，

存史、育人、资政，做好编修工作十分重要”，要求各级编撰人员“以更加饱满的热情、以求真存实的作风进一步做好地方志编纂、管理和开发利用工作，为弘扬优秀传统文化、服务经济社会发展作出新的贡献”。刘延东副总理在与部分会议代表座谈时强调，各地区各部门要认真贯彻落实李克强总理重要批示精神，进一步推动地方志事业的发展和繁荣，为全面建成小康社会和全面深化改革提供历史借鉴和智力支持，强调要着眼于提高国家文化软实力、展示中华文化独特魅力的大局，将地方志工作视为公共文化服务体系中重要的一环。同时，刘延东副总理在谈话中特别强调，“各级地方志工作机构要发挥优势，创新服务手段和方式，拓宽服务渠道，用人们喜闻乐见的方式利用地方志、传播地方志。地方志资源开发利用是一篇大文章，比如教育系统应当鼓励全国青少年阅读地方志，有些地方志内容应纳入当地中小学教材或课外读物，把当地历史上有特色的东西挖掘出来，让每一个出生在这片土地上的人都不能忘记祖宗、忘记历史。”

2014 年 2 月 25 日至 26 日，习近平总书记在北京考察工作时的讲话中指出：“历史文化是城市的灵魂，要像爱惜自己的生命一样保护好城市历史文化遗产……搞历史博物展览，为的是见证历史、以史鉴今、启迪后人。要在展览的同时高度重视修史修志，让文物说话、把历史智慧告诉人们，激发我们的民族自豪感和自信心，坚定全体人民振兴中华、实现中国梦的信心和决心”。2014 年 3 月 27 日，习近平总书记在联合国教科文组织总部的演讲指出：“中国人民在实现中国梦的进程中，将按照时代的新进步，推动中华文明创造性转化和创新性发展，激活其生命力，把跨越时空、超越国度、富有永恒魅力、具有当代价值的文化精神弘扬起来，让收藏在博物馆里的文物、陈列在广阔大地上的遗产、书写在古籍里的文字都活起来，让中华文明同世界各国人民创造的丰富多彩的文明一道，为人类提供正确的精神指引和强大的精神动力。”

据不完全统计，汉文古籍超过 20 万种，地方志约占 5%。地方志同时具备了地域性、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，建立资源库和知识库，并通过数据挖掘技术和可视化技术使“古籍里的文字都活起来”。

1.2 课题设置

项目分解为 3 个课题：

课题 1：地方志资源调查数字化加工规范研究（2015BAK07B01），由国家图书馆负责，汉王科技股份有限公司和中国传媒大学共同承担；

课题 2：地方志数字化与知识抽取技术研究（2015BAK07B02），由汉王科技股份有限公司负责，国家图书馆共同承担；

课题 3：地方志可视化技术研究与演示平台实现（2015BAK07B03），由华中师范大学负责，北京图捷讯通软件技术有限公司和武汉华大国家数字化学习工程技术有限公司共同承担。

2.研究方案

2.1 课题 1

本课题主要目标是通过点面结合的调查方式，研究清康熙时期纂修方志的存藏情况和文献特性，重点考察山西介休历代方志的纂修与存藏情况；利用国家图书馆数字方志项目与本课题相关的地方志图像 44000 筒子页，再依据课题需要补充 6000 筒子页清康熙时期纂修方志、介休县志等资料，作为本课题的数据基础；依据项目需求建立 8 个数字化加工规范和 4 个相关软件；在调查的基础上，遵循数字化加工规范，开发语料和本体加工软件，研制地方志语料库和时间本体库、地点本体库；为了演示平台和应用示范的需要，基于方志文献特性研究的成果，选择方志中专题内容进行文化创意，完成 4 个可视化脚本创意设计。如图 1 所示。

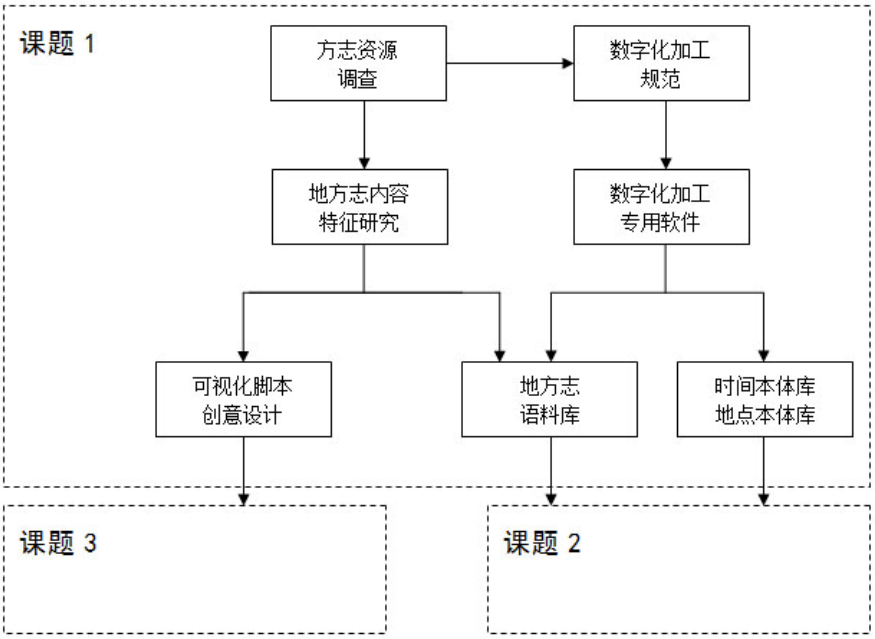

图 1　课题 1 结构图

调查现存清康熙时期纂修方志的存藏状况、编纂体例和内容特性，鉴定地方志版本，厘清增修、补修、续修、重修等关系，研究地方志记录的基本内容、主要形式和文字描述方式，同时注重个案研究，分析地方志内容与纂修时间、地域、文化等方面的内在联系。

依据地方志文献特性和项目需求建立5类8个地方志数字化加工规范，包括：文字处理规范 2 个：汉字集外字描述规范和文字认同描述规范；元数据规范 2 个：地方志专门元数据规范和地方志卷目数据标引规范；对象数据规范 2 个：地方志图像数据规范和地方志文本数据规范；语料数据规范 1 个：地方志语料数据规范；知识库数据规范 1 个：古今地名数据规范。地方志专门元数据规范用于整体描述地方志图像数据和文本数据，内容包括书目信息、数字化参数、版权信息、用户使用提示等。

依据地方志图像数据规范，开发自动拼图软件，将半叶（筒子页）、双半叶、部分叶等多种类型的地方志图像，依据版心、鱼尾、纸张纹理等特征，自动拼接成筒子页，同时具备手动拼接等辅助功能。依据地方志语料数据规范，开发语料数据标引软件，自动过滤文本数据中的版式标记或符号，可以将指定的版式标记或符号按规则转换成语义连续标记，支持自动或半自动语料块切分和语境标注。依据集外字描述规范，开发集外字描述软件，依据集外字位置描述自动截取集外字图像，支持集外字人工拆分、IDS 描述、IDS 动态显示等，同时支持 IDS 检索和文字部件辅助拆分。依据地名数据规范，开发地名数据规范加工软件，输入、编辑、查询、显示古今地名规范数据，支持地名的自动或半自动匹配。

依据语料数据规范，以地方志文本数据为基础，完成单字索引、语料块切分和语境标注，语料库数据量不少于 2 千万字。依据项目需求，研制时间本体库，时间起于汉高祖元年（前 206），自汉平帝元始元年（前 1）起加入公历，自唐高祖武德五年（622）起加入回历，南朝宋后废帝元徽四年公历（476）前并注明罗马历；研制地名本体库，包含古今地名（行政区划地名）不少于 1 万个，满足课题 2 和课题 3 的数据需求。

基于地方志文献特性和项目演示的需求，以山西介休方志为数据基础，加入周边地区部分有特色的方志，提炼地方文化的亮点，进行文化创意，组织与之相关的各类数字资源，形成某一主题的数据集合，完成 4 个地方志可视化脚本，包

括山西介休与介子推、山西介休与方言、山西洪洞与大槐树、山西太谷与秧歌。

2.2 课题 2

本课题以地方志数字化技术和知识抽取技术研究为核心，由三部分组成：地方志文献数字化技术研究与实现、地方志数据抽取技术研究与实现和方志文本数据、知识库数据等示范数据建设。

地方志的原始文本大多是古籍类的资料，年代比较久远，纸张比较薄、透光性较强、比较脆、颜色发黄、字迹偏淡；版式上有横排文字、竖排大字、单（多）行竖排小字、表格、图像等多种类型；文字字符集大，变形多，各种刻本字迹风格各不相同。本课题涉及的文字识别部分限定在版刻方志资料，手写地方志资料的文字识别不在本课题的研究范围内。基于以上特点，地方志数字化的过程包括图像处理、版式切分、文字识别、文字校对、集外字处理、数据标引等一系列过程。可以把这些技术大略划分为 OCR 识别、集外字处理技术和数据标引技术三部分。

OCR 识别是将纸质文件扫描生成图像，采用适合古籍版式的自动版面分析技术和文字切分技术，对图像进行自动分割；对于分割得到的文字图像，采用大字符集汉字 OCR 方法，进行文字识别。为了保证识别的准确性和可靠性，集成多种识别方法，实现分层分类识别。如图 2 所示。

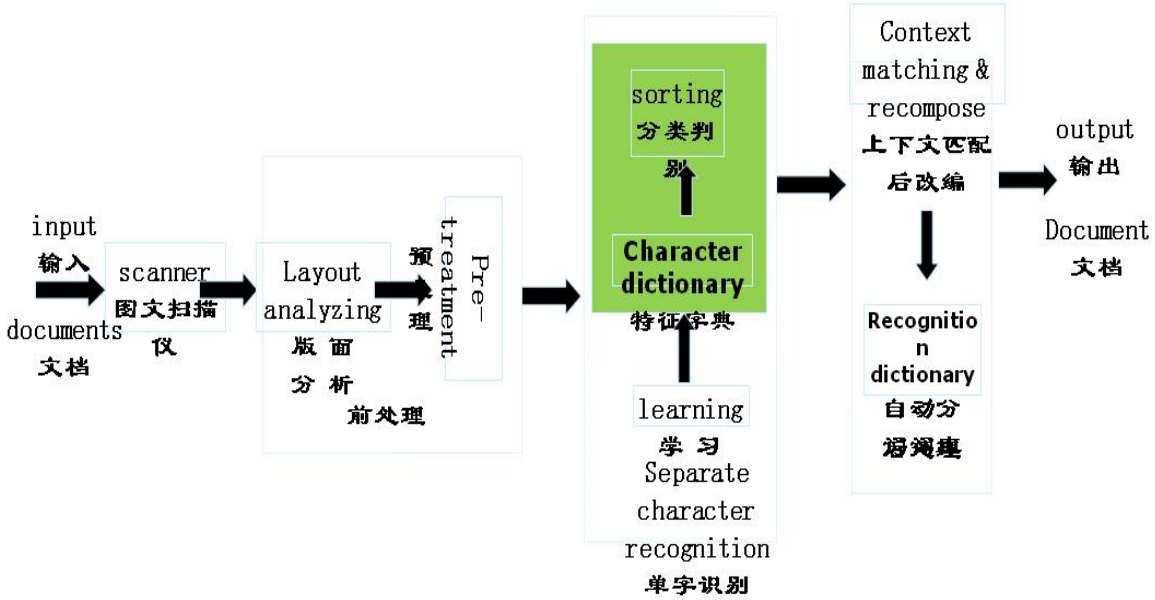

图 2　OCR 系统模块流程图

集外字处理技术是在对地方志文本进行数字化转换的过程中，对于出现的集外字，拟采取基于 IDS（Ideographic Description Characters，表意文字描述序列）的集外字描述方法。本课题中将开发相应工具，实现 IDS 描述及集外字检索技术。

数据标引技术是围绕地方志文本规范标准，结合方志资料特征，采用自动版式分析技术和语义理解技术进行数据标引，包括文字标引，符号标引、版式标引、语义标引等。

地方志知识抽取技术是以地方志语料库为基础进行数据抽取，得到时间、地点、人物、事件等地方志元数据关系网络，抽取感兴趣的内容，分析其特点、关系，从而发现隐藏在多本地方志之间的规则与事实。地方志数据抽取的工作包括知识本体的标志、抽取规则的设置、分析器的建立、元数据标准建立元数据聚类以及元数据关系网络的生成。在地方志数据抽取技术研究中，需要利用地方志专家知识、语义理解技术、统计技术、数据库技术、聚类技术、图书文献数字化标准制定及其它信息学科技术。半自动本体构建流程如图 3 所示。

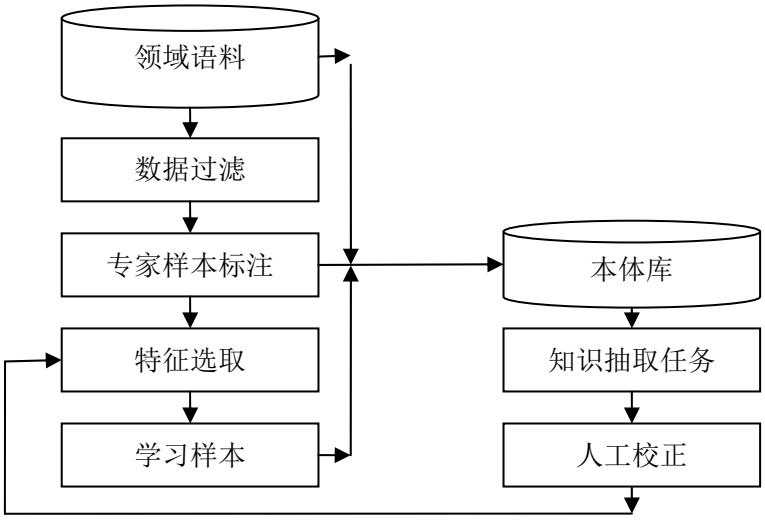

图 3　半自动本体构建的流程图

方志文本数据、知识数据的建设，在对本次选定的地方志文献全文数字化后，标志并抽取时间本体、人物本体等能够反映当时人物风情的相关知识元，建设方志知识数据。还需要根据业务需求，标注各知识元之间的关系，为后续知识元网络的建立和相关数据挖掘作好数据准备。

2.3 课题 3

本课题是以中国地方志可视化关键技术及演示平台研究为导向，研制地方志可视化演示平台，包括地方志数据多媒体展示技术、地方志地理信息可视化、地方志知识可视化、多源知识融合可视化等技术，在此基础上形成地方志数据相关技术的解决方案。

地方志资源具有多种类型，资源呈现富媒体的特点，不同平台对各自所能展示的资源具有不同的要求，针对智能手机（PDA 终端）、PC 等不同平台如果抽取出相应的资源与信息是地方志数据跨平台可视化展示技术研究的第一个技术点。本课题拟采用富媒体文本解析技术、多媒体解析技术、语义挖掘技术等信息抽取方法将富媒体资源中的信息进行分离，抽取成不同粒度的信息单元。对地方志数据跨平台可视化展示技术的研究，拟基于 epub3 标准，采用 Html5，java script 等技术，首先根据地方志专家的需求和内容设计，对地方志中的数字化文本、图片、音频、视频与动画等多种类型的知识媒体进行信息同步与图文混排，然后利用脚本语言针对地方志数据媒体的交互需求编写动态交互脚本，最后，将实现的地方志可视化内容形成 epub3 格式的媒体类型，并发布到移动终端上，研发出基于移动终端的地方志可视化案例，满足跨平台用户对动态化、立体化地方志数据可视化浏览的需求。

本课题拟通过一种统一的“类别—属性—词条—属性内容”层次框架，构建统一的分类体系，抽取类别属性建立“类别—属性”模板，并最终提取属性内容。解决多源语义知识库融合中存在的知识不一致问题，为建立大规模计算机“可读”的语义知识库资源提供可行性解决方法。针对多源语义知识库中“类别—属性”模板的构建问题，本课题拟采用基于多过滤器驱动和分散度统计类别属性的方法和基于语义关联挖掘类别属性的方法把类别属性抽取问题转换到模糊集中的求解上下近似问题进行求解，采用代数计算的方法来实现知识融合中的推理。本课题针对知识可视化中的两个关键环节进行重点研究，首先采用语义模型抽取出地方志资源中描述信息的主题词，在此基础上结合文本挖掘的相关技术，挖掘出主题词之间的关联关系并构建知识地图。同时调用 GIS 地理信息的输出接口，在统一框架下，基于文本索引、时间索引和空间索引导航方式，实现地方志的文本、时间、空间信息的有机融合，并对地方志空间信息进行三维建模与呈现，如图 4

所示。

图 4　多源知识融合效果图

基于 GIS 的地方志信息系统，以 API 接口的方式对外提供地图服务的基本功能，主要包括：地图切片生成，根据地方志显示的特点，生成地方志基础电子地图切片数据；地图查看，提供地图查看时的基础功能，包括平移、缩放功能、地图比例尺、全屏显示等；标注，使用地图 API 可以在地图上绘制点并录入说明等信息；折线，使用地图 API 可以在地图上绘制折线并计算距离；绘制区域，使用地图 API 可以在地图上绘制多边形并测算总面积；地图上叠加点、线、面（几何图形），在地图上描画点、线、几何图形；计算两点间距离，计算地球上两点的球面距离；计算多点间距离，计算地球上多点之间的折线距离；判断位置关系，判断点与几何图形的位置关系；计算面积，计算几何图形的面积；地名搜索功能，以 API 接口的方式对外提供地名搜索的服务。

地方志数据演示平台的系统架构，如图 5 所示，包含数据层、服务层和应用层，其中数据层实现地方志演示数据及知识的集成与存储，服务层基于 Web Service 技术，开发地方志知识体系、地方志资源管理和多维检索等服务组件，集成地方志可视化业务系统和 GIS 地方志信息业务系统，面向地方志专业人员、管理人员和普通民众对地方志的应用需求。

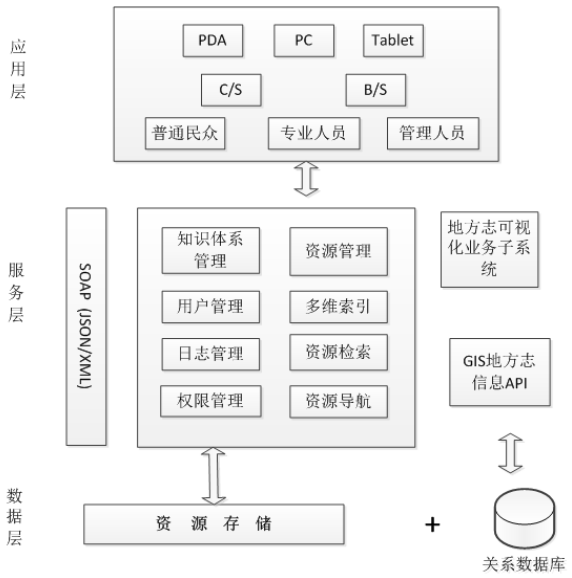

图 5　地方志数据演示平台架构图

3.预期成果

开展清康熙时期纂修地方志及山西介休历代方志的资源调查，归纳地方志文献特性并结合相关史料，编制专题目录。依据地方志文献特性，研制 8 个地方志数字化加工规范，包括：地方志专门元数据规范、地方志卷目数据标引规范、汉字集外字描述规范、文字认同描述规范、地方志图像数据规范、地方志文本数据规范、地方志语料数据规范和古今地名数据规范等，并规范，开发自动拼图软件、语料数据标引软件、集外字描述软件和地名数据规范加工软件。

依据项目需求完成 5 万叶地方志样例数据，包含目录、索引、图像和文本数据，并在此基础上完成语料数据库 1 套，语料总量超过 2 千万字。研建时间本体数据库 1 套，时间起于汉高祖元年（前 206），自汉平帝元始元年（前 1）起加入公历，自唐高祖武德五年（622）起加入回历，南朝宋后废帝元徽四年公历（476）前并注明罗马历。研建地名本体数据库 1 套，包含古今地名 1 万个。

研究地方志数字化相关技术，研发地方志数字化工具软件 1 套，满足地方志数字化过程中的图像处理、版式分析、OCR 文字识别、文字校对、集外字处理、IDS 描述、数据标引等功能，支持生产流水线管理和数据包流转管理。其中，古籍汉字识别软件支持超大字符集汉字识别，支持的字符集为 GB18030—2005 规定的字符集；对于版刻古籍汉字，识别率达到 80%以上。

研究地方志文本知识挖掘技术，研发地方志文本知识抽取系统，支持对特定

地方志文本的专家分析、概念抽取、基础本体知识创建、特定内容抽取等。提交 1 套知识库，包括人物知识 1000 条，地点知识 10000 条，事件知识 1000 条。

研究地方志可视化技术，构建 1 套地方志可视化演示平台，包括地方志数据前端展示系统、地方志 WEBGIS API 及展示系统。支持多维度地方志文本、图片、视频和元数据的导入和管理；支持一套地方志 WEBGIS API，能够以 API 接口方式进行基础地图服务；支持一套地方志位置编辑工具，能够采集地方志数据位置数据，并与方志数据库进行关联；支持多操作系统，包括 Windows7 及以上版本、IOS 6.0 及以上版本、Android3.2 及以上版本；支持多浏览器，包括 Windows Internet Exporer8.0 及以上版本、Google Chrome10.0 及以上版本、Firefox4.0 及以上版本；满足地方志数据的可视化展示及地图浏览、地图缩放、地名查询、信息标注、距离量算等地方志 WEBGIS 服务。

完成地方志可视化脚本 4 个，包括山西介休与介子推、山西介休与方言、山西洪洞与大槐树、山西太谷与秧歌等。

申请发明专利 12 项；获得软件著作权 8 项；公开发表相关学术论文 22 篇、出版专著 6 部；滚动培养研究生 15 人。