基于本体的地学数据建库方法

杨先洪[1] 诸云强[2,3,4] 朱 腾[5] 杨秀权[1] 杨 杰[2,6] 孙 凯[2,6]

（1.贵州国土资源技术信息中心，贵州贵阳 550000；2.中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室，北京 100101；3.白洋淀流域生态保护与京津冀可持续发展协同创新中心，河北保定 071002；4.江苏省地理信息协同创新中心，江苏南京 210023；5.河南理工大学测绘与国土信息工程学院，河南焦作 454000；6.中国科学院大学资源与环境学院，北京 100049）

摘要：地学数据资源具有多源、分散、异构等特征，为了能够充分共享和利用地学数据资源，必须有效集成并建立地学数据库，但是传统地学数据建库方法存在着难以系统地认知建库对象，缺乏统一的语义支撑等问题。针对这些问题，本文提出并定义了地学数据本体模型、概念体系及其构建方法，阐述了基于地学数据本体的数据建库方法和流程。该方法在数据本体明确语义的支撑下，抽取出完善的数据实体关系模型，实现多源、异构数据的映射、转换与规范化处理，并利用数据本体准确表达数据库的语义信息。最后以贵州岩溶地下水数据库构建为例，进行了方法的应用验证。研究表明：地学数据本体通过对数据资源全生命周期共识概念、属性、实例及其关系的形式化描述，可以有效克服传统地学数据建库方法的缺陷，更好地支撑和促进地学数据资源的集成建库。

关键词：数据本体；数据库建设；岩溶地下水；语义

中图分类号：P208 **文献标识码**：A **DOI**：10.3772/j.issn.1674-1544.2017.05.004

Ontology-based Geoscience Database Building Method

YANG Xianhong[1], ZHU Yunqiang[2,3,4], ZHU Teng[5], YANG Xiuquan[1], YANG Jie[2,6], SUN Kai[2,6]

(1.Guizhou land and Resources Technology Information Center, Guiyang, Guizhou 550000; 2. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Nature Resources Research , Beijing 100101; 3. Collaborative Innovation Centre for Baiyangdian Basin Ecological Protection and Jingjinji Regional Sustainable Development, Hebei University, Baoding, Hebei 071002; 4. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing Normal University, Nanjing, Jiangsu 210023; 5. School of Surveying and mapping, Henan Polytechnic University, Jiaozuo, Henan 454000; 6. College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049)

作者简介：**杨先洪**（1977—），男，贵州国土资源信息中心高级工程师，研究方向：地理信息系统；**诸云强**（1977—），男，中国科学院地理科学与资源研究所研究员，研究方向：地球数据本体与共享、资源环境信息系统（通讯作者）；**朱腾**（1992—），男，河南理工大学硕士研究生，研究方向：地理信息系统；**杨秀权**（1981—），男，贵州国土资源信息中心，研究方向：地理信息系统；**杨杰**（1990—），男，中国科学院地理科学与资源研究所硕士研究生，研究方向：地学模型数据匹配方法；**孙凯**（1990—），男，中国科学院地理科学与资源研究所博士研究生，研究方向：地学本体及数据关联。

基金项目：科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”（2013FY110900）；贵州省公益性基础性地质工作项目“贵州省岩溶地下水系统功能可持续利用性研究”（黔国土资地环函〔2014〕23 号）；2016 年贵州省公益性基础性地质工作项目“贵州省国土资源可持续发展战略研究”（黔国土资源函〔2016〕269 号）。

收稿时间：2017 年 7 月 31 日。

Abstract: Geoscience data (Geo-data) resources have many characteristics, such as multi-sources, dispersed and heterogeneous. In order to fully share and make use of Geo-data, such datasets must be effectively integrated into databases. For traditional methods of Geo-data integration and database construction, they are difficult to systematically cognize target objects and lack unified semantic support. Firstly, this paper proposes the model, concept architecture and construction method of the Geo-data ontology. And then the data integration and database building method and process are introduced based on Geo-data ontology. Finally taking Guizhou karst groundwater database construction as an example, the proposed method is applied and verified. The result shows that Geo-data ontology which is the formal description of consensus concepts, instances, attributes and relations related to the whole life cycle of Geo-data resources, can effectively support and promote Geo-data integration and database building, Geo-data exchanging, sharing and application.
Keywords: data ontology, database construction, karst groundwater, semantics

地学数据是人们从事地球科学理论研究、调查考察、观测探测、试验实验、模拟分析等活动所产生的数据，具有来源分散、时空跨度广、类型格式多样、数据量大等特点[1−2]。为了能够有效利用地学数据资源，往往需要将这些多源、分散、异构的数据进行集成并建立规范化的数据库。已有的传统数据建库方法，通常是利用实体关系模型（E−R），即先设计统一的关系型数据库，然后利用ETL（抽取、转换、装载）、互操作或网络服务等方法，将数据集成导入数据库中[3−4]。但是，该方法存在两大问题：一是由于缺乏对实体对象本质的深刻认知，在设计阶段，实体属性及其关系并没有被系统梳理出来，导致实体关系模型设计的数据库结构往往只能反映实体对象的某个或某几个方面的特性；二是由于缺乏对实体概念、属性及其属性值语义信息的准确描述，导致不同来源语义异构的数据很难集成在同一个数据库中。本体作为领域共识的概念及其相互关系的形式化说明，已经被广泛用于解决领域地理对象识别、实体语义模糊性、语义搜索、数据分类、集成与关联等方面的研究中[5−16]。然而，上述基于本体的建库方法主要是利用领域本体解决专题数据分类、集成与关联应用等问题，还没有将本体应用到数据库设计阶段，也缺乏基础的地学数据本体模型的支撑，导致领域本体间难以融合和转换的问题。为此，本文开展了基于本体的地学数据建库方法的研究。

1　地学数据本体模型

地学数据本体是以地学数据为核心，地学数据全生命周期所涉及的、通过形式化方式表达的共识概念、实例及其关系。数据本体涉及数据采集生产、加工处理、存储管理、交换共享和应用分析等过程的所有概念和实例，包含数据特征概念，以及支撑或作用于数据的责任者、仪器设备和模型工具等。面向地学数据全生命周期的数据本体抽象模型如图 1 所示。

地学数据特征概念是地学数据本体的核心，通过无歧义的语义表达，支撑地学数据的分类集成、交换共享和分析应用等。地学数据特征包括反映地学数据本质的时间、空间、主题特征概念，以及反映地学数据形态的数学基准、类型格式、语言字符、属性语义等特征概念。责任者、仪器设备和模型工具也是地学数据本体的重要组成部分，主要对数据采集生产、加工处理等过程进行准确描述，有利于数据源追溯和质量控制等。地学数据本体概念体系如图 2 所示。

依据图 2 的数据本体概念体系，将地学数据本体分为本质本体、形态本体和来源本体。其中，本质本体由反映地学数据本质特征的概念组成，包括时间本体、空间本体、要素内容本体；形态本体由反映地学数据形态特征的概念组成；来源本体由反映数据采集、处理的仪器设备、模型工具和责任者等概念组成。从本体构建的角度，可以进一步将数据本体分为基础本体、领域

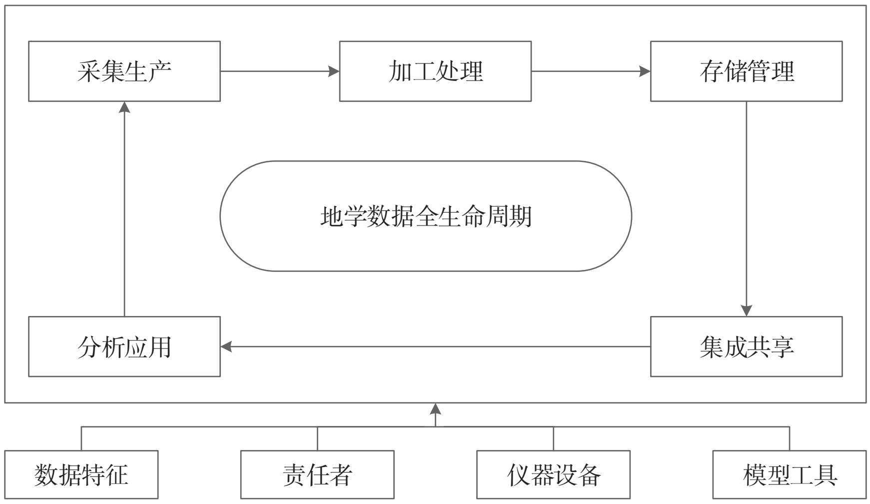

图1　地学数据本体抽象模型

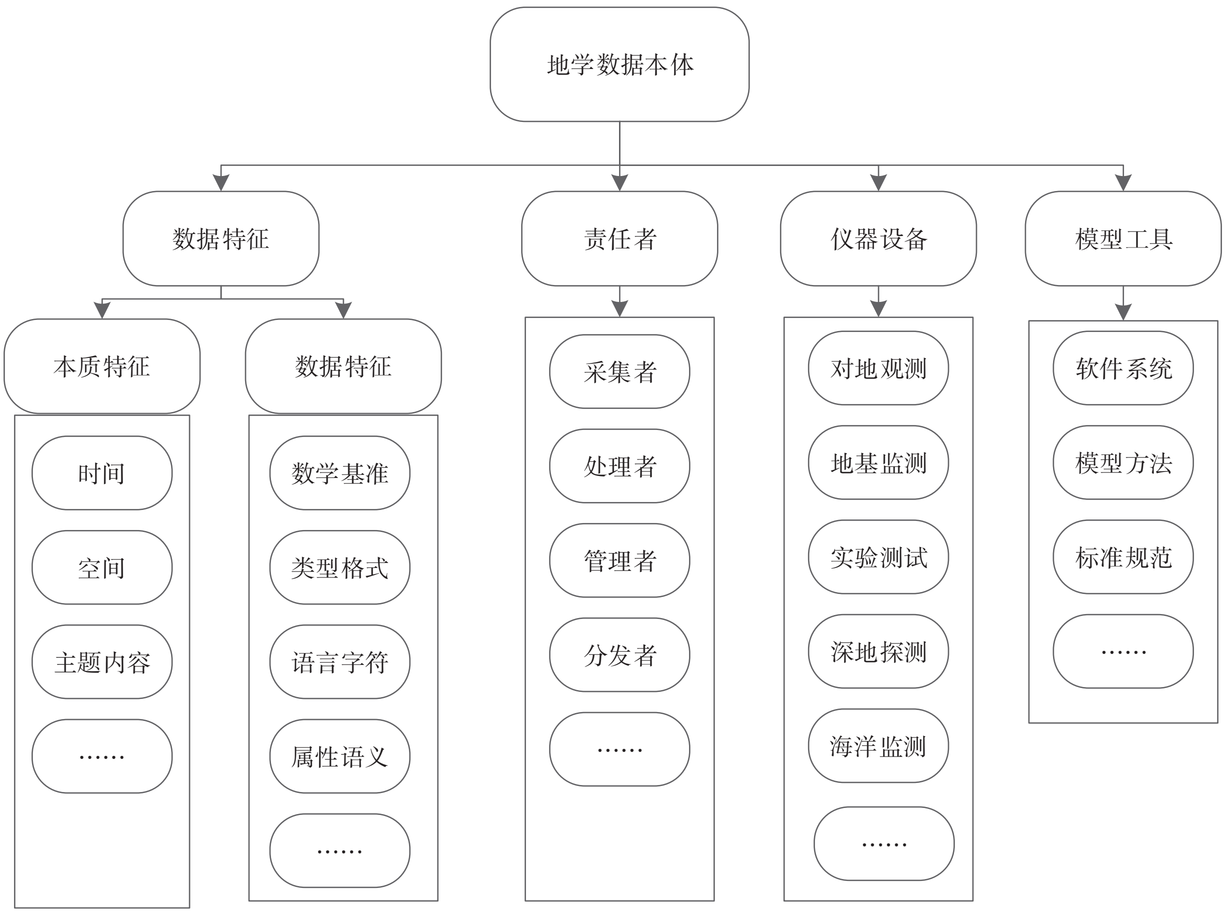

图2　地学数据本体概念体系

本体和应用本体。基础本体是指与领域无关的共性本体，包括：时间、空间、数学基准、语言字符等本体；领域本体是指领域内共识的本体，包括：主题内容、类型格式、属性语义（属性值单位、分类代码等）、仪器设备、模型工具等本体；应用本体是指与具体相关的本体，包括：责任者或其他与某个应用有关的本体。数据基础本体可以充分重用现有的DAML、KSL、SWEET、 GeoNames、EPSG等时间、空间和数学基准本体；数据领域本体和应用本体一般需要领域和应用研究者根据领域知识和应用场景由自己进行构建。

2 地下水领域本体

地下水是水资源的重要组成部分，指赋存并运移于地面以下岩土空隙中的水。地下水是在一定水文地质基础（条件）下孕育产生的，并通过各种开发利用手段，实现水资源供给和环境支撑等功能。在开发利用过程中，将产生一系列环境地质问题，并通过保护治理措施加以解决。因此，地下水领域本体抽象模型如图 3 所示。

依据地下水本体抽象模型，地下水本体涉及的顶层概念主要包括[17−21]：水文地质基础、地下水资源、地下水环境、人类活动、环境地质问题。水文地质基础包括：地层、地质构造、包气带水、潜水、承压水、孔隙水、裂隙水、岩溶水、暗河、泉、地下水系统、地下水类型、含水岩组、水文地质参数、含水层等二级概念。地下水资源包括：地下水水位、水量、地下水补给、地下水消耗等二级概念。地下水补给又包括降水入渗、地表水补给、侧向补给、越流补给等，而地下水消耗包括地下水开采、蒸发、地下水溢出、侧向排泄、越流排泄等。地下水环境包括：地下水化学类型、水温、水质、地下水污染等二级概念。人类活动包括：责任者和事件两个二级概念，责任者包括钻孔实施者、地下水采样者、测试分析者、地下水资源评价者等实例，而事件包括水文地质调查、地下水勘察、钻孔施工、地下水监测、水质测试、水资源评价等实例。环境地质问题包括：地面塌陷、降落漏斗、地裂缝、地下水污染、盐渍化等二级概念。地下水核心概念、属性、关系及其实例示例如表 1 所示。

3 地学数据库建库步骤

数据本体支持下的地学数据建库方法主要包含四大步骤（图 4）：（1）依据领域本体概念、概念属性、实例及其关系，抽象出数据库实体、属性及其实体关系，形成实体关系模型；（2）基于实体关系模型设计数据库结构；（3）依据基础和领域本体，对多源数据进行语义消歧、数据格式等规范化转换并入库；（4）基于数据本体明确的语义描述对最终建成的数据库进行形式化表达，以便后继数据的持续集成。

4 案例研究：贵州岩溶地下水数据库的构建

贵州省地处我国西南连片岩溶分布区的核心部位，岩溶地貌分布面积占全省 61.9%，岩溶地下水极为丰富，岩溶地下水资源量占全省总水资源量的 46%。为了合理开发利用岩溶地下水，自上世纪 80 年代中期，贵州省开展了水文地质调查、地下水监测、地下水资源评价、勘察找水等工作[22]，形成了全省 1:20 万的水文地质普查、部分 1:5 万的水文地质调查、6 个中心城市地下水长期监测数据成果以及各类地下

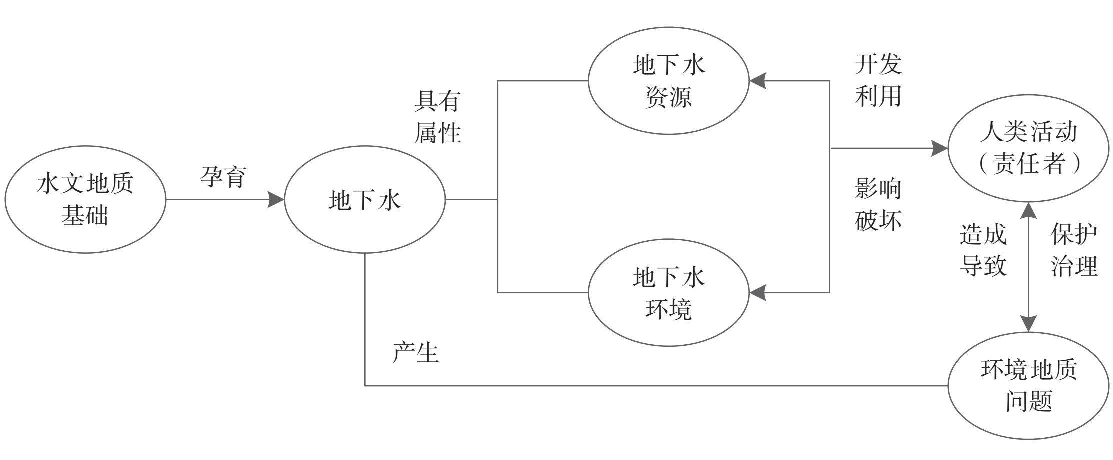

图 3 地下水本体抽象模型

表 1　地下水本体概念、属性、关系及实例示例（以贵州岩溶地下水为例）

地下水概念		核心属性	关系	实例
水文地质基础	地下水系统	流域系统名称、代码、面积、级别、地形、地貌、岩性、地下水类型、含水岩组、补给条件、径流条件、排泄条件等	地下水系统{具有}地下水类型、地下水系统{具有}含水岩组等	长江流域（一级）、乌江（二级）、思南以上（三级）、南明河（四级）等
	地下水类型	岩石类型、化学成份、含水介质组合、水动力特征等	地下水类型{表征}地下水系统等	溶洞–管道水、溶隙–溶洞水、溶孔–溶隙水等
	地下水水位含水岩组	岩性组合、含水介质组合方式等	含水岩组{表征}地下水系统等	纯石灰岩岩组、灰岩夹碎屑岩岩组、白云岩夹碎屑岩岩组等
	……	……	……	……
地下水资源		监测点、水位、监测时间	监测点{属于}地下水系统等	
	地下水补给	地下水系统、大气入渗补给、地表水补给、侧向补给、越流补给等	地下水补给{注入到}地下水系统等	
	地下水消耗	地下水系统、蒸发、溢出、开采、侧向排泄、越流排泄等	地下水消耗{排泄于}地下水系统等	
	……	……	……	……
地下水环境	地下水化学类型	主要阳离子及浓度、主要阴离子及浓度、同位素、微量元素等	地下水系统{具有}地下水化学类型等	重碳酸盐型水、硫酸盐型水、氯化物型水等
	地下水水温	监测点、水温、监测时间	水温{具有}单位	
	地下水水质	取样点、监测项目、监测值、取样时间、测试时间、水质类别等	监测值{具有}单位、水质类别{具有}类别代码等	
	……	……	……	……
人类活动	责任者	名称、唯一编码、通讯地址、联系电话	责任者{实施}事件	钻孔施工者、地下水采样者、监测者、开采者等
	事件	名称、事件描述、事件发生时间、地点、责任者	事件{由…实施}责任者	地下水勘查、监测、开采等
	……	……	……	……
环境地质问题	地面塌陷	塌陷区位置、塌陷面积、深度、监测时间	地面塌陷{位于}地下水系统	
	降落漏斗	降落漏斗位置、面积、监测时间	降落漏斗{位于}地下水系统	
	地裂缝	地裂缝位置、宽度、深度、长度、监测时间	降落漏斗{位于}地下水系统	
	……	……	……	……

水研究报告等。这些数据资料和成果主要分散在贵州省地质环境监测院、贵州省地质资料馆、贵州省地矿局111地质大队、114地质大队等。

为了整合集成上述地下水数据资料，构建形成统一的贵州省岩溶地下水数据库，主要采用了以下3种方式：一是收集并规范化处理分散在各单位的电子数据文件；二是通过互操作技术，实现地下水监测数据库的互连互通；三是通过文本挖掘技术，提取各类研究报告中的地下水数据。在具体实现时（图5），首先建立地学时间、空间、形态本体以及地下水领域本体[13-16]；然后基于本体完成地下水数据库结构的设计，保证数据库结构的系统性和完整性；在此基础上，在本体的支持下，消除数据时空基准、类型结构和语义等方面的异构性，实现数据的转换处理、互操作和挖掘抽取；最后对地下水数据库中的各类数据进行形式化的语义描述，建立数据字典。截止到2017年5月，贵州省地下水数据库整合集成了自上世纪80年代起的贵州省水文地质调查、地下水勘察、监测和评价分析等方面的研究成果以及数据资料99GB，10万多条数据记录，并建立了贵州地下水数据资源管理系统（图6）。

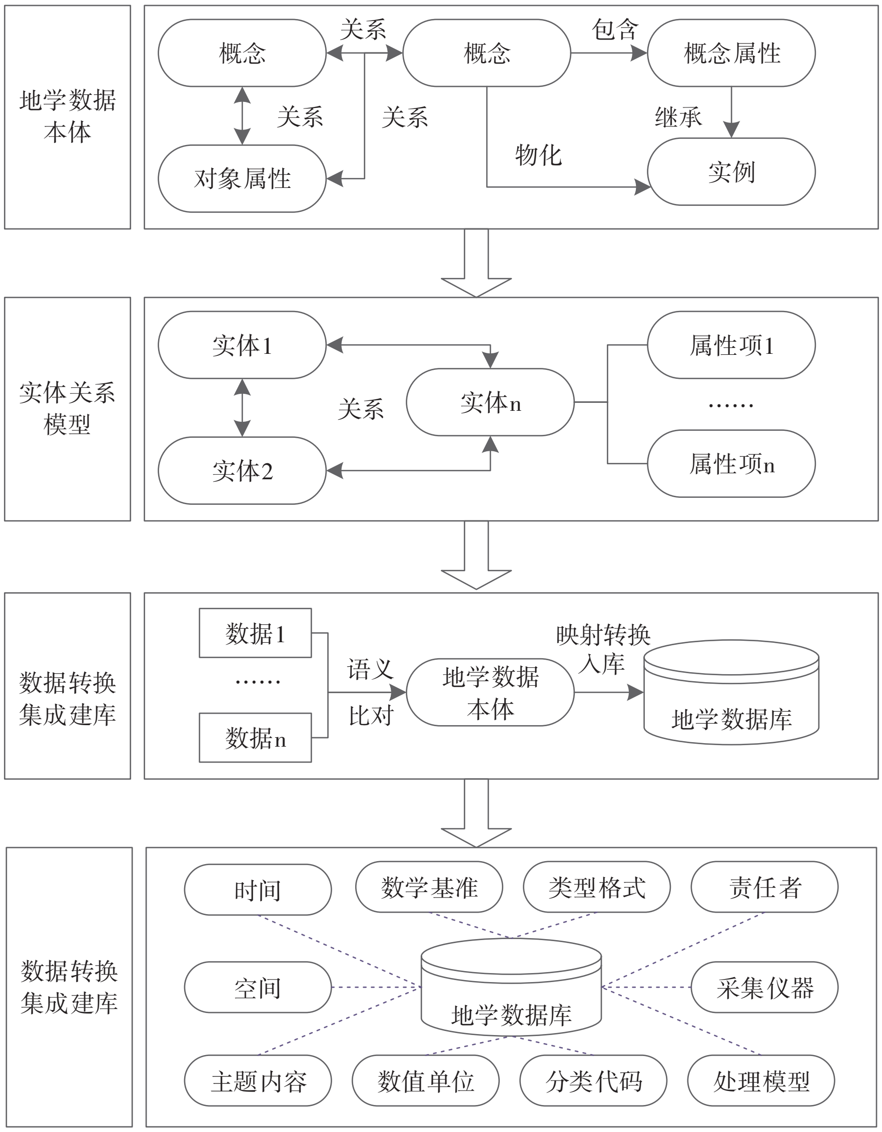

图 4　基于本体的地学数据建库方法流程

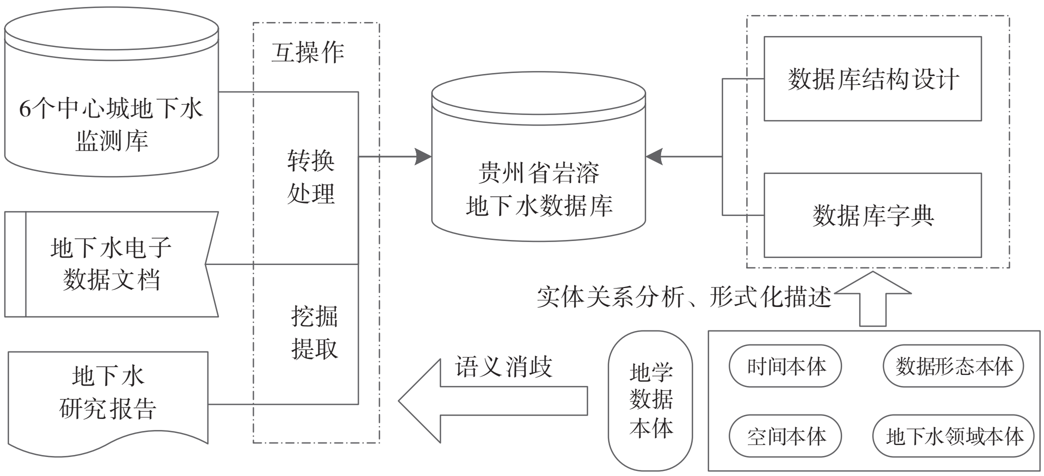

图 5　基于本体的贵州岩溶地下水数据库构建

图6　贵州地下水数据资源管理系统

5　结语

本文从数据集成共享的角度，提出地学数据本体模型，并以贵州岩溶地下水数据库构建为案例，提出基于地学数据本体模型的领域本体构建方法及其在数据集成建库中的应用。

一是提出并定义了数据本体，认为数据本体是对数据资源全生命周期涉及的共识概念、实例及其关系的形式化描述，可以为数据资源的分类集成、数据库设计与建设，甚至是数据的采集处理、存储管理和交换共享等应用提供有效的语义支撑。地学数据本体是数据本体在地球科学领域的应用，由时间、空间等基础本体和主题内容等领域本体构成。基础本体可以引用集成现有的知识本体，而后者则需要依据学科领域的特点，进行手工或半自动的构建。

二是基于本体提出数据建库的方法：首先基于数据本体构建实体关系模型，然后利用实体关系模型设计完善的数据库结构，接着在本体的支持下完成数据的映射、转换与规范化处理并入库，最后利用数据本体记录和表达数据库结构和语义信息。

三是以贵州岩溶地下水项目为对象验证了基于本体的地学数据库构建方法。该方法在数据本体明确语义的支撑下，首先抽取数据实体关系模型，指导数据库结构的设计；然后利用数据本体实现多源、异构数据的映射、转换与规范化处理；最后通过数据本体标识数据库的语义信息。从而克服了传统数据建库方法难以系统地认知建库对象，缺乏统一的语义支撑等问题，并通过将地学数据本体与地下水领域本体与数据库设计相结合，更好地支撑和促进建立地学数据资源库。

尽管笔者在已往的研究中已经构建了时间、空间和形态本体，下一步还应继续加强领域本体的研究与构建，并研发地学数据本体标注和转换工具，提高数据语义标注和转换的效率与正确性。

参考文献

[1]　孙九林, 施慧中. 中国地球系统科学数据共享服务网的构建[J]. 中国基础科学, 2003(1): 76−81.DOI: 10.3969/j.issn.1009−2412.2003.01.015.

[2]　黄鼎成. 科学数据共享的理论基础与共享机制[J]. 中国基础科学, 2003(2): 22−27.DOI: 10.3969/j.issn.1009−2412.2003.02.005.

[3]　龚健雅, 贾文珏, 陈玉敏, 等. 从平台GIS到跨平台互操作GIS的发展[J]. 武汉大学学报信息科学版, 2004, 29(11): 985−989.DOI: 10.3321/j.issn: 1671−8860.2004.11.011.

[4]　张鸣之, 诸云强, 罗德利, 等. 地质环境数据集成服务及其系统实现[J]. 中国地质灾害与防治学报, 2013, 24(3): 84−89.

[5]　何建邦, 李新通, 毕建涛, 等. 资源环境信息分类编码及其与地理本体关联的思考[J]. 地理信息世界, 2003(5): 6−11.DOI: 10.3969/j.issn.1672−1586.2003. 05.003.

[6]　李宏伟, 李勤超. 基于本体的土地利用数据分类分级表达[J]. 地域研究与开发, 2009, 28(1): 82−85.DOI: 10.3969/j.issn.1003−2363.2009.01.017.

[7]　刘耀林, 李红梅, 杨淳惠. 基于本体的土地利用数据

（下转第51页）

发展报告（2011−2012）[M].北京：科学技术文献出版社，2013: 2−9.
[4] 曹一化，刘旭，等.自然科技资源共性描述规范[M].北京：中国科学技术出版社，2006: 85−86.
[5] 国家岩矿化石标本资源共享平台[DB/OL].[2017−07−03].http: //nimrf.net.cn.
[6] MUSEUM COLLECTION RECORDS.Smithsonian National Museum of Natural History[DB/OL].[2017−07−03].http: //collections.nmnh.si.edu.
[7] DIVISION OF PALEONTOLOGY.American Museum Natural History[DB/OL].[2017−07−03].http: //research.amnh.org/paleontology/search.php.
[8] THE DATABASE RRUFF™. The University of Arizona Mineral Museum[DB/OL].[2017−07−03].http://rruff.info.
[9] DATA PORTAL.Natural History Museum[DB/OL].[2017−07−03].http: //data.nhm.ac.uk/.
[10] EARTH COLLECTIONS: SEARCHABLE DATA-BASES.Oxford University Museum of Natural History [DB/OL].[2017−07−03].http: //www.oum.ox.ac.uk/collect/earthcoll2.htm.
[11] RBINS−COLLECTION DATABASE.Royal Belgian Institute of Natural Sciences[DB/OL].[2017−07−03].http: //darwin.naturalsciences.be/search/geoSearch.
[12] PALEOBIOLOGY DATABASE. Fossilworks[DB/OL].[2017−07−03].http: //fossilworks.org/?a=collectionSearchForm&type=view.
[13] GEOCASE DATA PORTAL.Geocase[DB/OL].[2017−07−03].http://www.geocase.eu/portal? terms Accepted =true.
[14] 何明跃，吴淦国.岩矿化石标本资源收集整理保存技术规程[M].北京：地质出版社，2015: 4−75.
[15] 左群超，杨东来，宋越，等. 中国矿产资源潜力评价成果数据质量控制及方法技术[J].中国地质，2013, 40(4): 1314−1328.
[16] 许嘉玲，陈柏堃，王东，等. 地面观测月数据质量控制的人机交互问题与处理[J].气象研究与应用，2011, 32(Z2): 147−148.

（上接第 36 页）
综合研究[J]. 武汉大学学报信息科学版，2010, 35(8): 883−886.
[8] 周熙然，邵振峰，周寿章. 基于地理本体的空间传感网数据处理[J]. 云南大学学报自然科学版，2011(Z2): 196−201.
[9] 陈玉鑫，李景文，崔爽，等. 基于地理本体的空间数据组织方法与应用[J]. 测绘与空间地理信息，2011, 34(6): 47−49.DOI:10.3969/j.issn.1672−5867.2011. 06.015.
[10] 宋佳，王卷乐，诸云强，等. 基于地理空间本体的语义检索相关度研究[J]. 计算机工程与应用，2011, 47(5): 114−117.DOI: 10.3778/j.issn.1002−8331.2011.05.035.
[11] 张朴，王锡洁. 基于地理本体建模的河流阶地识别方法研究：以汉中盆地为例[J]. 安徽农业科学，2014 (10): 3080−3082.DOI: 10.3969/j.issn.0517−6611. 2014.10.093.
[12] 詹长根，严盼，涂李蕾，等. 基于地理本体的土地空间面对象模型[J]. 测绘与空间地理信息，2015(9): 12−16.DOI: 10.3969/j.issn.1672−5867.2015.09.005.
[13] 侯志伟，诸云强，高星，等. 时间本体及其在地学数据检索中的应用[J]. 地球信息科学学报，2015, 17(4): 379−390.DOI: 10.3724/SP.J.1047.2015.00379.
[14] 王东旭，诸云强，潘鹏，等. 地理数据空间本体构建及其在数据检索中的应用[J]. 地球信息科学学报，2016, 18(4): 443−452.DOI:10.3724/SP.J.1047.2016.00443.
[15] 罗侃，诸云强，程文芳，等. 极地科学数据关联方法及应用研究[J]. 极地研究，2016, 28(3): 361−369.DOI: 10.13679/j.jdyj.2016.3.361.
[16] 孙凯，诸云强，潘鹏，等. 形态本体及其在地理空间数据发现中的应用研究[J]. 地球信息科学学报，2016, 18(8): 1011−1021.DOI:10.3724/SP.J.1047.2016.01011.
[17] 张永波，梁国玲，张礼中，等. 中国地下水资源空间数据库标准化研究[J]. 地球学报，2003, 24(4): 371−374. DOI: 10.3321/j.issn:1006−3021.2003.04.015.
[18] 彭淑惠. 云南岩溶区地下水及地质环境数据库建设[J]. 云南地质，2005, 24(2): 232−239.DOI: 10.3969/j.issn. 1004−1885.2005.02.013.
[19] 张礼中，周小元，张永波，等. 西北地下水资源数据库及其网上发布[J]. 地球学报，2001, 22(4): 307−310. DOI: 10.3321/j.issn:1006−3021.2001.04.005.
[20] 梁国玲，张永波，张礼中，等. 基于GIS的中国地下水资源空间数据库建设[J]. 地球学报，2007, 28(6): 572−578.DOI: 10.3321/j.issn:1006−3021.2007.06.010.
[21] 梁国玲，张永波，张礼中，等. 区域地下水资源数据库标准建设问题探讨[J]. 工程勘察，2010, 38(6): 31−34.
[22] 杨胜元. 贵州环境地质[M]. 贵州：贵州科技出版社，2008.