

引用格式:李威蓉,诸云强,宋佳,等.地理空间数据来源本体及其在数据关联中的应用[J].地球信息科学学报,2017,19(10):1261-1269. [ Li W R, Zhu Y Q, Song J, et al. 2017. Geospatial data provenance-ontology and its application in data linking. Journal of Geo-information Science, 19(10):1261-1269. ] DOI:10.3724/SP.J.1047.2017.01261

# 地理空间数据来源本体及其在数据关联中的应用

李威蓉<sup>1</sup>, 诸云强<sup>2,4,5</sup>, 宋佳<sup>2,4,5\*</sup>, 孙凯<sup>2,3</sup>, 杨杰<sup>2,3</sup>

1. 山东理工大学建筑工程学院, 淄博 255000; 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. 中国科学院大学, 北京 100049; 4. 白洋淀流域生态保护与京津冀可持续发展协同创新中心, 保定 071002; 5. 江苏省地理信息资源开发与利用协同创新中心, 南京 210023

## Geospatial Data Provenance-Ontology and Its Application in Data Linking

LI Weirong<sup>1</sup>, ZHU Yunqiang<sup>2,4,5</sup>, SONG Jia<sup>2,4,5\*</sup>, SUN Kai<sup>2,3</sup> and YANG Jie<sup>2,3</sup>

1. School of Architecture Engineering, Shandong University of Technology, Zibo 255000, China; 2. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China; 3. University of Chinese Academy of Sciences, Beijing 100049, China; 4. Baiyangdian Lake Ecological Protection and Sustainable Development of Jing-jin-ji Collaborative Innovation Center, Baoding 071002, China; 5. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application Nanjing 210023, China

**Abstract:** Data provenance is an important reference factor of data reliability evaluation and important research content of geospatial data ontology. Taking consideration of provenance, an important research object of geospatial data, we constructed a geospatial data provenance conceptual model based on systemic analysis of the meaning of geospatial data provenance. Based on it, we put forward geospatial data Provenance-Ontology concepts system and the formalization method for constructing geospatial data Provenance-Ontology. Finally, we take the data materials in “special work of the science and technology basic work” as an example. Based on Provenance-Ontology library, using RDF to link geospatial data and D3.js to achieve the data provenance visualization. The result shows that data linking based on Provenance-Ontology can effectively solve the problem of the nonstandardization in the description of data provenance information. It can support geospatial data semantic retrieval, intelligent recommendation and other applications. It also provides new ideas for geodata sharing and data linking.

**Key words:** geospatial data; provenance; ontology; data linking

**\*Corresponding author:** SONG Jia, E-mail: songj@lreis.ac.cn

**摘要** 数据来源是数据可靠性评价的重要参考因素,是地理空间数据本体的重要研究内容。本文针对来源这一重要的地理空间数据研究对象,系统地分析了地理空间数据来源的涵义,建立了地理空间数据来源本体模型,在此基础上,提出了地理空间数据来源本体的概念体系和来源本体概念间关系及其属性的形式化表达方法,并构建出地理空间数据来源本体。最后,以

收稿日期 2017-06-12;修回日期:2017-08-11.

基金项目 科技基础性工作专项重点项目(2013FY110900);国家自然科学基金重点项目(41631177);贵州省公益性基础性地质工作项目(黔国土资地环函[2014]23号);贵州省公益性基础性地质工作项目(黔国土资资源函[2016]269号);国家自然科学基金项目(41371381)。

作者简介 李威蓉(1991-),男,江西萍乡人,硕士生,研究方向为地学数据来源及数据关联。E-mail: liwr@lreis.ac.cn

\*通讯作者 宋佳(1980-),男,山西太原人,博士,助理研究员,研究方向为地球信息科学。E-mail: songj@lreis.ac.cn

“科技基础性工作专项”项目数据资料为例,基于来源本体库,利用RDF从来源角度实现数据的语义关联,通过web前端框架D3.js技术实现数据与其来源信息的可视化。结果表明,基于来源本体的数据关联可以有效解决数据来源信息描述不规范的问题以及能够支持地学数据语义检索、智能推荐等应用,为促进地学数据共享和数据关联应用提供了一种新方法和新思路。

**关键词** 地理空间数据;来源;本体;数据关联

## 1 引言

近年来,对地观测、地球深部、表层系统、空间环境探测系统、物联网、web 2.0等技术的发展,极大地提升了地球科学数据获取的能力,使地理空间数据呈爆炸式增长。海量的数据资源由于其采集和加工处理过程无法得知,用户往往都会对这些数据的可靠性提出质疑,使用户不得不关注数据的来源。现有对数据来源信息的描述主要以元数据或者数据文档的形式存在,且描述不规范,难以进行数据的可靠性评估,这严重阻碍了数据的使用。本体(Ontology)是共享概念模型明确的形式化规范说明<sup>[1]</sup>,能够对领域内的概念以及概念之间的关系进行明确的表达,且具有强大的语义推理能力,能够建立数据之间的语义关系,实现数据之间的智能关联。将来源这一重要的地理空间数据研究对象与本体理论相结合进行研究,构建出来源本体,对数据来源信息进行规范化处理,具有非常重要的意义。

目前,国内外针对来源本体展开了大量的研究,主要包括以下3个方面:

(1)来源信息认知和描述。ISO19115元数据标准<sup>[2]</sup>通过数据志对数据源、数据处理方法和数据质量控制等信息进行描述。Di等<sup>[3]</sup>总结了现有的数据来源理论方法和技术,提出了针对来源研究的5大方向,即来源信息的表达、获取、存储、查询以及可视化。戴超凡等<sup>[4]</sup>系统阐述了数据来源的概念及其内容。

(2)来源模型研究。陈颖等<sup>[5]</sup>将流起源信息模型、时间-值中心起源模型、四维起源模型等不同领域中的来源模型进行了系统地对比分析,利用生物学当中的DNA双螺旋结构,提出了基于DNA双螺旋结构的来源模型,给来源模型的建立提供了一种新思路。李文燕等<sup>[6]</sup>对来源的概念进行了系统的阐述,详细对比分析出了OPM模型<sup>[7]</sup>、Provenior模型<sup>[8]</sup>、CRM<sub>dig</sub>模型<sup>[9]</sup>、PROV模型<sup>[10]</sup>的特点,并应用示例说明及汇总出了PROV模型应用情景,即以代理为中心、以对象为中心、以过程为中心。戴超凡等<sup>[4]</sup>系统地描述了2个典型的形式化模型,OPM模型<sup>[7]</sup>和

Provenior模型<sup>[8]</sup>,总结出了Chimera<sup>[11]</sup>、CMCS<sup>[12]</sup>、MyGrid<sup>[13]</sup>、PASOA<sup>[14]</sup>、Kepler<sup>[15]</sup>等多个追溯系统的技术特点,并提出目前该领域需要解决的3个问题,统一业界标准、保证来源信息安全以及解决数据流在多系统之间传递的来源追踪问题。贾君枝等<sup>[16]</sup>详细分析了W7模型<sup>[17]</sup>和PROV模型<sup>[10]</sup>中各个结构、语义特征以及它们之间的关系,并设计了PROV数据来源应用情景,结合应用情景对模型的语义特征进行了说明。

(3)来源本体应用。Harting等<sup>[18]</sup>对关联数据的来源问题进行了系统地阐述和讨论,且提出了来源模型,使用该模型可以获取网页上描述数据的创建信息,从而评估数据的质量和可靠性。Jane Hunter等<sup>[19]</sup>提出了一种有关来源的资源管理器,结合用户输入的内容、语义推理以及访问策略,能够自动产生具有来源关系的个性化视图。乐鹏等<sup>[20]</sup>针对地理空间数据来源3方面的问题进行总结,即采集、管理和服务,提出面向SOA的空间数据来源系统框架。

目前,国内外针对来源本体的研究虽然已有丰富的成果,但还存在不足之处:现有的来源本体主要从数据质量控制和追溯的角度,关注数据的来源、处理方法和规范等,较少从数据发现共享与关联利用的角度,围绕数据资源全生命周期,系统研究和构建地理空间数据来源本体。即使是现有的地理空间元数据中已经有大量的数据来源信息,但缺乏对来源概念间多维关系和语义化表达,如不同来源、类型、尺度地理空间数据引用参考、时间序列化、更新修订、融合同化、尺度变换、矢栅转换等关系的精准定义。在此背景下,本文在已有来源本体研究的基础上,从数据发现共享和关联利用的全新视角,以地理空间数据全生命周期主客体为核心,结合地理空间数据特征,提出地理空间数据来源本体模型,构建较为完整的来源本体概念体系,对来源本体概念间关系及其属性进行精确的形式化表达,并以“科技基础性工作专项”项目数据资料为例,开展基于地理空间数据来源本体的关联应用实践。

## 2 地理空间数据来源本体概念体系

### 2.1 地理空间数据来源本体模型

来源一词最初是源于法语单词“provenior”,意为“to come from”。目前,国内外学者提出了多个模型对来源进行定义,应用较为广泛的有W7模型<sup>[17]</sup>、PROV模型<sup>[10]</sup>、OPM模型<sup>[7]</sup>、Provenior模型<sup>[8]</sup>等,它们从不同的角度对来源进行定义,都有各自的优缺点(表1)。

为了更好地应用到地理空间数据中,本文综合上述模型对数据来源涉及的人、机构以及活动、过程等要素的考虑,从数据发现共享和关联利用的全新视角,以地理空间数据全生命周期为核心,结合地理空间数据显著的时空特征,构建从数据源,到数据采集、加工、管理、分发等数据活动,以及数据活动过程各阶段采用的工具(方法)和涉及的各类主体(责任者)为框架的地理空间数据来源本体模型(图1)。在图1的地理空间数据来源本体模型中,依据应用目的,数据责任者使用工具直接采集数据,或基于数据源通过加工处理、融合同化、尺度变换、矢栅转换等数据活动生产出新的数据,数据活动过程每一项操作的时间、空间和责任者都被记录,以便进行数据质量的控制、追溯与关联发现。因此,本文将地理空间数据来源定义为在一个数据的完整生命周期(数据从创建到销毁的过程)内,记录在数据活动(数据采集、数据加工、数据分发、数据管理等)过程中所涉及到的责任人、责任团体、数据源、时间、空间等相关信息。它是能够判定数据可靠性的重要参考以及实现责任制的重要基础。

数据责任者起着至关重要的作用,它是所有动作的执行者,是在整个数据活动中,所有与数据有关的机构和个人,是在数据活动中承担责任或功能的一种描述,表明了它与数据活动之间的密切联系,指明了它是如何参与数据活动,并承担其

相应的责任。

数据活动指的是对地学数据产生影响的一系列动作的集合。从来源本体模型(图1)中可看出,数据活动是整个来源本体模型的中心,是生产数据的过程,数据活动过程中的每一步、每一个细节的好坏或者更改都会影响到最终数据产品的精度和质量。

时间是来源本体模型中的一个重要要素,指的是数据活动的发生时间,即数据活动是何时开始和何时结束以及数据是何时生产和何时利用等有关的时间信息。其中,所涉及到的时间概念不包括数据本身的时间(数据活动的时间和数据时间一致的情况除外)。例如,“2015年中国1:10万土地覆被数据”以及“1978–2012年中国长时间序列深雪数据集”,这2个数据集名称中所涉及的时间“2015年”和“1978–2012年”不属于来源概念模型所涉及的时间,因为数据内容本身的时间与“来源”无关,即数据内容本身的时间不会影响来源这一过程。

空间与数据活动密切相关,是对数据活动的描述,指的是数据活动发生的具体地点。来源本体模型中,空间与地学数据中的空间概念有所区别,不包括“投影系统、高程系统、坐标系统”等对数据特征的描述。

工具是人利用外界物体作为自身功能的一种延伸,在来源本体模型中,指的是数据责任者在进行数据活动生产数据的过程中所使用的器具。工具是来源的一个重要要素,工具的选择是否合理与最终数据产品的生成有着密切的联系,合理地选择工具能极大地提升工作效率,也直接影响着数据产品的质量和精度。

数据源指的是在数据加工过程中所使用的原始数据资料。目前,大部分数据产品都经过一系列的加工处理操作,这些数据都存在数据源,数据源本身的质量在一定程度上也会对最终的数据产品

表1 来源模型间的优缺点对比

Tab. 1 The advantages and disadvantages between provenance models

模型名称	模型描述	优点	缺点
W7	由7个相互关联的要素组成,即 what、where、why、how、which、when、who,详细地描述了它们之间的相互关系	来源要素完整	通用模型,难以应用于具体领域
PROV	W3C标准,计算机可以读取和处理的来源框架,支持 owl、XML 等多种格式,定义了如何获取、利用以及验证来源信息	完整定义了人、机构以及活动之间的关系	通用模型,难以应用于具体领域
OPM	由 Artifact、Process、Agent 3个要素组成,定义某个对象在不同状态时的因果关系	完整定义了某个对象在不同状态时的因果关系	缺少时间、空间等重要来源要素
Provenior	一种描述工作流的来源模型,由 data、agent、process 3个要素组成	完整的工作流过程	缺少数据间关系的描述



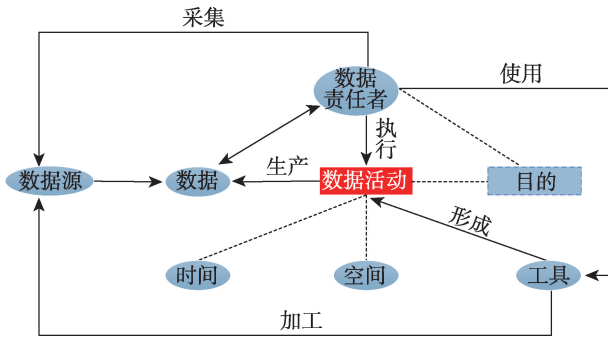


图1 地理空间数据来源本体模型

Fig. 1 The model of provenance-ontology of geospatial data

造成一定的影响,所以数据源这一概念在来源本体模型中也极其重要。

## 2.2 地理空间数据来源本体概念体系

地理空间数据来源本体是对地理空间领域内共享的来源概念明确的形式化规范说明。来源概念体系是构建地理空间数据来源本体的重要基础。根据图1,地理空间数据来源本体以数据本身

为核心,描述数据全生命周期涉及的数据源、数据活动及其各阶段利用的工具和涉及的责任者等信息。因此,地理空间数据来源本体一级概念由:数据源、数据、数据活动、工具、数据责任者组成,如图2所示。数据责任者包括组织机构和个人2种类型,它们在数据生命周期中充当数据的采集者、处理者、管理者、分发者、使用者等;数据活动是对数据从最初的采集、加工到最终的分发、管理等全生命周期过程的完整概括。因此,数据活动包括数据采集、数据加工、数据分发、数据管理4大概念。其中,数据采集和数据加工最为重要,其直接影响到最终数据产品的质量和精度,这也是数据生产者和数据使用者最为关注的内容。对于原始数据而言,数据采集过程尤为重要,采集过程中所涉及到的采集环境、采集方法、采集工具等都与数据质量密切相关,先进的采集工具、正确的采集方法以及良好的采集环境能很好地提升数据的质量;对于二次加工数据而言,在对数据加工的过程中,涉及到的一

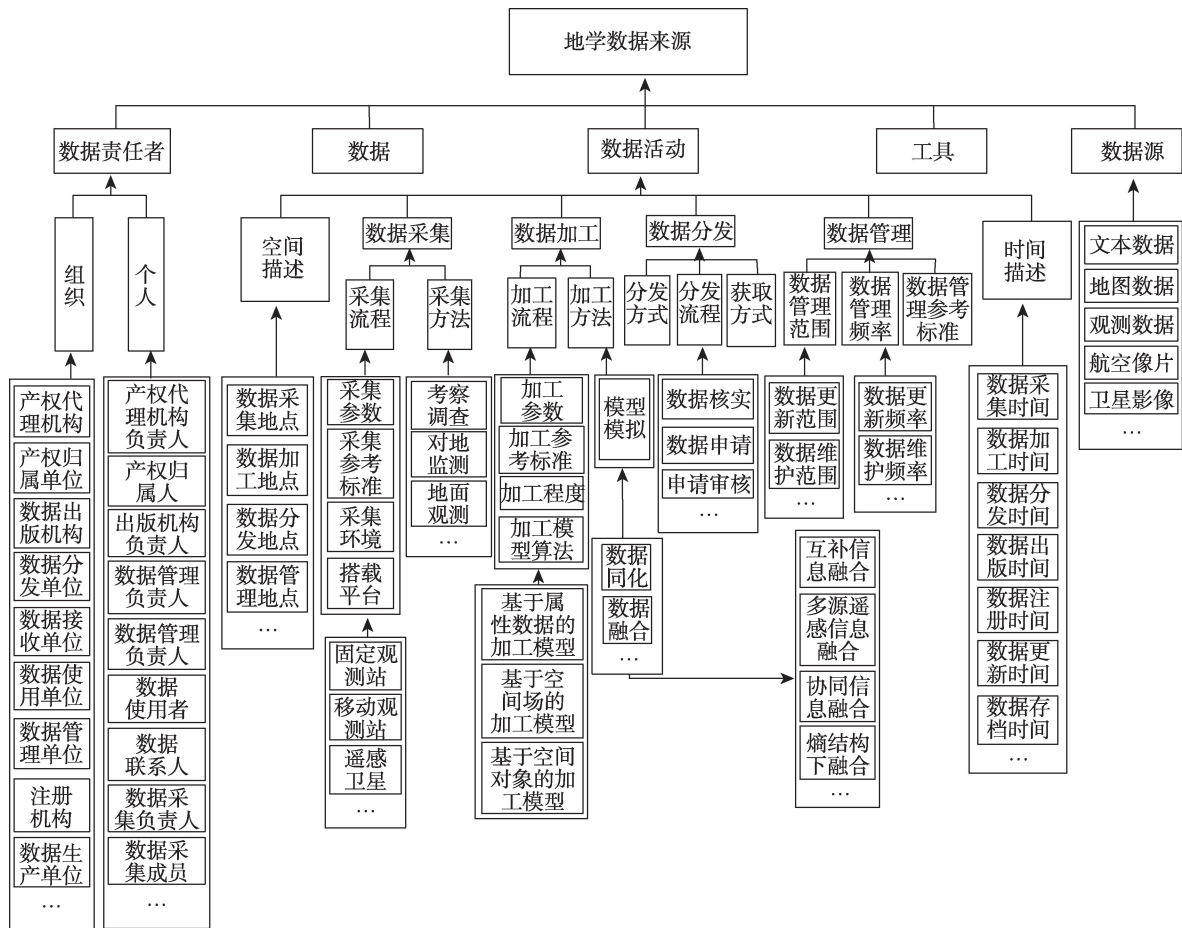


图2 地理空间数据来源概念体系

Fig. 2 The concept system of geospatial data provenance

些算法、模型以及对数据源的加工程度等都会对后期数据产品的质量造成影响。数据分发和数据管理在数据共享的过程中同样发挥着及其重要的作用,数据分发和数据管理的好坏对于用户能否顺利使用到可靠的数据有着密切联系,数据分发主要包括分发方式、方法流程以及获取方式;数据管理主要包括数据管理范围、数据管理频率以及管理过程中依照的参考标准。时间和空间是对采集、加工、分发、管理等数据活动的描述。时间包括数据采集时间、加工时间、分发时间、更新时间、存档时间等内容;空间包括数据采集地点、加工地点、分发地点、管理地点等内容。

### 3 地理空间数据数据来源本体建模

#### 3.1 地理空间数据来源本体建模原语

地理空间数据来源本体是对地理空间领域内共享的来源概念及其关系的明确的形式化规范说明。来源本体建模的核心是确定地理空间领域内来源的概念、属性、概念之间的关系。目前使用最为广泛且最为全面的是Perez等<sup>[21]</sup>提出的五元组方法。其利用分类法进行本体的组织,提出了本体的五元组建模原语,即本体可以表示为 $Ontology=\{C, R, F, A, I\}$ 。其中,C代表概念或者类;R代表了概念之间的关系,即在某一领域中概念之间的相互作用;F代表了函数,是一种特殊的关系,这种关系中的 $n-1$ 个元素可以决定第 $n$ 个元素;A代表公理;I代表实例。

结合五元组建模原语和地理空间数据来源的概念体系,本文将地理空间数据来源本体的逻辑结构定义为五元组模型,即概念、属性、关系、约束和实例。具体表示如式(1)所示。

$$P_{GDO}=\{PC, PR, PP, PC_{on}, PI\} \quad (1)$$

式中:PC(Provenance Concepts)表示来源本体的概念。PR(Provenance Relations)表示来源本体的概念与概念、概念与实例、实例与实例之间的关系。PP(Provenance Properties)表示来源本体的属性,包括对象属性和数据属性,对象属性表示实例与实例之间的关系,如数据A与它的数据源B之间的关系是“衍生”,数据属性主要表示实例与数值的关系,例如J6经纬仪的测角精度是6 s;PC<sub>on</sub>(Provenance Constraint)表示来源本体的规则集合。即对某一属性的值进行约束,让某一个属性值只能处于某一个

范围之内。PI(Provenance Individuals)表示来源本体的实例。

#### 3.2 来源本体概念间关系及属性的形式化表达

##### 3.2.1 来源本体概念间关系的形式化表达

来源本体概念间关系(对象属性)指的是在数据完整生命周期内,对所有来源要素间的相互联系和相互作用的描述,能够反映完整的来源过程,是地学领域内研究数据间相关性的重要基础。本文从来源过程中涉及动作的执行程度大小的角度对来源本体概念间的关系进行划分,主要划分为静态、半动态以及动态3种类型:①静态关系不涉及执行动作,侧重于描述对象间的关系,主要包括数据间的关系、数据责任者之间的关系、数据与数据责任者之间的关系;②半动态关系涉及的动作程度较弱,主要描述对象与过程间的关系,包括数据与数据活动间的关系和数据责任者与数据活动间的关系;③动态关系涉及较强的动作执行程度,重点描述过程间的关系,主要指多个数据活动之间的关系。针对上述3种类型关系,本文对来源实体间的具体关系进行汇总,关系较多,仅列出其核心关系,如表2所示。

为了更便捷地表示来源关系,本文针对来源本体概念间的关系定义一种形式化的表达方法,其表达式如式2所示。

$$P_{ATT-OBJ}(A)=WasAttributeTo\{B, C, D, \dots\} \quad (2)$$

式中:“P”代表Provenance,即来源,“ATT-OBJ”代表对象属性,即关系,P<sub>ATT-OBJ</sub>指的就是来源关系;“A、B、C、D”代表数据、数据活动、数据责任者中任意一个,“B、C、D”必须属于同一类,如“B、C、D”都代表数据或者数据活动;“WasAttributeTo”代表的是数据“A”与“B、C、D”等多个数据责任者之间的关系是“属于”。

##### 3.2.2 来源本体属性的形式化表达

来源本体属性(数据属性)指的是来源实体的性质,本文从数据性质角度对属性进行划分,主要分为3类:定位属性、定量属性、定时属性。定位描述的是实体的位置特征,例如“某个企业的地址为江西省南昌市高新二路18号”;定量反映的是实体的数量特征,例如“某个采集仪器的工作范围是300 m”;定时描述的是实体的时间特性,例如“某个企业的成立时间为2015年8月”。综合上述3类属性,对核心属性进行汇总,部分来源本体属性示例如表3所示。

由于来源本体所包含的属性较多,为了更便捷地表示实体的特性,本文针对属性提出一种形式化

表2 来源本体概念间的核心关系  
Tab. 2 The core relations between provenance entities

关系	关系简述	图示
引用	多个数据源合并成一个新数据,侧重于数据的复制,新的数据中存在旧的数据源	
更新	在已有数据上添加新的信息	
融合	多个数据源合成一个新数据,新数据中不存在旧的数据源	
修订	修复数据中的某些错误	
衍生	单个数据经过加工后生产新的数据,侧重于数据一对一的形成	
使用	利用已有数据源进行数据活动,利用数据前,数据活动不会被数据源所影响	
生成	通过数据活动完成新数据的生产,生产之前不存在,生产之后可供使用,主要针对原始数据的产生	
共生	数据生产过程中,涉及多个数据活动,相互之间缺一不可	
授权	数据责任者A委托数据责任者B进行数据活动	
属于	数据责任者对数据具有所有权	
负责	数据责任者在数据活动中承担任务或者责任	
贡献	数据责任者参与数据活动,对数据的生成起有利作用	

表达方法,如式(3)所示。

$$P_{ATT-DAT}(A)(B)(C)(D)=Value \quad (3)$$

式中:P代表来源;ATT-DAT代表数据属性;A代表实体名称;B代表属性类名,属性类名可存在多个,存在多个类名时,前后类名是子类关系;C代表属性名称;Value代表属性值,如 $P_{ATT-DAT}(Tools)(ProcessingTools)(ArcGIS)(Version)=10.4$ ,该公式表示在“工具”实体中,ArcGIS版本号为10.4,该属性为加工工具的属性。

## 4 地理空间数据来源本体构建与应用实践

### 4.1 地理空间数据来源本体的构建实践

本体的构建是根据本体模型,确定出该领域内

研究对象所涉及的概念、属性、关系及规则,然后利用本体构建工具建立基于某种本体描述语言的本体文件,以便于在数据关联等应用中使用,本体的构建是个系统化、工程化的过程,在构建的同时,需要确保本体的可扩充、可维护。

地理空间数据来源本体的构建采用owl本体描述语言,以来源概念模型为基础,按照上述方法将上文中所提到的概念、属性、关系等进行明确表达。来源本体的构建充分考虑了对各大标准及模型的继承,如“IOS19115”和“OPM模型”等,从而保证来源本体概念设计合理、集成依赖程度最小以及良好的可扩展性。本文利用开源的Protégé工具实现来源本体的构建,在类模块中构建出来源概念;在对象属性模块中添加来源概念间的关系,在数据属性中添加来源本体所涉及的属性,由于来源本体属于应用本体,即在应用时才能找到实例,因此在



表3 来源本体中的核心属性示例  
Tab. 3 The core properties of provenance ontology

实体名	属性类名	属性名
工具	采集仪器	型号
		唯一标识
		标称精度
		应用领域
		工作范围
	加工工具	采集对象
		运行环境
		工具版本
		加工精度
		模型提出者
数据责任者	个人	模型提出时间
		模型版本
		职务
		入职时间
		联系方式
		联系地址
		员工编号
		所在部门
	机构	单位法人
		成立时间
		单位类型
		单位规模
		业务范围
		服务时间

工具中暂不建立实例。目前,已构建了地理空间数

据来源本体的100多个概念、20多种关系以及30多个属性,详细情况如图3所示。

4.2 地理空间数据来源本体在数据关联中的应用

本文从“科技基础性工作专项”项目数据资料中挑选了来自多个不同项目的数据集为例,将来源本体库作为参考,人工提取数据集中的来源信息,并构建来源关联规则。本文主要对数据进行了2个方面的关联,即数据间的关联和数据与其来源信息之间的关联。

在数据间关联方面,本文使用web前端框架技术D3.js对数据间关系进行可视化,如图3所示;利用RDF(Resource Description Framework)对数据进行关联,关联规则主要有6种,如表4所示。为了方便说明数据间的关系,数据名称以大写英文字母代替。从图4和表4中,清晰地反映了各个数据间的多种关联关系和关联规则,利用该关联网,从来源角度进行数据间的关联,能够支持地理空间数据的语义检索和智能推荐等应用,有利于解决数据的语义异构造成的基于关键字匹配的数据检索效率低的问题。

在数据与来源信息关联方面,从图4可看出,数据集为“典型县高分辨率土地覆被分类数据”的全

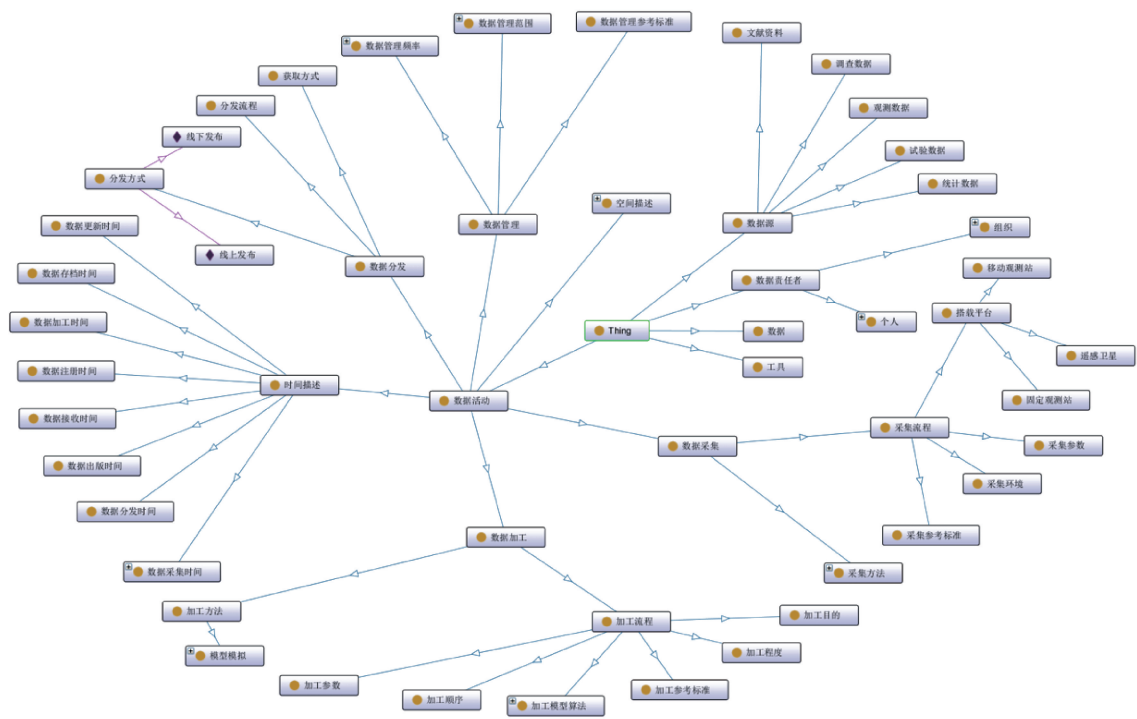


图3 地理空间数据来源本体可视化

Fig. 3 Geospatial data provenance-ontology visualization

表4 数据关联规则  
Tab. 4 Data linking rules

数据	关联的数据	关联规则
A	B、C	B、C是A的数据源
F	D、E、G	F是D、E、G的数据源
H	I、J	I、J是H的数据源
K	L	K是L的数据源
A	K	同一加工工具
A	D	来自同一个项目
H	D、N、G、K	产权归属单位相同
N	S、M	产权归属单位相同
G	P、Q、R	依托单位相同
K	M	依托单位相同
N	O	依托单位相同
P	Q	P和Q的采集方法相同
M	U、V	M和U、V的采集方法相同

部加工过程为:基于 SPOT 和 Rapideye 遥感影像数据的近红外、红、绿、蓝4个波段,首先在 ENVI 平台上对2个影像进行投影转换和格式转换,然后使用 eCognition 平台对影像进行分类处理和人机交互解译,最后使用 ArcGIS 软件平台进行分类后处理,并

且构建误差混淆矩阵进行精度验证。因此,将数据与其来源信息进行关联,实现来源信息的规范化,对提升数据的可靠性评估等服务具有重大意义。

5 结语

本文以地理空间数据来源作为研究对象,深入分析了地理空间数据来源的涵义和重要概念;建立了地理空间数据来源本体模型及其概念体系;提出了来源本体五元组形式化表达模型,构建形成了由100多个概念、20多种关系组成的地理空间数据来源本体。最后,以“科技基础性工作专项”项目数据资料为例,进行了基于来源本体的数据关联实践。通过关联数据网络能够清晰地反映数据详细的来源信息,用户可以直观地看出数据间的多种关联关系,整个数据采集、加工过程能够完整呈现,从而解决数据来源信息描述不规范的问题,帮助用户对数据的可靠性进行判定,同时有效支持地理空间数据的语义检索等应用。因此,将来源本体应用到地理

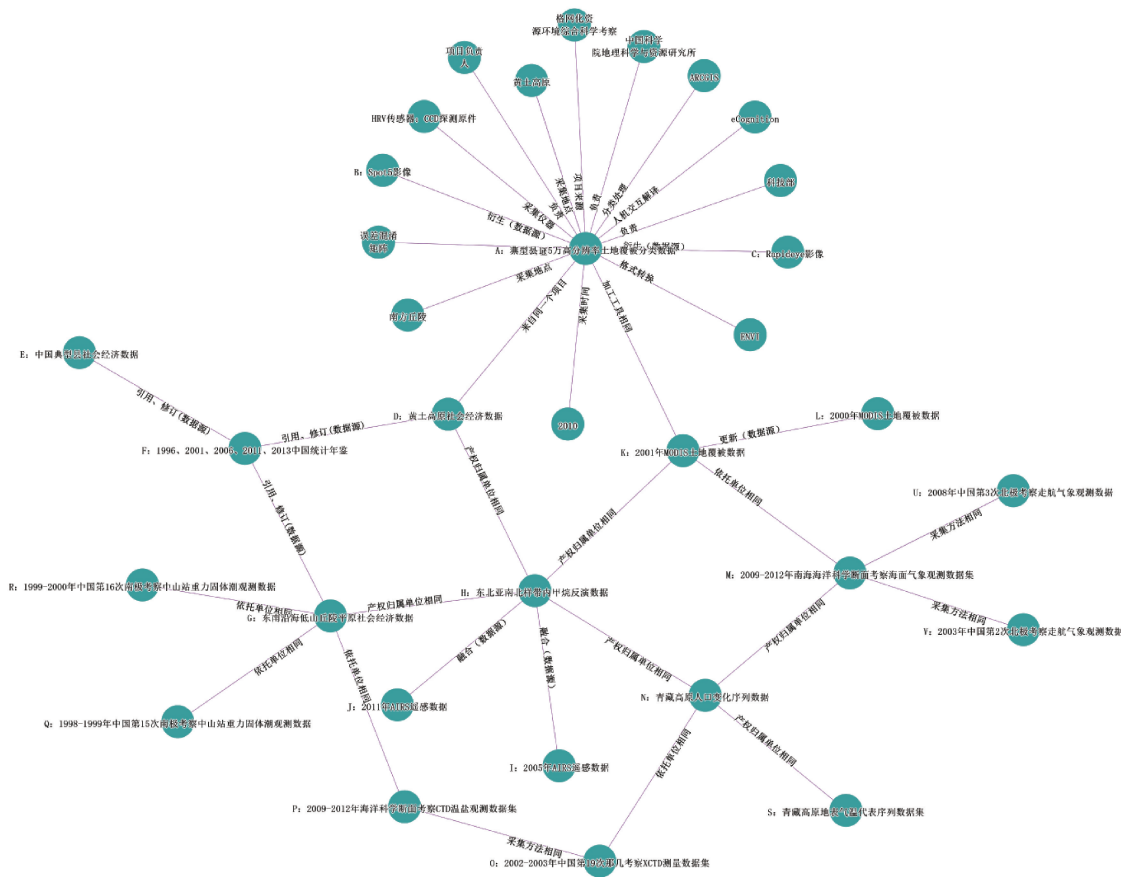


图4 基于来源信息的地理空间数据关联网络

Fig. 4 Geospatial data linking network based on provenance information



空间数据的数据关联中,对于促进数据共享和利用有着重要的实用价值。

基于地理空间数据来源本体的数据关联应用仍然存在很多需要解决的难题:来源本体的概念体系还有待完善;来源本体的构建、来源信息的获取仍然依赖于人工操作,尚未实现自动化,针对来源信息变更和扩展来源本体实现自动更新的技术尚未实现;全面覆盖地理空间数据来源、时间、空间、内容、形态等完整的数据关联指标体系研究进展缓慢,尤其针对地理空间数据来源语义的关联指标体系,即地理空间数据来源关联权重的计算研究等。

#### 参考文献(References):

- [1] Studer R, Benjamins V R, Fensel D. Knowledge engineering: Principles and methods[J]. Data & Knowledge Engineering, 1998,25(1-2):161-197.
- [2] ISO19115- 2- 2009, Geographic Information- Metadata [S].2009
- [3] Di L, Yue P, Ramapriyan, et al. Geoscience data provenance: An overview[J]. IEEE Transactions on Geoscience & Remote Sensing, 2013,51(11):5065-5072.
- [4] 戴超凡,王涛,张鹏程.数据起源技术发展研究综述[J].计算机应用研究,2010,27(9):3215-3221. [ Dai C F, Wang T, Zhang P C. Survey of data provenance technique[J]. Application Research of Computers, 2010,27(9):3215-3221. ]
- [5] 陈颖.一种基于DNA双螺旋结构的数据起源模型[J].现代图书情报技术,2008,24(10):11-15. [ Chen Y. A data provenance model based on the double-helical structure of DNA[J]. Data Analysis and Knowledge Discovery, 2008,24(10):11-15. ]
- [6] 李文燕,吴振新.起源信息模型及标准PROV的研究分析[J].情报理论与实践,2015,38(4):23-29. [ Li W Y, Wu Z X. Research and analysis of provenance information model and standard PROV[J]. Information Studies: Theory & Application, 2015,38(4):23-29. ]
- [7] Moreau L, Freire J, Futrelle J, et al. The open provenance model: An overview. proceedings of the provenance and annotation of data and processes, F[C]. 2008.
- [8] Sahoo S S, Sheth A P. Provenir ontology: Towards a framework for eScience provenance management. proceedings of the Knoesis Publications, F[C]. 2009
- [9] Doerr M, Theodoridou M. CRMdig: A generic digital provenance model for scientific observation; proceedings of the TaPP, F[C]. 2011.
- [10] LEBO T. PROV-O: The PROV ontology:W3C recommendation 30 April 2013[J]. Journal of Surgical Research, 2013,147(2):194-199.
- [11] Foster I, Vockler J, Wlode M, et al. Chimera: a virtual data system for representing, querying, and automating data derivation; proceedings of the International Conference on Scientific and Statistical Database Management, 2002 Proceedings, F[C]. 2002.
- [12] Myers J D, Pancerella C, Lansing C, et al. Multi-scale science: supporting emerging practice with semantically derived provenance[C]. Proceedings of the ISWC 2003 Workshop on Semantic Web Technologies for Searching and Retrieving Science Data, Sanibel Island, Florida, October 2003.
- [13] Stevens R D, Robinson A J, Goble C A. MyGrid: Personalised bioinformatics on the information grid[J]. Bioinformatics, 2003,19Suppl 1(suppl\_1): i302-4.
- [14] Miles S, Groth P, Branco M, et al. The requirements of using provenance in e-Science experiments[J]. Journal of Grid Computing, 2007,5(1):10-1007.
- [15] Bowers S, Mcphitips T, Riddle S, et al. Kepler/pPOD: Scientific workflow and provenance support for assembling the tree of life. Proceedings of the provenance and annotation of data and processes, second international provenance and annotation workshop, IPAW 2008, Salt Lake City, UT, USA, June 17- 18, 2008 Revised Selected Papers, F[C]. 2008.
- [16] 贾君枝,寇蕾蕾.基于W7模型的数据起源本体语义分析[J].情报理论与实践,2016,39(3):118-121. [ Jia J Z, Kou L L. Semantics analysis of data provenance ontology based on W7 model[J]. Information Studies:Theory & Application, 2016,39(3):118-121. ]
- [17] Ram S, Liu J. A new perspective on semantics of data provenance. Proceedings of the international workshop on the role of semantic web in provenance management, F [C]. 2009.
- [18] Hartig O. Provenance information in the web of data. Proceedings of the Linked Data on the Web Ldow Workshop at WWW, F[C]. 2011
- [19] Hunter J, Cheung K. Provenance explorer: A graphical interface for constructing scientific publication packages from provenance trails[J]. International Journal on Digital Libraries, 2007,7(1):99-107.
- [20] 乐鹏,彭飞,龚健雅.基于SOA的空间数据起源研究[J].地理与地理信息科学,2010,26(3):6-10. [ Yue P, Peng F F, Gong J Y. Research on SOA-based geospatial data provenance[J]. Geography and Geo-Information Science, 2010,26(3):6-10. ]
- [21] Benjamins V R, Gómez A. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods[J]. Pérez, 1999,8(1):11-1.