

多源异构下区域性专题数据库建设研究

——以京津冀协同发展专题数据库为例

陈晨

(中共北京市委党校图书馆/北京市市情研究中心, 北京 100044)

摘要: 通过探讨基于多源异构的区域性专题数据库的技术实现, 为图书馆开展数字资源服务提供新的思路与方法、满足相关用户的实际使用需求, 进一步完善资源建设制度。从区域性专题数据库建设现状、多源异构专题数据的需求设计与技术实现等角度出发, 全面介绍多源异构专题数据库的建设与实践过程, 并对数据库在实际运行中的效果进行分析与探讨, 提出专题数据库建设的若干问题与建议。中共北京市委党校图书馆(以下简称“北京市委党校图书馆”)自建的京津冀专题数据库具有用户体验友好化、功能设计灵活化、栏目划分清晰化等特点, 契合专题数据库建设服务要求, 内容上更贴近研究者的研究需求。

关键词: 多源异构; 区域性; 专题数据库; 京津冀

中图分类号: G250.7 **DOI:** 10.3772/j.issn.1673-2286.2020.12.005

引文格式: 陈晨. 多源异构下区域性专题数据库建设研究——以京津冀协同发展专题数据库为例[J]. 数字图书馆论坛, 2020 (12): 31-37.

在专题数据库建设过程中, 数据通常来源于不同的信息系统, 这些数据在数据命名、数据格式上都存在较大差异, 而这些数据支撑着不同的服务, 这就使得数据融合在综合性数据库建设中十分必要。在图书情报领域, 进行知识融合是为在大数据环境下对多源异构的知识资源进行科学、有效的利用, 最大化地挖掘知识价值, 知识融合的结果用来进一步作用于知识服务^[1]。

在数据库建设向着数据化、智慧化转型发展的时代背景下, 北京市委党校图书馆以承担中央党校四大专题数据库建设为契机, 进行京津冀协同发展专题数据库(以下简称“京津冀数据库”)的开发建设。本文以京津冀数据库的构建思路与关键技术实现为例, 探索基于多源异构数据来源的区域性专题数据库建库方式与方法, 总结建库过程中的不足与待改进之处, 以期达到数据库满足各类用户的实际需求的目标, 在丰富自身数字资源建设策略的同时为其他图书馆或相关机构开展数据库自建和研究提供参考。

1 区域性专题数据库建设的现状

1.1 区域性数据库建设现状

2018年,《中共中央 国务院关于建立更加有效的区域协调发展新机制的意见》指出, 推动国家重大区域战略融合发展, 以“一带一路”建设、京津冀协同发展、长江经济带发展、粤港澳大湾区建设等重大战略为引领, 以西部、东北、中部、东部四大板块为基础, 促进区域间相互融通补充^[2]。政府及科研机构陆续将区域性专题数据库建设作为重点工作。其中建设相对成熟的主要有“一带一路”专题数据库^[3]与长江经济带数据库^[4](见表1和表2)。

在多源异构数据库建设过程中, 有学者在理论层面提出实现异构多信息源集成的关键在于找到一个合适的公共数据模型^[5], 并研究了异构数据集成的3种方法, 即XML、数据仓库和Web Service。这3种方法直到今天仍然具有较高的参考价值。在知识融合领域, 唐晓波等^[6]提出基于两层知识融合的金融知识服

表1 “一带一路”专题数据库

数据库	网 址	访问性	内容特征
中国一带一路网	https://www.yidaiyilu.gov.cn/	免费访问, 免费下载	新闻、政策、报告等内容服务
国研网“一带一路”研究与决策支撑平台	http://ydyl.drcnet.com.cn/	免费访问, 免费下载	可视化呈现为主, 内容服务为辅
社会科学文献出版社一带一路数据库	https://www.ydylcn.com	免费访问, 付费下载	新闻、政策、报告等内容服务

表2 长江经济带专题数据库

数据库	网 址	访问性	内容特征
复旦大学长江经济带大数据平台	http://yreb.sozdata.com/	免费访问, 内网下载	政策文件等文献资料与可视化呈现的综合数据库
上海社会科学院长江经济网	https://www.yangtze.org.cn/	免费/收费访问, 未提供下载	新闻、政策导航为主

务模型, 包括知识获取层和知识处理层的知识融合, 这两种融合方式都是基于本体库的构建。多源数据由不同的用户和不同的来源渠道产生^[7], 这些数据通常存储在不同的数据库中, 包括Oracle、SQL Server、MySQL、DB2、Sybase等。多源数据融合主要涉及数据唯一识别、数据记录滤重、字段映射与互补、同名消歧、别名识别、数据加权等多个方面^[8]。

在实践层面, 余鹏等^[9]设计研究的高校科研服务系统通过遵循统一的数据标准来解决数据交换时带来的问题, 针对不同平台与不同语言带来的不规范数据调用问题。于亚秀等^[10]采用Web Service技术解决异构数据融合问题, 同时针对不同数据库中的数据需要对命名、类型等问题进行统一清洗, 确保数据的完整性、规范性。赵捷等^[11]提出可采用基于规则库和规范库实现多源异构元数据的统一规范。丁道劲等^[12-13]指出, 文献元数据集成管理需要针对既定多来源元数据标准格式, 制定统一的元数据标准和映射体系, 其中包括书目元数据、文摘元数据、关联元数据等, 并进一步提出通过数据转换、信息整合、语义关联对多源异构元数据进行深度整合。鲜国建等^[14]从实际操作层面根据NSTL数据本身的特点, 制定了多源异构数据匹配规则, 并通过数据采集、汇聚、匹配、融合等数据治理核心流程, 使多源异构文摘数据实现精准匹配和字段级融合。

1.2 京津冀专题数据库建设现状

2014年, 京津冀协同发展被列为重大国家战略后, 相关研究直线上升, 但没有成熟可用的综合性专题数据库, 主要的京津冀数据库研究成果有旅游类数据库、

气候类数据库、金融类数据库、人才类数据库、物流类数据库、教育资源类数据库, 且大多局限于理论层面, 可访问的数据库并不多。其中, 张婕^[15]针对体育信息资源共享平台建设提出了相对完备的京津冀专题数据库建库方案, 其建库定位、信息收集原则等内容都具有一定的现实指导意义。河北省千年古县特色数据库建设研究以京津冀协同发展为背景, 在明确的建设目标指导下, 提出了具体栏目建设规划^[16]。韩兆柱等^[17]提出京津冀整体性技术协调模式, 强调信息技术的整合、网络简化和一站式服务。

京津冀数据库以前人研究与实践为基础, 在内容收集与加工上保证文本、图片、音频、视频等多源异构数据统一规范标准, 并能与参建单位进行数据交换、共享。通过关键词表定期采集、匹配互联网资源信息, 整合网络资源和馆藏自有数据, 保障了数据库内容的全面性与针对性, 同时节省了建库成本与时间, 也更利于数据库后期维护的规范化与一致性。

2 京津冀数据库建设方案

2.1 建库思路

为贯彻落实《中共中央关于加强和改进新形势下党校工作的意见》和全国党校工作会议精神, 中央党校发布了《全国党校系统数字资源建设规划(2016—2020年)》, 提出要在我国“十三五”期间重点建设好党校系统的四大专题数据库, 北京市委党校图书馆承担了京津冀数据库建设的主要任务, 天津市委党校图书馆与河北省委党校图书馆协助参与数据库建设。

2019年10月发布的《中国共产党党校（行政学院）工作条例》中提到，党校（行政学院）应当重视图书馆（室）建设，加强图书文献和信息的采集、整理与开发，积极推进数字资源共建共享工作。根据已公开的京津冀相关数据情况，结合北京市委党校图书馆馆藏特色资源，北京市委党校图书馆京津冀项目团队及第三方公司开发团队决定以此为契机，开发建设区域性专题数据库，充分利用包括党校图书馆全文数据库（如皮书数据库等）、京津冀主题书籍资源等在内的三地馆藏资源、三地政府公开的政府文件、统计公报等开放数据，以重大事件为纲，借助时间轴形式显示京津冀发展至今的重要活动轨迹。在首都人才济济的基础上，充分发挥党校京津冀研究领域的专家优势，在数据库栏目制定及内容选择上引入专家观点，为数据库的权威性提供有力保障。考虑到数据库内容涉及京津冀三地庞大数据量的情况，确定以“总分库”的形式建设数据库，即以北京为主的京津冀总库，天津、河北为分库形式进行数据库底层设计。京津冀数据库建库目的是在为京津冀相关领域学者提供文献与数据参考的同时，也向普通大众用户传达京津冀协同发展战略的历史背景、发展脉络及未来方向，为党校数字图书馆智库建设添一份力。

在内容资源数字化方面，京津冀数据库力争深入到词语的粒度级别，对若干京津冀关键词建立词表。在内容呈现方面，京津冀数据库要体现更方便地获取资料、节省用户获取和利用信息资源的时间及精力的特点，并对相关经济指标信息进行可视化呈现。在数据服务方式与目标群体方面，京津冀数据库根据地域特点采取“总分库”的形式，专注于京津冀三地经济建设、政治建设、文化建设、社会建设和生态文明建设五位一体协同发展的信息采集与整合，旨在为不同学科背景的学者提供更直观、更深层次的数据服务，方便决策者一目了然地获取所需信息，辅助相关领域学者发现京津冀协同发展过程中的特征，节省学者前期海量数据的查阅、整理及统计指标数据的时间。

2.2 功能设计

京津冀数据库结合数据库建库特点，保障数据库资源类型的完备性、数据库栏目层级的清晰化、面向用户的易用性，在此基础上进行多样化功能设计。建库初期的目标为实现以下功能。

（1）“总分库”结构下的资源管理功能。京津冀资

源类型多样且分布广泛，数据库按“总分库”结构整合多源异构数据，包括文献资料、图片、音频、视频，全面支持各类资源信息的发布与管理，并可直接上传导入以下文件。①文献资料：txt文本、MS Office系列文件、WPS系列文件、pdf文件、ePub规范格式文件、xml格式文件包；②图片：兼容bmp、jpg、png、tiff等常见格式；③音频：兼容MP3及wav格式；④视频：兼容MP4及flv格式。如图1所示，系统针对北京市委党校图书馆存量资源和增量资源来源情况，依据国际通用标准和全国党校（行政学院）系统数据标准规范进行标引、入库、排重、收录等相关数据采集工作，在此基础上，支持多类型文献资料的聚类存储，支持图片库、文献库、电子图书馆、多媒体资料库、人物库等专题子库，同时各库之间相互关联，形成知识链，并提供相关附件免费下载。系统底层采用HDFS（Hadoop分布式文件系统）作为存储方案。异构数据主要来源于非结构化数据，包括资源文件、图片文件、标准文件等。对Web服务器来说，不论是Tomcat、IIS，还是其他容器，图片是最消耗资源的，因此系统采用将图片与页面进行分离的方案，这样的架构可以降低提供页面访问请求的服务器系统压力，并且可以保证系统不会因为图片问题而崩溃；同时这部分数据量较大，因此采用分布式文件系统HDFS作为存储方案。在内容组织上，京津冀数据库总库的前端页面资源内容均采用导航和聚类模式进行组织和呈现，内容组织聚类和导航支持后台灵活按需自定义。内容架构与河北省委党校建设的京津冀数据库-河北的内容架构可互联互通，形成总库与分库资源的有机交互。

（2）数据分类与标引。京津冀数据库对知识体系进行标引，结合阅读、检索等功能提供应用服务。通过词间关系自动构建、关联标识符编码嵌入等技术从大量京津冀相关文本数据抽取关键词，根据需要标引的结构化关键词的数量，选择使用人工或者半自动化知识标引，标引结果经过人工辅助复核，形成知识化资源。标引后的结构化数据为数据库实现一站式检索提供了保障，除标题（标题、眉题、副题）、作者、日期、摘要、关键词、来源、正文等常规检索字段外，还可提供热词、推荐词检索，保证检索结果快速响应的同时，内容资源保障查全查准。

（3）共建共享原则下用户及权限管理功能。京津冀三地党校在后台内容管理上互相独立，各自处理、加工、上传内容。用户和管理员可以通过网络在身份验证后按权限使用或管理数据库，管理员和用户权限可实现动态管理，同时资源使用权限可动态管理，实现京津

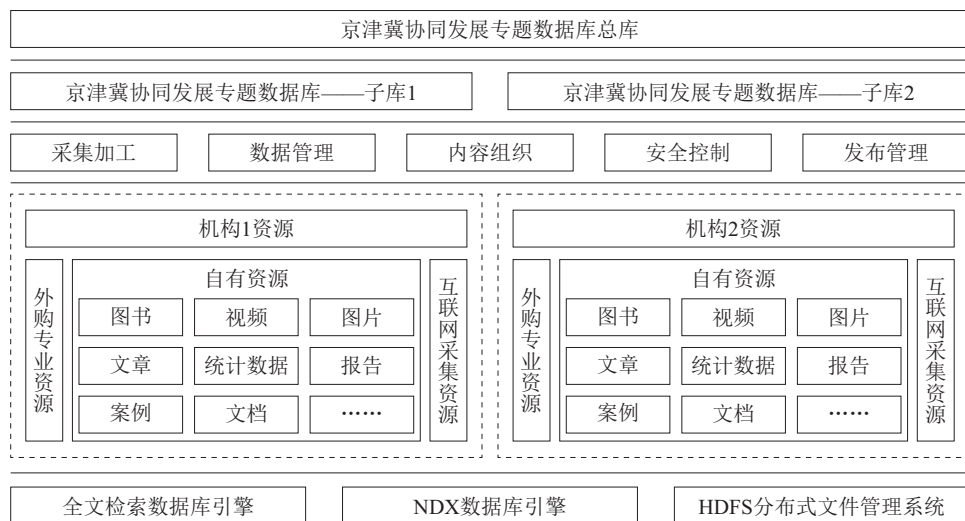


图1 专题数据库资源管理结构

冀三地数据库共建共享。

(4) 统计分析功能。其主要针对京津冀数据采集、存储、标引、发布过程日志信息进行统计分析,实现对于数据管理过程的有效监控。主要功能包括操作日志浏览、各类型资源存储量和发布量统计分析等。

(5) 前端用户服务系统。京津冀数据库前端界面栏目设计清晰明了、层次分明。在专家意见与项目组多次讨论会议后制定了常规栏目与特殊栏目。常规栏目展示以信息流为主,如高层关注、媒体聚焦、政策法规、专题研究、协同合作,再根据个别页面的特点配以个性化UI展示,如数据分析栏目以树状图、饼状图呈现出京津冀三地GDP、人口、CPI等经济社会指标数据,同时用户可以根据地域、指标类型、年份等维度自由组合来获取所需信息;发展大事记栏目打破常规信息罗列的呈现方式,按时间顺序筛选京津冀概念提出以来发生的重大事件,以时间轴形式展开,用户可以直观便捷地了解京津冀发展的历史脉络。在保证用户界面友好化的前提下,京津冀数据库开放京津冀相关电子书、期刊、论文、专家分析等数据(包括pdf、word、ppt等格式)的下载和数据分析图表的图像导出。

2.3 系统架构

考虑到数据库覆盖知识领域广阔,京津冀数据库针对三地不同的业务需求以及实现这些需求所需要的功能、软硬件环境、系统环境等,制定了不同的技术方案以满足不同的业务场景,同时详细分析不同技术方

案、仔细论证各个技术方案,以达到各技术方案完美融合,从而形成整个平台的完整技术方案(见图2)。

在设计系统架构时,充分运用先进性原则,预留足够的系统扩展空间,并提供丰富的接口,以便其他业务功能模块的快速调用。可以看出,底层数据来源多样化,同时在数据采集与结构化处理上,数据库充分做到异构数据融合。第一,所有数据的加工处理符合全国党校(行政学院)系统数字图书馆数据规范标准,确保中共北京市委党校(北京行政学院)京津冀数据库项目与中央党校(国家行政学院)图书和文化馆四大专题数据库的平滑对接和顺利迁移。第二,文献基本字段信息(如标题、作者、来源、出版日期、摘要、关键词等)可根据自建库需要灵活定制。

数据库前端发布页面系统采用HTML5实现,支持Web界面的操作,所有页面展现的资源内容组织均采用导航和聚类模式,检索系统提供全文检索,检索范围包括标题(标题、眉题、副题)、作者、日期、摘要、关键词、来源、正文等。在知识组织层面,支持多用户在线管理知识元的增删改查和知识元关联关系的增删改查及可视化呈现功能。系统提供知识元标引功能,对已经标引的知识元的修改、删除操作。

3 区域性专题数据库建设效果与展望

3.1 京津冀数据库建设效果

京津冀数据库通过人工与网络相结合的方式共



图2 京津冀数据库系统整体架构

收录2 384条数据；已将118本京津冀相关书籍加工成PDF；涵盖专题研究子栏目8个，分别为北京核心区、北京城市副中心、河北雄安新区、北京·张家口冬奥会、曹妃甸、大兴机场、通武廊、北三县。可分析的经济社会指标111个，关键词标引1 134条数据，人工标引后的关键词在数据库文章中匹配最多的前16个关键词如表3所示，可以看出，关键词基本覆盖京津冀区域专题数据库的发展方向。

京津冀数据库建库期间，项目团队多次向领域专家请教当前京津冀研究的热点与前沿，并根据反馈意见完善资源收集策略。在获得京津冀领域专家认可的同时，有学者提出数据库在重数据分析深度的同时应当引入国内经济圈与国际经济圈用于数据对比，查找差距。社会学专家建议数据库应该配以京津冀三地相关领导足迹，有利于研究者快速索引、查找对应资料。

截至2020年底，数据库建设取得初步成效。数据分析栏目已经将京津冀三地相关经济、社会、人口等领域的指标导入完成，为保证指标数据权威且能反应经济社会发展的相关情况，指标数据来源主要为三地年度统计公报以及国家统计局。如图3所示，所有指标覆

盖京津冀五位一体全面发展的各个方面，其中，地区生产总值类指标9个、常住人口类指标8个、财政类指标6

表3 京津冀主题部分关键词

关键词	文章匹配次数/次
京津冀	151
京津冀协同发展	133
京津冀一体化	76
协同发展	70
京津冀地区	40
城市群	31
京津冀区域	27
区域经济一体化	21
雄安新区	19
功能定位	18
长江中游城市群	16
粤港澳大湾区	16
中原城市群	16
区域发展	14
区域经济	13
国家战略	6

个、居民消费指数类指标7个、农业类指标6个、工业类指标9个、交通运输类指标13个、金融类指标11个、固定资产投资类指标5个、市场消费类指标4个、对外经济类

指标4个、旅游类指标6个、交通类18个、教育类28个、资源和环境类17个。



图3 数据分析子栏目概况

专题研究8个子栏目共收录文章条目269条,如表4所示,其中,河北雄安新区数据量相对较多,可以看出河北雄安新区与京津冀协同发展在国家未来战略发展中的重要地位。通过详细对比各子专题内容发现,区域协同发展主要集中在交通运输、人口等重要领域的合作共赢,通过加强区域间的合作交流,实现经济社会的高质量发展。

表4 专题研究子栏目概况

专题名称	文章数量/条
北京核心区	6
北京城市副中心	25
河北雄安新区	146
北京·张家口冬奥会	17
曹妃甸	50
大兴机场	7
通武廊	7
北三县	11

他山之石栏目以国内经济带与国外经济带为研究目标,展示世界各地域协同发展的相关研究资料,以作为京津冀协同发展战略的重要参考对象。其中,国内经济带相关数据主要来源于长三角城市群、中原城市群、长江中游城市群、成渝城市群等,这些城市群在长期发展过程中积累了重要的发展经验,相关数据具有较高的参考价值。国外经济带数据采集以北美、欧洲各

大都市经济圈为主体,呈现了不同外部环境下的协同合作,相关发展经验可资借鉴。

3.2 存在的问题与展望

京津冀数据库虽取得初步成果,但在数据与功能层面还有许多待完善的地方。在数据方面,京津冀书籍加工采用低成本加工方式,最终生成的电子版为单层PDF,相比矢量PDF,单层PDF无法切割到目录与正文级别,导致对于书籍内容的检索只能到达书籍名称、ISBN号等字段,无法更进一步对目录进行检索。在功能方面,标引后的关键词只用来给文章标签化处理,没有更进一步加以利用挖掘深层次的功能,如“词云”“文本共现”等可视化的数据呈现功能,用以辅助研究者发现和梳理复杂数据中的关联关系。

纵观国内已建成或者正在建设的一些区域性数据库,虽然数据体量庞大、内容完备,但仍然有不少数据库在有意或者无意进行内容保护。基于多源异构数据融合的数据库相对较少,更多的是新闻、政策内容的堆砌。在数据库建设向数据化、智慧化转型的关键时期,结合京津冀数据库的建设实践,未来区域性专题数据库的建设可以考虑在以下3个方面加以强化与升级。

(1) 引入文本挖掘技术。着重加强非结构化数据向结构化数据的转化,深挖文本内容中潜在的信息价值,通过程序编码实现资源内容的层级化展现,如思维

导图的生成。

(2) 体现数据库共建共享理念。数据库建设后期加强专题数据库的宣传推广,完善数据库页面中搜索引擎优化的元素,让用户能更快找到、参与、共享数据库并进行相关反馈,使用户间接加入数据库的共建。

(3) 建立后期维护保障团队。数据库后期数据与功能的质量需要保持一致性,体现专题数据库的建设价值。

参考文献

- [1] 张心源,邱均平.国内图书情报领域知识融合研究的发展与分析[J].数字图书馆论坛,2016(3):17-23.
- [2] 新华社.中共中央 国务院关于建立更加有效的区域协调发展新机制的意见[EB/OL].[2020-11-29].http://www.gov.cn/zhengce/2018-11/29/content_5344537.htm.
- [3] 严丹,马吟雪.“一带一路”专题数据库的建设现状及开发策略研究[J].图书馆学研究,2017(12):40-47.
- [4] 吴利.学习贯彻党的十九届四中全会精神建设高质量长江经济带数据库[J].学习月刊,2019(12):32-35.
- [5] 王兰成,敖毅,曾琼.异构多信息源组织与集成技术的研究现状及其进展[J].现代图书情报技术,2006(3):68-71.
- [6] 唐晓波,刘广超.基于两层知识融合的金融知识服务模式研究[J].图书馆学研究,2018(16):79-85.
- [7] 化柏林,李广建.大数据环境下多源信息融合的理论与应用探讨[J].图书情报工作,2015,59(16):5-10.
- [8] 化柏林.多源信息融合方法研究[J].情报理论与实践,2013,36(11):16-19.
- [9] 余鹏,李艳,万晨.融合多源异构教育大数据的高校科研服务系统设计研究[J].图书情报知识,2019(1):32-43.
- [10] 于亚秀,李欣,鲁丹,等.基于多源异构数据库融合创新决策支持服务——以华东师范大学学科评估系统为例[J].图书馆学研究,2017(4):74-78.
- [11] 赵捷,董微.面向发现服务的图书馆元数据集成管理系统构建研究[J].数字图书馆论坛,2018(7):11-21.
- [12] 丁道劲,曾建勋.文献元数据集成管理研究[J].情报学报,2019,38(6):568-577.
- [13] 丁道劲,苏静,曾建勋.国家元数据库及其协同构建框架研究[J].情报理论与实践,2020,43(10):88-92.
- [14] 鲜国建,罗婷婷,赵瑞雪,等.从人工密集型到计算密集型:NSTL数据库建设模式转型之路[J].数字图书馆论坛,2020(7):52-59.
- [15] 张婕.京津冀区域一体化体育信息资源共享平台建设研究[J].河南图书馆学刊,2018(3):72-73.
- [16] 韩曾丽,赵蕾霞.京津冀协同发展背景下河北省千年古县特色数据库建设研究[J].农业图书情报学刊,2016,28(7):32-35.
- [17] 韩兆柱,单婷婷.基于整体性治理的京津冀府际关系协调模式研究[J].行政论坛,2014,21(4):32-37.

作者简介

陈晨,男,1988年生,硕士,副研究馆员,研究方向:信息管理,E-mail:jxxycc@qq.com。

The Construction of Regional Thematic Database under Multi-source Heterogeneity:
Taking the Beijing-Tianjin-Hebei Collaborative Development Database as an Example

CHEN Chen

(Library of Beijing Administration Institute, Beijing Research Center of Beijing Administration Insititute, Beijing 100044, China)

Abstract: By discussing the technical realization of multi-source heterogeneous regional thematic databases in an open access environment, providing new ideas and methods for libraries to develop digital resource open access services, and meeting the actual use needs of related users, and further improve the resource construction system. From the perspectives of the current situation of regional thematic database construction in the open access environment, the design and technical realization of multi-source heterogeneous thematic data, etc., a comprehensive introduction to the construction and practice process of multi-source heterogeneous thematic database, and the effect of the database in actual operation is analyzed and discussed, and several issues and suggestions for the construction of thematic database to satisfy open access services are put forward. The Beijing-Tianjin-Hebei thematic database built by the Library of Beijing Administration Institute has the characteristics of user-friendly experience, flexible functional design, and clear column division. It meets the requirements of database built and is closer to researchers in content.

Keywords: Multi-source Heterogeneity; Regional; Thematic Database; Beijing-Tianjin-Hebei

(收稿日期:2020-11-21)