“中国地方志数字化关键技术研究与演示平台设计”完成情况述略

陈红彦

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3年，国拨及各单位配套项目经费总额超过1千万。本项目旨在面向地方志数字化工程，通过分析地方志文献特性，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，以深度知识挖掘利用为特征，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。

项目分解为3个课题：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；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；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。

本项目以1912年以前编纂或出版的地方志为研究对象，研究成果也适用或部分适用于于1912年至1949年间编纂或出版的地方志。

1.项目目标

本项目的目标旨在创立地方志数字化、知识化和可视化模式，促进地方志的应用与推广，使“书写在古籍里的文字活起来”。通过演示平台示范，将地方志目录、索引、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、文献三个维度的知识融合，并与方言和方块文字项目对接。

本项目的研究目标紧密结合《纲要》的要求和国家的实际需求，以方志学、图书馆学理论为指导，在已有的地方志数字化成果上，以中文信息处理、数据库、GIS、数据挖掘、人工智能等先进技术为手段，探索地方志数字化与资源服务的新方法，实现现代技术与传统文化的紧密结合，为社会发展、经济发展、文化教育、学术研究等提供强有力的支撑。通过演示平台的示范效应，带动地方志知识库的建设与应用，形成“信息与资源汇聚、管理与服务

融合、线上与线下互通”的地方志服务新模式。

2.项目完成情况

本项目的主要任务是在调查中国地方志现有资源的基础上，点面结合，以康熙朝地方志和山西介休历代方志为样本，建立标准规范体系服务于地方志数字化，研究地方志文献数字化技术、数据抽取技术、可视化技术等作为技术支撑，实现地方志 GIS 和演示系统，并进行示范应用。

2.1 地方志资源调查

项目组在充分调研国家图书馆藏方志的基础上，选择部分外省市开展实地考察。实地考察的目的主要是两个，一是进一步调查康熙朝地方志的存藏情况以及历代山西介休方志的编纂与存藏情况，并对部分版本进行考证，更加全面、准确地把握方志的内容特性；二是围绕项目内容，采集文物、非遗、方言、名胜古迹等方面的素材，作为 GIS 系统呈现和可视化展示的基础资料。所选的地方不仅方志收藏丰富，而且还需有丰厚的文化底蕴。

本项目选取康熙朝方志分布较多的几个省市，包括山西、江苏、浙江、上海、福建、甘肃 6 个省，并且选取这几个省中比较关键的 16 县市进行实地调研,如表 1 所示。

表 1 地方志资源调查表

省	市/县	考察时间
山西	太原	2016 年 7 月 4 日—2016 年 7 月 8 日
	临汾	2016 年 8 月 1 日—2016 年 8 月 5 日
	太谷	2016 年 10 月 10 日—2016 年 10 月 14 日
		2017 年 8 月 7 日—2017 年 8 月 11 日
	介休	2017 年 5 月 22 日—2017 年 5 月 26 日
		2017 年 9 月 18 日—2017 年 9 月 22 日
	洪洞	2017 年 7 月 24 日—2017 年 7 月 28 日
江苏	南京	2016 年 4 月 11 日—2016 年 4 月 15 日
	苏州	2016 年 6 月 13 日—2016 年 6 月 16 日
甘肃	兰州	2016 年 9 月 19 日—2016 年 9 月 23 日
浙江	绍兴	2016 年 10 月 24 日—2016 年 10 月 28 日
	宁波	2017 年 3 月 27 日—2017 年 3 月 31 日
	温州	2017 年 4 月 24 日—2017 年 4 月 28 日

	杭州	2017 年 6 月 19 日—2017 年 6 月 23 日
福建	福州	2017 年 2 月 20 日—2017 年 2 月 24 日
	泉州	2017 年 4 月 10 日—2017 年 4 月 14 日
	三明	2017 年 6 月 26 日—2017 年 6 月 30 日
上海		2017 年 3 月 6 日—2017 年 3 月 10 日

在前期调研的基础上，完成了 4 个地方志可视化脚本的创意设计与素材采集，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。基于可视化脚本创意设计与素材，项目组完成了地方志资源的可视化实现。

项目组完成现存清康熙时期纂修方志目录数据 1525 条，其中包括区域志目录数据 1410 条，专志目录数据 115 条。目录数据包含题名、卷数、著者、版本、藏地、残缺、备注等字段，如表 2 所示。

表 2 清康熙时期纂修方志目录数据样例表

字段	**著录内容**
序号	1
题名	[康熙]順天府志
卷数	八卷
著者	（清）張吉午纂修
版本	清康熙間抄本
藏地	國圖
残缺	存卷二至八
备注	

2.2 地方志数字化规范研制

项目组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。上述规范作为汉王科技股份有限公司的企业标准，在企业标准信息公共服务平台（http://www.cpbz.gov.cn/）发布，如表 3 所示。

表 3 标准规范成果表

标准名称	标准号	发布时间
地方志文本数据规范	Q/HWSZT0005-2017	2017 年 1 月 22 日
地方志语料数据规范	Q/HWSZT0006-2017	2017 年 1 月 22 日
地方志卷目数据标引规范	Q/HWSZT0001-2017	2017 年 2 月 6 日
中国古今地名数据描述规范	Q/HWSZT0002-2017	2017 年 2 月 6 日
文字认同描述规范	Q/HWSZT0003-2017	2017 年 2 月 6 日
汉字集外字描述规范	Q/HWSZT0004-2017	2017 年 2 月 6 日
地方志图像数据规范	Q/HWSZT0007-2017	2017 年 8 月 11 日
地方志元数据规范	Q/HWSZT0008-2017	2017 年 8 月 11 日

在 8 个地方志数字化加工规范中，项目组选取了地方志文本数据规范、中国古今地名数据描述规范、文字认同描述规范和汉字集外字描述规范申报文化行业标准。文化部图标委专家认为，上述 4 个规范适用于或基本适用于汉文古籍，建议调整规范名称和适用范围。如表 4 所示。

表 4 标准规范成果表 2

序号	立项时间	计划编号	项目名称	标准性质	制/修订	完成年限	技术归口单位	起草单位
1	2016 年	WH2016-01	汉语文古籍文字认同描述规范	推荐	制定	2 年	全国图书馆标准化技术委员会	国家图书馆
2	2016 年	WH2016-05	中国古今地名数据描述规范	推荐	制定	1 年	全国图书馆标准化技术委员会	国家图书馆
3	2017 年	WH2017-03	汉文古籍集外字描述规范	推荐	制定	1 年	全国图书馆标准化技术委员会	国家图书馆
4	2017 年	WH2017-04	汉文古籍文本数据规范	推荐	制定	1 年	全国图书馆标准化技术委员会	国家图书馆

目前，“中国古今地名数据描述规范”已在网上公示，“汉语文古籍文字认同描述规范”在进行专家审核，“汉文古籍文本数据规范”和“汉文古籍集外字描述规范”在进行草案修订。

项目组依据地方志数字化规范，完成图像自动拼接软件、集外字描述软件、语料数据加工软件和古今地名数据加工软件。

2.3 语料库、本体库研制

项目组依据地方志语料数据规范，构建地方志语料库，处理地方志 34 种 57595 叶，完成语料数据 2710 万字。

语料库建设分为两个阶段：第一阶段，经过数字化系统生成地方志文本数据；第二阶段，在地方志文本数据基础上引入地方志目录、凡例等信息，生成地方志语料数据。

第一部分，地方志文本数据建设中，经过排版工具使用人工对切分、识别、校对后的文字和图像对照原图进行排版，使得结果数据中文字和图像的大小，位置等与原图在最大程度上保证一致，结果数据遵循《地方志文本数据规范》，内容包括文本数据格式、文字描述、版式描述、图形图像描述等。该结果数据适用于各类古籍文图内容数字化加工结果的描述的规范，可以还原古籍原貌，支持全文检索所需信息，适合古籍文献长期存储与使用。

第二部分，根据地方志文本数据和地方志特有的目录、凡例、样式等对文本数据进行重组，生成语料数据，内容包括文本、文字描述、版式描述、图形图像描述等。描述碎片化后的地方志数据的规范。语料数据将每本书组织成一棵有逻辑层级关系的 xml 树，如图 1 所示。

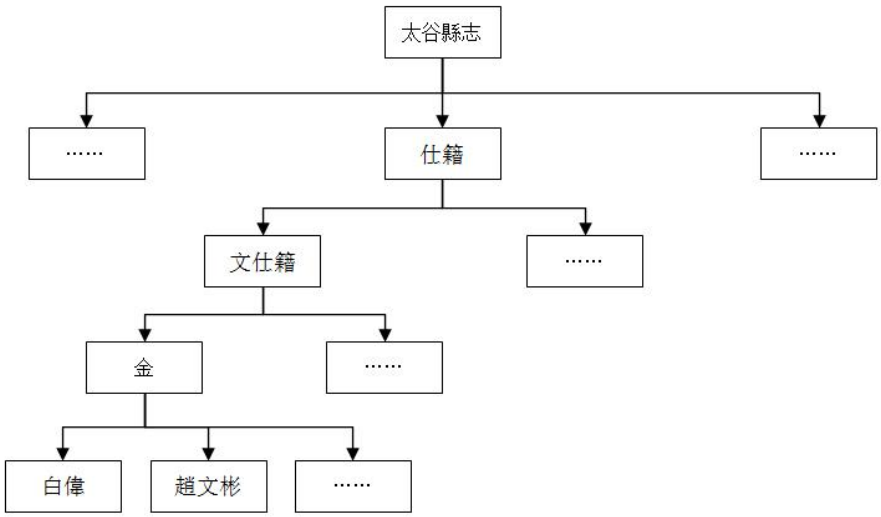

图 1 语料数据结构示意图

项目组完成时间本体库 1 套，以年为单位的数据起于公元前 1046 年，止于公元 2100 年，共 3146 条；以日为单位的数据起于公元 1 年 2 月 11 日，止于公元 2100 年 12 月 31 日，共 766970 条。参考陈垣所著的《二十史朔闰表》进行本体实例加工，实现了从汉高祖元年

起，至 1949 年建国前的一套中历、西历、回历、罗马历的时间对照信息。并以此为基础建立了时间本体库，包括时间的概念、关系、规则，该本体库是目前时间跨度最大、最符合地方志时间抽取的时间本体库。为了便于对 2000 多年间，中、西、回、罗马四种历法相互转换的使用，开发了时间转换工具，该工具在古籍研究过程中，可以在时间方面有一个横向的比较，对古籍中的时间研究有所助益。

项目组完成地名本体库 1 套，地名数据 64328 条，沿革数据 172794 条。

参考上海辞书出版社的《中国古今地名大词典》进行本体实例加工，首先完成了《中国古今地名大词典》数字化，共计 19667000 字，包含古今地名 64328 个，其中古地名 9541 个、今地名 52020 个、旧地名 2767 个。沿革数据 172794 条。然后将每个地名按照二级义项进行切分，每个地名的一个二级义项作为一条；总结所有的地名类别属性，总计 62 种地名属性，并进行地名类别属性归纳。最后完成地名本体库概念、关系和规则的建立，并可以根据地名本体库对地方志中的地名信息进行抽取。

2.4 地方志文献数字化技术研究

在地方志数字化软件开发中，引进流水线思想，开发专业化工具，对地方志数字化过程中的图像处理、版式分析、文字识别、文字校对、集外字处理、数据标引各个功能，分别开发对应的工具。每个工具功能单一，明确，易于操作，可以提高效率。根据功能将数字化加工软件划分成管理端和客户端两部分：管理端负责处理角色、流水线、数据包等生产相关的信息；客户端处理图像、版式、文字、排版等业务相关的信息。

系统通过角色管理模块、基本工序维护模块、流水线管理模块等实现生产流水线管理和数据包流转管理的功能。角色管理模块可以对系统中的角色信息进行操作，包括：查询、添加、修改、 删除、管理端授权、客户端授权、配置人员。通过对不同角色赋予不同权限，再为操作员配置不同角色，以达到每个人有不同技能，不同权限的目的。流水线管理模块主要完成流水线的添加与配置，包含查询、添加、修改、删除、启动、暂停、完成、人员配置、查看详细、备份和强制删除等功能。工件管理模块的作用在于管理系统中工作包流转过程中生成的所有工件，包括工件的查询、返工审核、删除、查看相关工件、导入导出和分篇导出等功能。统计信息模块负责对工作量、返工、派工、抽检结果、工资等信息进行统计，具体可以统计的信息见上图菜单，所有的统计信息均可以导出为Excel。

地方志项目的流水线由导入、版式、自动切分、切分校对、自动识别、文字校对、集外字编辑、PDF排版以及成品导出等一系列工具组成。图像处理工具用于文字识别前的图像预处理，版式工具用于文字识别前的版式自动识别与手工调整，切分校对工具用于文字识别前的版式切分的校对与修改，文字校对工具用于文字识别后的文字校对与修改，集外字编辑工

具用于文字识别后的集外字处理，PDF排版工具用于文字识别后的版式处理与修改。

2.5 地方志知识抽取技术研究与实现

本项目采用基于规则的方法实现地方志领域知识抽取。首先构建知识本体，本体内部包括概念和模板，并且每类知识对应几个模板。然后对知识本体进行解析，在此基础上对地方志文本进行知识抽取，得到文本所包含的元数据及其类别。最后对初步抽取的结果进行补全处理和指代消解，并用于完善本体库。地方志知识抽取系统如图2所示。

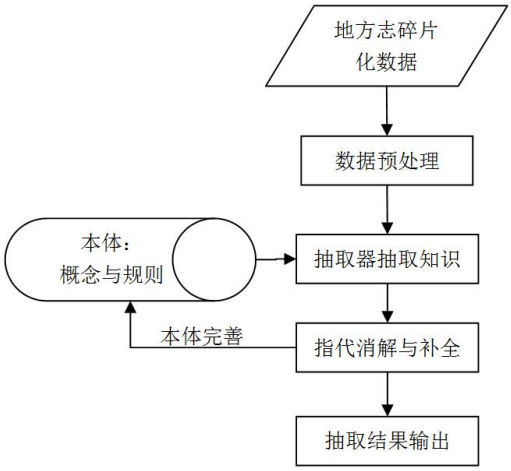

图 2 知识抽取流程图

基于语义的分析器根据知识本体和规则，基于一定程度的语义理解，快速地从文本中抽取出相关的元数据，并且需保证准确性、全面性及效率。本项目对34种地方志的数据碎片进行数据抽取，得到时间、地点、人物、事件等地方志元数据。

为评价知识抽取的结果，人工标记了[康熙]《介休县志》、[嘉庆]《介休县志》、[乾隆]《介休县志》三部方志，并与自动抽取的结果进行对比，结果如下：

- 按条目统计：

  地点：

  标记地点总数：988

  抽取地点总数：983

  正确抽取地点：905

  准确率：0.9206510681586979

  召回率：0.9159919028340081

  F值：0.9183155758498224

  人物：

标记人物总数：10419

抽取人物总数：10418

正确抽取人物：7755

准确率：0.7443847187559992

召回率：0.7443132738266628

F值：0.7443489945769544

事件：

标记事件总数：104

抽取事件总数：104

正确抽取事件：92

准确率：0.8846153846153846

召回率：0.8846153846153846

F值：0.8846153846153846

- 按属性统计：

地点：

标记属性总数：3952

抽取属性总数：3932

正确抽取属性数量：3693

准确率：0.9392166836215666

召回率：0.9344635627530364

F值：0.936834094368341

人物：

标记属性总数：52095

抽取属性总数：52090

正确抽取属性数量：50990

准确率：0.9788827030140143

召回率：0.9787887513197043

F值：0.9788357249124153

事件：

标记属性总数：208

抽取属性总数：208

正确抽取属性数量：199

准确率：0.9567307692307693

召回率：0.9567307692307693

F值：0.9567307692307693

2.6 方志文本数据、知识库数据等示范数据建设

项目组完成地方志图像数据采集 2 种 6206 筒子页，依据地方志文本数据规范，全文化 2 种 6206 筒子页，文本整理 32 种 51389 筒子页，如表 5 所示。

表 5 示范数据加工表

编号	题名	卷数	叶数	数据加工方式
1	新校天津卫志(康熙)	5	133	文本整理
2	浒墅关志	21	206	文本整理
3	浙江通志(康熙)	51	3415	扫描/全文化
4	灵寿县志（康熙）	10	224	文本整理
5	山东通志(康熙)	64	1785	文本整理
6	河南通志（顺治）	50	2049	文本整理
7	河南通志（康熙）	50	2214	文本整理
8	新修南乐县志(康熙)	2	198	文本整理
9	山西通志（光绪）	185	8351	文本整理
10	太原府志（万历）	5	241	文本整理
11	太原府志（乾隆）	60	1700	文本整理
12	太谷县志（乾隆）	8	552	文本整理
13	太谷县志	8	674	文本整理
14	介休县志(康熙)	8	310	文本整理
15	介休县志(乾隆)	14	544	文本整理
16	介休县志(嘉庆)	14	642	文本整理
17	介休县志	21	296	文本整理
18	汾州府志(乾隆)	34	1096	文本整理
19	平阳府志(雍正)	37	1864	文本整理

20	临汾县志(康熙)	9	271	文本整理
21	临汾县志	7	576	文本整理
22	洪洞县志	19	930	文本整理
23	陕西通志（康熙）	32	3548	文本整理
24	徽州府志(康熙)	18	1235	文本整理
25	琅盐井志	5	157	文本整理
26	云南通志	31	2190	文本整理
27	康熙台湾府志	15	458	文本整理
28	湖广通志（康熙）	81	2791	扫描/全文化
29	袁州府志（康熙）	21	1023	文本整理
30	大清一统志	356	11676	文本整理
31	鄢署杂钞	14	240	文本整理
32	江南通志（康熙）	72	3084	文本整理
33	增广洪洞古大槐树志	4	97	文本整理
34	大明一统志	90	2825	文本整理
合计		**1421**	**57595**	

项目组完成知识库一套，其中包括人物知识、地点知识、事件知识、物产知识。知识库的建立基于地方志语料数据，抽取其中对地方志研究非常重要的人物、地名、事件、物产信息，形成地方志知识库，可以供地方志领域专家查询与使用。

抽取人物数据155455条，抽取内容为：所属书目、类目 、人物名称、字、号、人物描述、籍贯、相关地名等。抽取地点数据117452条，抽取内容为：所属书目、类目 、地名、又名、地名描述 、方向、距离等。抽取事件数据8331条，抽取内容为：所属书目、类目 、时间、事件描述、相关地名等。抽取物产数据8521条，抽取内容为：所属书目、类目 、物产名称、详细信息、相关地名等。

2.7 地方志可视化技术与实现研究

2.7.1 地方志知识库可视化

地方志知识库包含人物、地点、事件、物产等数据，如图 3 所示，项目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，实现地方志知识库可视化。

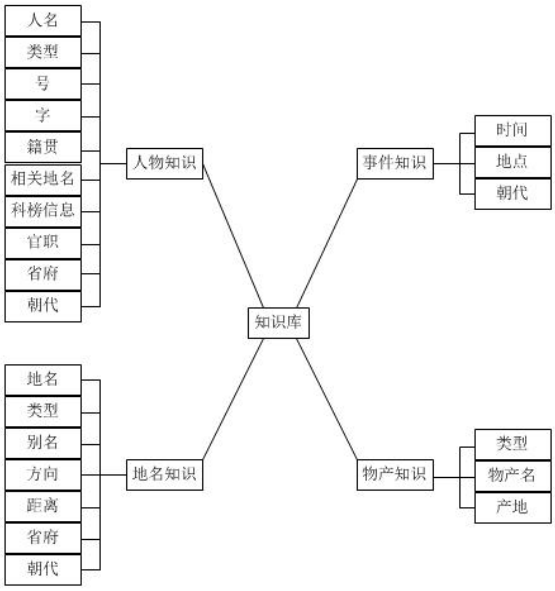

图 3 地方志知识库基本结构图

地方志知识库可视化的重点是数据间关系的可视化，项目组在开发过程中以 JavaScript 语言为基础，引入了 ECharts 插件，结合 Ajax 异步调用方式动态读取数据库，将数据信息用可视化的图形界面展示在前台，实现了对数据体系关联监控的可视化开发工作，并能够通过良好的界面达到与用户友好交互的目的。

以地方志人物关系的可视化为例，基于 Echarts 的地方志知识可视化展示效果如图 4 所示，对《大清一统志》中清朝康熙时期山西地区的人物关系数据进行系统和全面的分析，通过横向、纵向、多维度等方面进行比较，生成相关的可视化关系图谱，根据关系图谱反映人物关系及结果，例如黄廷栢、王之儀、周士章等同僚关系，点击相应人物则会显示人物的详细信息。通过地方志知识库可视化，可提供直观、生动、可交互、可高度个性化定制的可视化数据，图形化的表现形式更加一目了然， 也能够让用户更加清楚、形象、直观地了解到所查询的信息，从而提升用户体验。

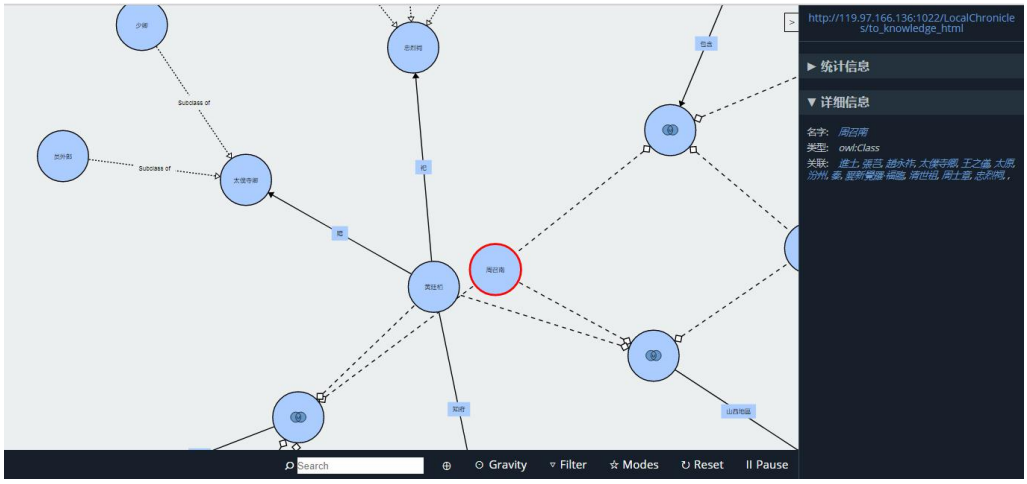

图 4 地方志人物关系可视化效果图

2.7.2 地方志脚本可视化

目前交互性、丰富性较强的地方志数字资源多是采用 flash 技术实现，但是跨平台性和交互性表现不佳。本项目基于 HTML5、CSS3、JavaScript 及相关技术实现地方志资源的交互式跨平台可视化。地方志资源的可视化过程，首先根据专家提供脚本（功能需求和内容设计），对文本、图片、音频、视频与动画等多种类型的地方志资源进行信息同步与图文混排，然后利用 JavaScript 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，并发布到用户终端上，即地方志脚本可视化。

基于 HTML5 和 JavaScript 的可视化系统主要包括：资源处理层、资源设计层和资源实现层，如图 5 所示。资源处理层是将传统的地方志资源进行数据化处理，主要包括首先对传统资源进行收集、整理、组织，然后将所需的传统资源数字化处理成相应的可用的数据资源。资源设计层主要是对处理好的数据资源的呈现进行设计，具体分为三个模块设计：信息模块设计，布局模块设计，交互模块设计。方志目录导航的设计需要以符合方志主题的目录的形式将各章节的主题内容和页码清晰的展示出来，使读者能直观的看出方志内容，跳转到自己想要阅读的章节。地方志的内容信息比较多样化，涵盖范围比较广，在对内容信息呈现形式设计时要根据所呈现的具体信息内容特点按照人们最能直观理解的方式和思维习惯进行设计，内容以图像、影像、声音、文字等多媒体形式相结合的方式进行呈现，使读者较为全面、直观地了解所反映的文化信息，同时提高他们的兴趣。资源实现层主要是根据相应模块的设计，选取相应的技术，对 pc 端和移动端分别进行设置，对各模块进行技术实现。

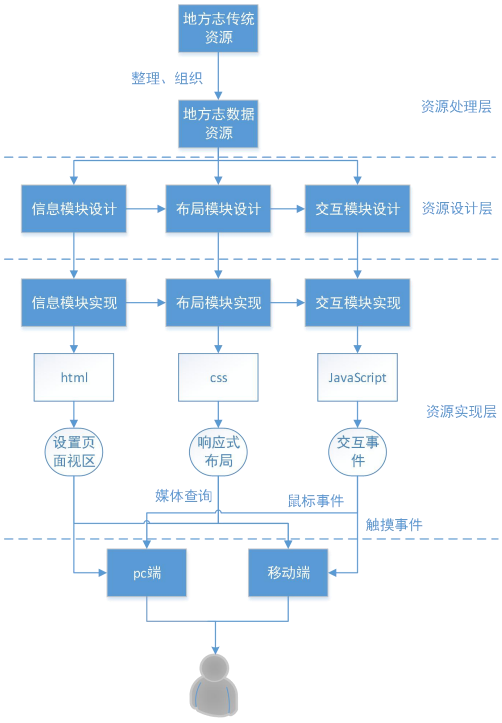

图 5 基于 HTML5 和 JavaScript 的可视化系统架构图

基于上述可视化系统，本项目采用电子书、游戏、动画、集成多媒体等方式完成了 4 个地方志脚本的可视化。以“山西介休与介子推”，基于 Epub3.0 标准，运用 HTML5 和 JavaScript 技术，设计并制作了地方志电子书，脚本可视化效果如图 6 所示。

图 6 脚本可视化效果图

2.8 基于 GIS 的地方志信息系统及示范数据建设

项目组基于栅格地图服务技术开发了 WebGIS API，如图 7 所示，同时在 WebGIS API 的基础上构建了地方志 WebGIS 展示系统。为了适应方志展示需要，项目组在地图制作过程中做了一些特殊处理，比如底图风格采用做旧处理，地图层级设置 14 级比例尺，方便用户浏览不同粒度的数据，山西晋中、介休等地采用精确到村庄切图，为方志及古今地名匹配提供参考点。在开发 API 的过程中，项目组借鉴了部分优秀开源框架的设计风格，比如 openlayers，leaflet 等，同时结合地方志特定设计了一套 API，其中包括常用的地图缩放，漫游，量测，拾取，标注，绘制等，也提供了对地方志特有的信息查询服务。

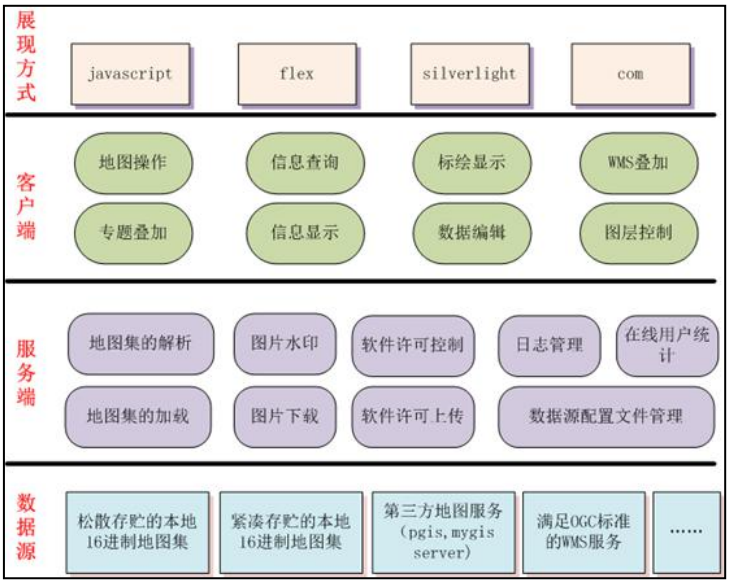

图 7 WebGIS 地图应用 API 架构图

地方志 GIS 数据采集与编辑系统是基于地方志 GIS 系统 API 接口的一种应用，由于 API 采用 JavaScript 编写，在页面代码中添加相应的 JavaScript 代码即可实现远程调用函数并显示效果。基本原理是在基础地图图层上，从数据库中将用户需要的数据抓取出来，并通过相应的方法对覆盖物图层或其它主题图层进行操作，形成由多个图层叠加而成的图像，针对不同的功能，调用相应的接口函数完成业务流程，具体描述如下：

查询：采用 sql 语句搜索数据库空间数据库的古今地名表，并在地图上添加 Marker 标记，跳转至该地，用户点击后从名称、简介、类型和位置四个方面展示该地的信息；

修改：修改功能是在地图上新增图层，将文本框、图片等元素在新图层中作为 html 元素进行排列呈现，用户可在文本框中进行修改操作，完成操作后点击修改按钮即可完成修改；

添加：用户点击地图后，后台会记录下该点的空间数据，此时用户在新增图层页面中将该点的简介信息添加完毕后保存即可。

基于 GIS 的地方志数据展示系统的实现是在上述地方志 GIS 系统的设计与实现基础上进一步的开发扩展，该部分主要是对方志的内容进行研究分析，做到可视化的效果。根据地方志文献内容的分析，其包含的信息有政治、经济、军事、文化等多个方面，如何将这么多信息进行有序、多维地向用户呈现，是一项极其重要和复杂的工作。在技术上，地方志 GIS 采用了栅格地图进行地图的展示，基于地图的二次开发接口，可以采用 JavaScript 将地方志记载的事件等信息呈现在地图上。

在内容呈现上，项目组将地方志知识库中的数据以合适的方式在 GIS 上呈现。以物产为例，以种类、地区为查询条件，展现出不同种类物产的空间分布及方志来源，点击查询结果列表，地图定位到物产具体分布区域并以气泡框显示物产详情以及方志来源。

图 8 地方志物产数据可视化效果图

2.9 地方志数据演示平台设计并实现

在地方志示范数据的建设上，项目组从地方志图文对照呈现、地方志语料库、地方志知识库与知识融合、WebGIS 展示数据和地方志可视化脚本几个方面进行了相应的建设工作。

演示数据建设包括清康熙时期纂修方志目录数据 1525 条；地方志图像数据 34 种 57595 筒子页；文本数据 34 种 57595 筒子页；古今地名规范数据 10023 条；语料数据 27101366 字，另外完成了山西介休与介子推多媒体电子书、山西洪洞与大槐树动画、山西太谷与秧歌音视频、山西介休与方言互动游戏和山西永济鹳雀楼与王之涣三维虚拟漫游场景。

项目组在对上述地方志资源进行梳理和分析后，设计了地方志可视化演示、地方志资源管理、地方志资源知识融合可视化和 WebGIS 展示四大模块共 15 项子功能，如图 9 所示，并采用了弹性搜索、分布式存储、缓存等技术研制了一套地方志可视化演示平台。在性能上，该平台支持多种多版本的浏览器；支持在低性能环境下的运行；支持包括地方志文本、图片、视频和元数据等多维度地方志数据的导入和管理功能；支持地方志多媒体数据的可视化展示功能。

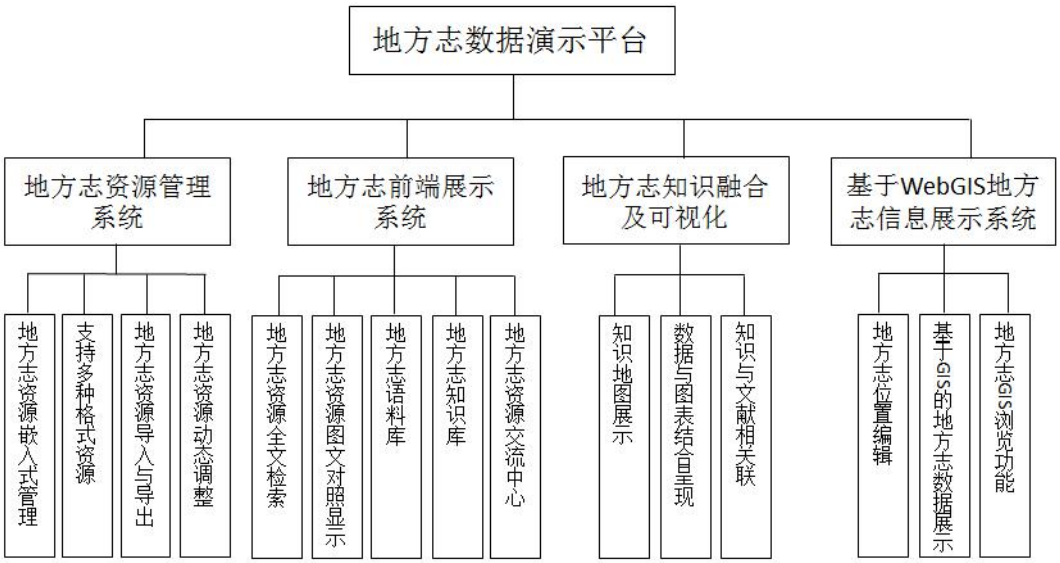

图 9 地方志数据演示平台功能架构图

地方志演示平台的架构设计考虑到地方志数据在当前环境下，数据量会呈现 PB 级规模，数据的呈现形式不再是单一的关系型数据，用户数量会越来越多，存在高并发，高访问量的请求。基于此，架构设计的耦合度要非常低，能够承受的访问压力要很大，数据的存储空间要易于扩展，架构要具备良好的可扩展性。在数据存储方面，采用典型的关系数据库 MySQL 和非关系数据库 MongoDB 以及 Hadoop 的文件系统来共同存储地方志相关资源。数据存储具备良好的可扩展性。地方志数据可视化服务和地方志数据编辑服务分开搭建，缓解服务器压力，服务器可选 Nginx 来达到负载均衡，可视化服务在处理请求量过大时，可以搭建消息队列比如 Celery 来异步处理消息，例如邮件服务对实时性要求不高的服务。如图 10 所示。

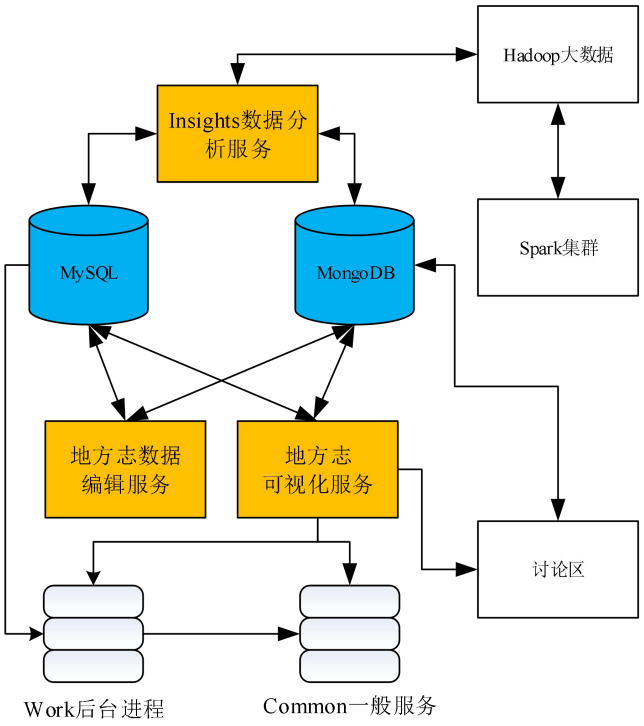

图 10 地方志数据演示平台网络部署图

地方志前端展示系统的全文检索界面如图 11 所示，地方志图文对照显示效果如图 12 所示、地方志语料库显示效果如图 13 所示。

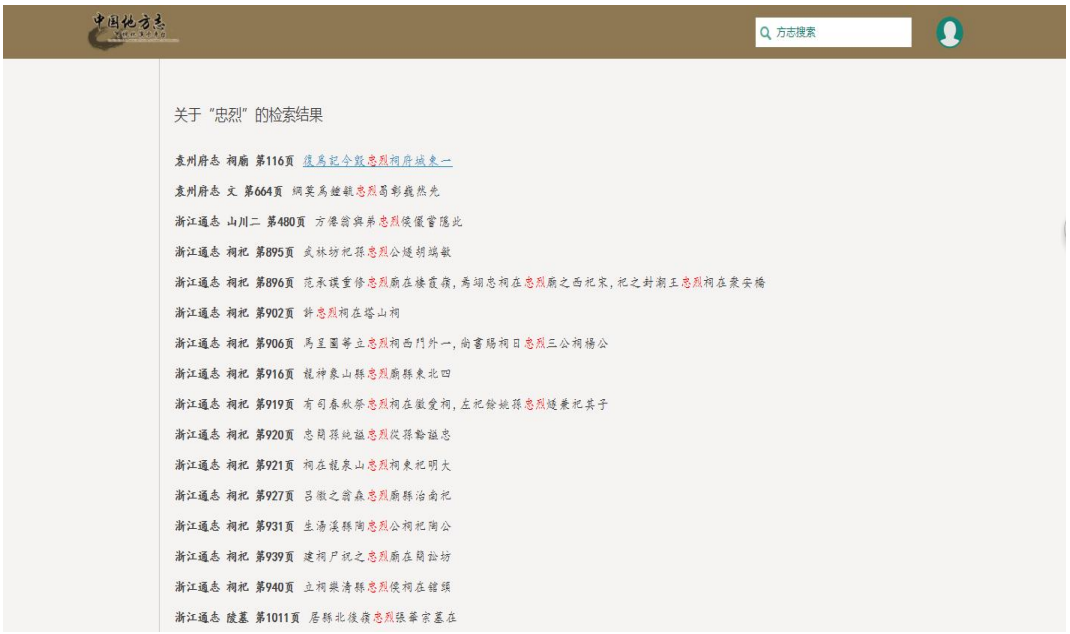

图 11 地方志前端展示系统界面 1

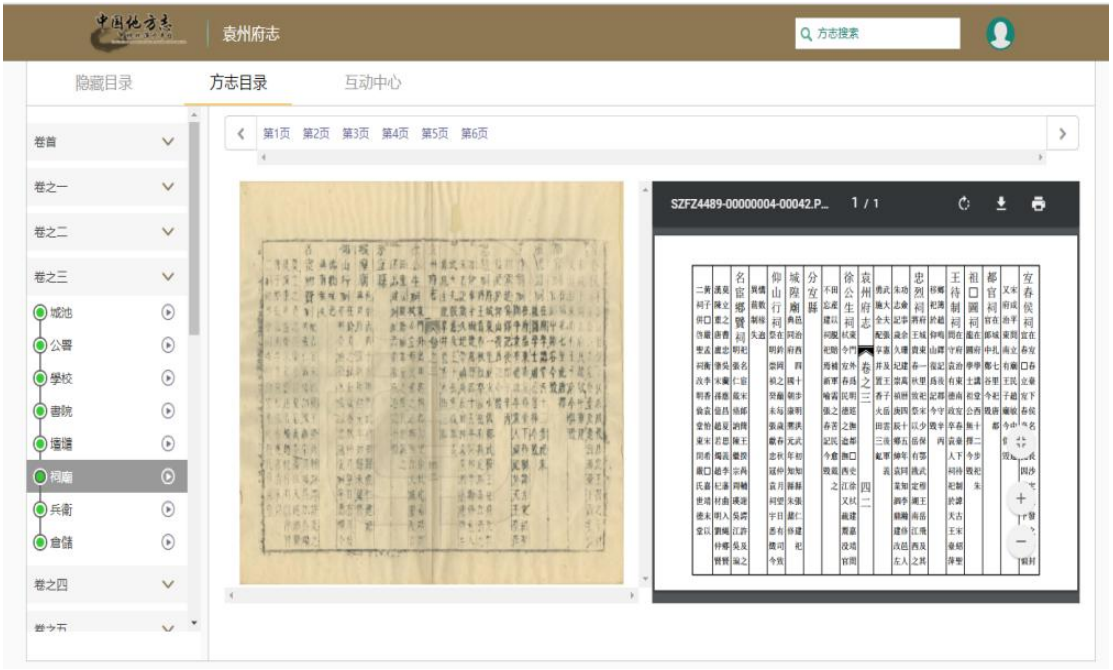

图 12 地方志前端展示系统界面 2

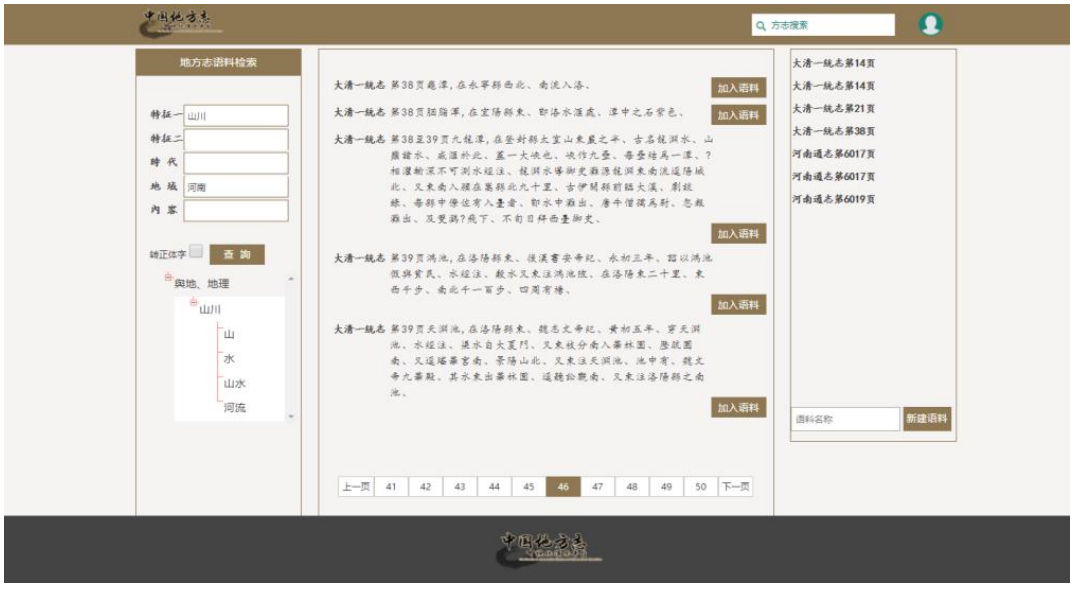

图 13 地方志前端展示系统界面 3

3. 已发布成果

在研究过程中，项目组积累了大量的技术、方法和经验，经过系统梳理后，针对不同的研究方向，申请了 13 项国家发明专利，如表 6 所示。

表 6 申请专利情况表

序号	申请号	专利名称
1	201510696674.4	版面分析方法和装置
2	201511009801.5	文字检索方法及装置

3	201611223323.2	图像处理系统及图像处理方法
4	201611225709.7	文字检索方法及文字检索装置
5	201710283168.1	一种异形字符输入方法、装置及电子设备
6	201710283791.7	一种信息处理方法、装置及电子设备
7	201610063529.7	一种知识地图映射生成方法
8	201710608341.0	一种地方志知识融合方法
9	201710608336.X	一种地方志资源的富媒化制作方法
10	201710607058.6	一种地方志资源跨平台可视化方法
11	201710608340.6	一种基于WebGIS的地方志文献可视化方法
12	201710608338.9	一种基于地方志研究的搜索优化方法
13	201710608348.2	一种面向地方志网站的混合推荐系统

在研究过程中，项目组，获得了 9 项计算机软件著作权，如表 7 所示。

表 7 获得计算机软件著作权情况表

序号	登记号	软件名称
1	2017SR692534	汉王数字方志数字化软件
2	2017SR692534	汉王数字方志知识抽取软件
3	2017SR664984	地方志知识可视化系统
4	2017SR666196	地方志资源互动系统
5	2017SR666201	介休方志可视化展示系统
6	2017SR663881	中国地方志可视化演示平台
7	2017SR680856	基于WebGIS的地方志信息展示系统
8	2017SR680886	WebGIS地图服务系统
9	2017SR680863	栅格地图生成系统工具

在研究过程中，项目组出版专著 6 部，发表论文 19 篇，如表 8 所示。

表 8 论著论文表

序号	成果类型	题名	完成情况
1	专著	IDS 与集外字处理方法研究	上海远东出版社（2017 年 3

			月）
2	专著	古籍文本数据格式比较研究	上海远东出版社（2017 年 4 月）
3	专著	方志文献特性与数据抽取研究	上海远东出版社（2018 年 1 月）
4	专著	清代顺治康熙时期地方志编纂研究	上海远东出版社（2018 年 4 月）
5	专著	地方志数据加工规范研究	学苑出版社（2017 年 12 月）
6	专著	地方志数据加工规范应用指南	学苑出版社（2017 年 12 月）
7	论文	云平台下基于隐私保护的桶划分方案	《计算机学报》（2016 年第 2 期）
8	论文	论面向读者需求的数字方志建设策略	《图书情报导刊》（2016 年第 4 期）
9	论文	国家图书馆藏《河北地理杂抄》小议	《中国地方志》（2016 年第 9 期）
10	论文	古籍索引数据应用研究	《新世纪图书馆》（2017 年第 5 期）
11	论文	地方志资源聚合方法与实现	《国家图书馆学刊》（2018 年第 2 期）
12	论文	面向方志文本的可视化应用研究	《国家图书馆学刊》（2018 年第 2 期）
13	论文	一种地方志资源的的混合推荐模型	《国家图书馆学刊》（2018 年第 2 期）
14	论文	Blind deconvolution using the similarity of multiscales regularization for infrared spectrum	Measurement Science & Technology, 26(11), 115502.
15	论文	A content-based recommendation algorithm for learning resources	Multimedia Systems, pp. 1-11, 2017.
16	论文	Blind image restoration with sparse priori regularization for passive millimeter-wave images	Journal of Visual Communication and Image Representation, vol. 40, Part A, pp. 58-66, 10// 2016.
17	论文	KDE based outlier detection on distributed data streams in multimedia network	Multimedia Tools and Applications, 2016: 1-19.
18	论文	Towards a resource migration method in cloud computing based on node failure rule	Journal of Intelligent & Fuzzy Systems, 2016, 31(5): 2611-2618.

19	论文	Blind Spectral Signal Deconvolution with Sparsity Regularization: An Iteratively Reweighted Least-Squares Solution	Circuits Systems and Signal Processing, 2016: 1-12
20	论文	Infrared spectrum blind deconvolution algorithm via learned dictionaries and sparse representation	Applied Optics, vol. 55, no. 10, pp. 2813-2818, 2016.
21	论文	A learning mode of interactive autonomy based on the smart learning environments	International Conference on Advanced Technologies Enhancing Education.2017.
22	论文	DBNCF: Personalized Courses Recommendation System Based on DBN in MOOC Environment	Educational Technology (ISET), 2017 International Symposium on. IEEE, 2017: 106-108.
23	论文	Interactive Learning Resources Based on Cognitive Load: Design and Application	Educational Technology (ISET), 2017 International Symposium on. IEEE, 2017: 208-211.
24	论文	Research on Rubbing Image Mosaic Based on SIFT Feature	2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 29-31 July 2017, Guilin, China.
25	论文	Harris Corner Detection based Leaf Image Segmentation for Ancient Chinese Books	2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI 2017), 14-16 October 2017, Shanhai, China.

4. 项目成果应用前景

国有史，地有志，家有谱。中国地方志是一种珍贵的文献资源，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；地方志内容极为广泛，且成系统，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，所录资料既要丰富，又要实事求是严加考证，去伪存真，人、时、地、事无差错，达到资料翔实。

本项目将地方志语料地方志文献转换成目录、图像、文本、知识数据等不同粒度的数据，并与 GIS 系统相结合，实现时间、空间、文献三个维度的智能检索、数据分析和可视化显示。

基于地方志数字化技术和演示平台，可以将地方志中蕴含的地理信息、历史信息、文化信息、科技信息等进行高度整合，并以可视化方式显示，从而形成高质量的数字资源，为社会发展、学术研究、文化教育等提供强有力的支撑。

利用这样的平台，将现有方志深度整合，复原已经消逝的名胜古迹，让文化和旅游更好结合；还原一方青山绿水，因地制宜挖掘各地的物产，让人民享受富足美好的生活；研究一地的灾异和变迁，以便防灾减灾，让百姓生活更加安定；梳理一地的经济、制度，让各级管理机构决策更加科学；了解家乡的著名人物、科学文化，增强自豪感，让爱国爱家成为游子发自内心的情怀……如此典籍中的故事不仅真的活起来，还能激活焕发新的生命力，造福一方。当然实现这样的目标还需要进一步投入、支持。我们对此充满期待。