

DOI: 10.16108/j.issn1006-7493.2021111

引用格式: 诸云强, 孙凯, 李威蓉, 王曙, 宋佳, 程全英, 杨杰, 牟兴林, 耿文广, 代小亮. 2023. 地球科学知识图谱比较分析与启示: 构建方法与内容视角[J]. 高校地质学报, 29(3): 382–394

地球科学知识图谱比较分析与启示: 构建方法与内容视角

诸云强^{1,2}, 孙凯^{1*}, 李威蓉^{1,3}, 王曙¹, 宋佳^{1,2}, 程全英^{1,3}, 杨杰¹,
牟兴林⁴, 耿文广⁵, 代小亮^{1,3}

1. 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;

2. 江苏省地理信息协同创新中心, 南京 210023; 3. 中国科学院大学, 北京 100049;

4. 自然资源部国土卫星遥感应用中心, 北京 100048; 5. 山东理工大学建筑工程学院, 淄博 255000

摘要: 地球科学(以下简称地学)知识图谱将地学知识以有向图的方式进行形式化表达, 具有强大的知识表达能力、开放互联能力和推理预测能力, 是地学与人工智能交叉融合发展的基础设施之一, 已成为当前地学研究中重要的研究热点。因此, 国际上很多科学组织或团队先后开展了地学领域的知识图谱研究, 并构建了一系列具有代表性的知识图谱。然而, 目前尚缺乏对这些知识图谱的深入研究和分析。文章从基本情况、构建方法、主要内容及特点等方面, 对当前国际上主要的地学领域知识图谱进行了系统的比较分析, 并在此基础上, 指出了对未来地学知识图谱研究的启示: 从构建方法上, 应构建地学知识图谱统一表达模型, 建立融合多源、多模态数据的知识源, 研究地学知识表示与计算方法; 从内容上, 应加强地学知识时空特征描述, 考虑地学知识复杂时空关系和推理规则; 从应用上, 应发展地学知识质量评估和修正方法, 提升地学知识图谱应用成效。

关键词: 地学知识图谱; 地学知识; 地学本体; 形式化表达; 人工智能

中图分类号: P628+.4

文献标识码: A

文章编号: 1006-7493(2023)03-382-13

Comparative Analysis and Enlightenment of Geoscience Knowledge Graphs: A Perspective of Construction Methods and Contents

ZHU Yunqiang^{1,2}, SUN Kai^{1*}, LI Weirong^{1,3}, WANG Shu¹, SONG Jia^{1,2}, CHENG Quanying^{1,3}, YANG Jie¹,
MU Xinglin⁴, GENG Wenguang⁵, DAI Xiaoliang^{1,3}

1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 10010, China;

2. Jiangsu Center for Collaborative Innovation in Geographic Information Resource Development and Application, Nanjing 210023, China;

3. University of Chinese Academy of Sciences, Beijing 100049, China;

4. Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China;

5. School of Architecture Engineering, Shandong University of Technology, Zibo 255000, China

Abstract: Geoscience knowledge graphs (GKGs) formally represent geoscience knowledge in a way of directed graph and have strong capabilities in knowledge representation, openness and interconnectivity, and reasoning and prediction. GKGs have been one of the important infrastructures for the development of combining geoscience and artificial intelligence, thereby

收稿日期: 2021-10-22; 修回日期: 2021-12-21

基金项目: 国家重点研发计划项目(2022YFB3904200); 国家自然科学基金原创探索计划项目(42050101)联合资助

作者简介: 诸云强, 男, 1977年生, 博士, 研究员, 主要从事地理空间数据本体和地学知识图谱等相关研究; E-mail: zhuyq@lreis.ac.cn

*通讯作者: 孙凯, 男, 1990年生, 博士后, 研究方向是地学知识图谱构建及应用; E-mail: sunk@lreis.ac.cn

becoming one of the important research focuses in geoscience. Therefore, many international scientific organizations or groups have successively carried out studies in this domain, and constructed some representative GKGs. However, there is a lack of an in-depth study and analysis of these existing GKGs. To this end, this paper makes a systematic comparative analysis on their general information, construction methods, and main contents. On this basis, some enlightenments about future research of GKGs are discussed. In terms of the construction method, a unified representation framework for GKGs should be built, the source of geoscience knowledge should be enhanced by conflating multi-source and multimodal data, and methods for the representation and computation of geoscience knowledge should be studied. Regarding the contents of GKGs, complex spatio-temporal characteristics, relations, and reasoning rules should be considered. From the perspective of application, methods for assessing quality and making correction for geoscience knowledge should be developed, and application effects of GKGs should be improved.

Key words: geoscience knowledge graphs (GKGs); geoscience knowledge; geoscience ontology; formal representation; artificial intelligence

Corresponding author: SUN Kai, Postdoc Researcher; E-mail: sunk@reis.ac.cn

现代地球观测手段的不断丰富使得地球科学（以下简称地学）研究进入了空天地、全天候、全要素覆盖和观测的地学大数据时代（周成虎等，2021）。在地学大数据的驱动下，地学研究进入数据密集型的第四科研范式（Hey et al., 2009），开始逐渐注重从海量数据中挖掘和发现地学规律，并形成了地学大数据挖掘的系统理论与方法（裴韬等，2019）。与此同时，人工智能的技术进步和迅速发展给地学研究带来新的发展动力，地学研究逐步进入与人工智能的交叉融合发展，并形成地学人工智能的重大交叉学科研究方向，聚焦于利用人工智能技术提升地学研究的自动化和智能化（高松，2020）。

无论是地学大数据挖掘还是地学人工智能的发展均需要系统化、形式化、规范化的地学知识的支撑（陆锋等，2017）。地学大数据挖掘需要地学知识驱动以发现更多隐藏在结构化或非结构化的地学数据中的新知识；地学人工智能则需要可计算、可推理的地学知识作为基础设施，辅助人工智能深刻理解地学规律，提升其地学思维能力（高松，2020）。然而，地学研究涉及地理学、地质学、地貌学、海洋科学、大气科学等分支学科，具有极为庞大和复杂的学科知识体系（孙鸿烈，2017），如何自动化地获取地学知识，并将其用一种形式化的方式进行表达，以能够支撑地学知识的计算和推理，成为亟待解决的问题。

知识图谱作为一种高效的知识挖掘、组织和表示方式，可为上述问题提供完整的解决方案（周成

虎等，2021；王志华等，2021）。知识图谱将知识以有向图（节点-边）的方式进行组织，并以形式化、机器可理解的语言进行表示（Hartig, 2017），可支持知识的智能查询与推理（Hey et al., 2009）。正因为知识图谱强大的知识表达、查询和推理能力，它在通用领域的智能问答与搜索、个性化推荐等方面已有广泛应用（徐增林等，2016；Shin et al., 2019；Shao et al., 2021）。地学知识图谱（Geoscience Knowledge Graph, GKG）可视为知识图谱技术在地学领域的延伸（Sun et al., 2020），主要通过结构化和形式化的方式来组织、表示和存储地学知识，并服务于基于地学知识的查询发现、推荐等应用（陆锋等，2017；Zhang et al., 2018；Zheng et al., 2021）。

现有地学知识图谱研究主要从方法和应用两个方面展开。方法上，考虑地学知识图谱的构建过程，从知识抽取（包括实体抽取（Wang et al., 2018；Qiu et al., 2019b；Wang et al., 2020）和关系抽取（余丽等，2016a；余丽等，2016b））、知识融合（Sun et al., 2019；Sun et al., 2020）、知识表示（Mai et al., 2020；Santos et al., 2018；Gao and Yan, 2018）、知识补全（Qiu et al., 2019a）和知识质量评估（Senaratne et al., 2017；Degrossi et al., 2018）等方面开展研究。应用上，在地学信息的智能关联和发现（赵红伟等，2016）、推荐（Zhu et al., 2017）以及可视化（蒋秉川等，2020）等方面已有系列研究。这些研究以及国际上很多科学组织或团队构建了很多地学知识图谱，可分为任务型和通用型两种。前者通常用于为具体应用任务提供知识服务；后者则是为地学领域

的某方面知识提供统一的知识框架。在任务型地学知识图谱研究方面,蒋秉川等和陈晓慧等均以新型冠状病毒肺炎的大量疫情时空数据为数据源,构建疫情地学知识图谱,基于此,对病患关联关系、疫情防控态势等进行分析(蒋秉川等,2020;陈晓慧等,2020)。刘俊楠等则借助于维基百科构建地学知识图谱以完善地学实体的语义信息(刘俊楠等,2020a,b)。此外,还有地学数据关联知识图谱和应急减灾知识图谱(赵红伟等,2016;陶坤旺等,2020;杜志强等,2020)等研究。然而,这些研究仅是针对特定场景的案例应用,其中绝大多数随着学术论文的发表而停止维护和更新,无法从整体上推动地学与知识图谱的融合发展。

相比来说,通用型地学知识图谱,例如 GeoNames Ontology、OSM (OpenStreetMap) Semantic Network 等,通常包含大量地学知识,且由稳定的社区或团体维护和运营,可对地学知识图谱的发展起到极大的推动作用。这些地学知识图谱,采用了哪些构建技术方法、侧重于哪方面的主题内容、知识来源是什么等?这些问题对未来地学知识图谱研究至关重要,其回答需要对已有知识图谱的构建方法及内容等做全面系统的比较分析,但是目前缺乏相关研究。为此,本文针对现存主要的地学知识图谱,从其基本情况、构建方法及内容等方面进行深入的比较分析和研究,基于此,指出当前地学知识图谱研究中存在的问题以及对本领域未来发展的启示。本研究的重要意义体现在:首先,基于对现有地学知识图谱构建方法和内容的系统分析,可为未来地学知识图谱的构建提供借鉴;其次,从中获取的启示可指引和加速地学知识图谱及其高级应用的发展和实现;最后,对于刚刚进入或者准备进入本领域的研究人员或组织,可帮助其全面掌握已有大型地学知识图谱的概况,以快速融入本领域研究。

1 地学知识图谱

地学知识图谱可视为地学实体及其属性和实体间相互关系的集合(陆锋等,2017),可形式化为式(1):

$$GKG = \langle E, R, P \rangle \quad (1)$$

其中 E 表示实体,主要包含概念和实例。概念,又称类,是指一类事物或对象的统称,例如平原。实

例则是概念具体的实现,例如华北平原。 R 表示实体间的关系,主要是时空关系以及语义关系,例如华北平原与太行山、秦岭等山脉以及与北京、天津等地间的空间关系。 P 表示实体的属性,指对实体特征的定义和描述,包含本质属性以及其他描述属性,例如华北平原的面积、空间位置等。具体到实现上,所有知识都可采用资源描述框架(Resource Description Framework, RDF)的主谓宾三元组表示(式(2))。

$$Triple = \langle Subject, Predicate, Object \rangle \quad (2)$$

从知识的不同层次上区分,地学知识图谱可分为模式层和实例层(刘峤等,2016),如图1所示。其中模式层,又称本体层,指知识图谱所涉及的概念、属性、关系和规则等的形式化定义,它构成知识图谱的概念模型和逻辑基础;实例层,又称数据层,是模式层的实例化,包含概念的实例及其属性,实例间的关系等。模式层是实例层的约束和抽象化;实例层是模式层的实现。地学知识图谱可仅包含模式层,此时可理解为地学本体;也可仅包含实例层,此时由于没有模式层对于关系和规则等的定义,将无法支持推理。

2 材料与方法

本文采用 Kitchenham 和 Charters (2007) 提出的文献综述方法对现有地学知识图谱进行比较分析,共包含三个步骤。(1)研究对象初步筛选。典型文献综述以文献分析或总结为目的。与此不同,本文的目标并不是对地学知识图谱相关文献进行综述,而是对典型地学知识图谱进行比较分析,因此研究对象筛选的目的是找出具有代表性的地学知识

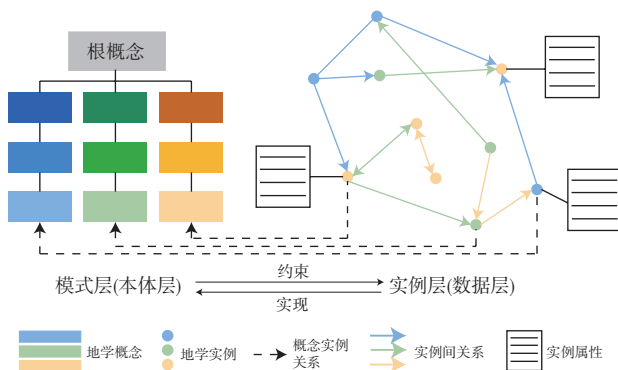


图1 地学知识图谱示意图

Fig. 1 The illustration for a geoscience knowledge graph

图谱。同时,本文主要聚焦在具有一定知识规模的通用型地学知识图谱上,对于面向案例的小型知识图谱不做研究。然而,大多数相关文献是研究案例型的地学知识图谱,鉴于此,本文未通过文献检索来发现研究对象,而是基于专家知识初步确定现有的地学相关知识图谱。(2)研究对象确定。指根据本文地学知识图谱比较分析的目标,按照制定的原则,从现有地学相关知识图谱中进一步选择典型目标作为研究对象。(3)比较分析。针对研究对象,结合数据解析和统计、文献资料阅读和理解等方式从构建方法和内容视角对地学知识图谱进行比较分析。

2.1 研究对象初步筛选

首先由5位本领域专家(覆盖固体地球科学、地理学等学科领域)按照知识图谱的代表性分别列举主要的地学知识图谱,经过分析讨论,初步筛选出16个研究对象,包含8个地学知识图谱:GeoSciML (Geoscience Markup Language)、OSM Semantic Network、GeoNames Ontology、LinkedGeoData、SWEET、GeoWordNet、LinkedEarth Ontology 和 Global Change Information System (GCIS) Ontology;8个通用领域知识图谱:CYC、DBpedia、YAGO、Freebase、Nell (Never-Ending Language Learner)、Wikidata、Knowledge Vault 和 Microsoft Concept Graph。本文将通用领域知识图谱包含在初步筛选的研究对象中,其原因在于,通用知识图谱发展较早,也更为成熟,通过与通用知识图谱的比较,可在多个方面对地学知识图谱的构建及应用有所启发:方法上,通用知识图谱构建过程中使用大量知识抽取、对齐融合等自动化方法,可为地学知识图谱的自动构建关键技术研发提供借鉴;时空信息表达上,通用知识图谱中时空信息的表示可为地学时空信息的表示提供启示;应用上,通用知识图谱已有广泛应用,可一定程度上启发未来地学知识图谱的应用推广。为了对上述知识图谱进行比较分析,还进一步获取其相关资料,主要包括知识图谱官方网站、知识图谱数据文件、相关文献及报告等。

为了进一步从上述知识图谱中遴选典型目标作为本文的研究对象,首先对上述知识图谱的基本情况进行研究,主要包括:知识图谱的定位目标、创建时间、创建者及所在国家、知识规

模、是否开源以及最新版本(本文后续以最新版本的知识图谱作为研究分析的对象)等。上述信息主要从知识图谱的官方网站获取得到,结果如表1所示。上述知识图谱的定位目标可大致分为两类,一类是知识的共享利用,另一类则是数据的共享集成。GeoNames Ontology、OSM Semantic Network、LinkedGeoData 和 GeoWordNet 均以地理信息的共享为目标,而 SWEET 则主要共享地学知识体系。全部8个通用领域知识图谱主要聚焦在开放网络资源蕴含知识的共享上。相比来说,GeoSciML、LinkedEarth Ontology 和 GCIS Ontology 分别以地学数据、古气候数据和全球变化相关数据资源的集成共享为目标。

从创建时间上看,除 CYC 外,其他知识图谱均创建于2000年以后。严格意义上,CYC 并不是知识图谱项目,而仅是一个知识库。2000年前后,本体及知识工程研究开始迅速发展,以此为时间节点,地学知识图谱陆续出现(Li et al., 2017)。从创建者看,创建人或团队主要来自发达国家,发展中国家无论在通用知识图谱或地学知识图谱的研究上均落后于发达国家。知识图谱的知识规模差别较大,其主要原因在于知识获取方式的不同。GeoSciML、SWEET、CYC 等依赖人工构建的知识图谱规模较小;OSM Semantic Network、YAGO、Freebase 等采用众包技术或自动知识抽取方法获取知识,具有更大的知识规模,可达数亿甚至数十亿的三元组。是否可开源获取对于知识图谱的共享利用极为重要,目前,除 Knowledge Vault 和 CYC 外,其他均可开源获取。版本信息通常预示着知识图谱维护更新的频度,总体来看,通用知识图谱的更新频率高于地学知识图谱。

2.2 研究对象确定

初步筛选的16个知识图谱中存在数据重复、活跃度低等问题,需根据一定原则进行剔除。本文考虑如下剔除原则:(1)通用性强:指研究对象需为通用型地学知识图谱,面向案例的任务型知识图谱将被剔除;(2)数据不重复:指当有多个知识图谱的数据相同时,仅保留其中一个;(3)可开源获取:指知识图谱可通过文件下载或查询服务等形式自由开放获取,无法开源获取的将被剔除;(4)仍处于活跃期:指知识图谱仍在维护或更新中,已

表1 初步筛选的知识图谱基本情况
Table 1 General information of all the selected knowledge graphs

知识图谱	发布网址	定位目标	创建时间	创建者及所在国家	知识规模	是否开源	最新版本
GeoSciML	http://geosciml.org/	创建一套服务于地学数据共享传输的数据模型	2003	国际地球科学信息委员会, 一	包含1772个概念	是	4.1
OSM Semantic Network	https://www.openstreetmap.org/	创建一个内容自由的、免费的且能让所有人编辑的世界地图。	2004	Stephen Coast, 英国	60亿个点状要素, 6.89亿个线状或面状要素。	是	2014*
GeoNames Ontology	http://www.geonames.org/	创建覆盖全球的地名词典。	2005	Marc Wick, 瑞士	全球约1200万个地理实体的约2500万个地名。	是	3.3
LinkedGeoData	http://linkedgeodata.org/	以OpenStreetMap为数据源, 创建大型知识库。	2009	莱比锡大学, 德国	超过30亿个节点和3亿条边, 约200亿个三元组。	是	2016*
SWEET	https://github.com/ESIPFed/sweet	创建一套描述地球科学领域的本体库。	2009	国家航空航天局, 美国	4533个概念以及359个属性。	是	3.5.0
GeoWordNet	—	通过集成WordNet和GeoNames形成一个语义信息更丰富的地名词典。	2010	特伦托大学, 意大利	3698238个实例, 334个概念, 182个概念间关系。	是	2016*
GCIS Ontology	https://data.globalchange.gov/gcis.owl	集成开源和基于网络的资源, 以协调和整合全球环境变化相关的数据资源。	2013	全球变化研究计划, 美国	2106个报告、7406篇文章、1277个期刊、3220个数据集等。	是	2.0
LinkedEarth Ontology	http://linked.earth/ontology/	以支撑古气候研究为应用场景, 创建古气候数据集成和归档的语义平台。	2015	国家自然科学基金会, 美国	6个子本体, 148个概念, 55种关系	是	1.2.0
CYC	https://cyc.com/knowledge-layer/	创建一个包含人类常识背景知识的本体库。	1984	Douglas Lenat, 美国	50万条术语和700万条断言。	否	2017*
DBpedia	https://www.dbpedia.org/	以Wikipedia为信息源, 从中提取结构化的知识并构建知识图谱。	2007	莱比锡大学等, 德国	1219个本体, 2.2亿个实体, 14.5亿个三元组。	是	Largest Diamond
YAGO	https://yago-knowledge.org/	创建包含人、城市、国家和组织等通用性知识知识图谱。	2007	Max Planck信息学研究所, 德国	超过5000万实体, 20亿事实。	是	4
Freebase	https://developers.google.com/freebase	创建一个允许所有人(机器)快捷访问的知识库。	2007	Metaweb公司, 美国	超过1.25亿个三元组, 4000个概念, 7000种属性。	是	2013*
Nell	http://rtw.ml.cmu.edu/rtw/	通过自学习的方式不断从网络资源上学习和抽取新的知识。	2010	卡内基梅隆大学, 美国	2810379个实例, 1186个概念及关系。	是	2018*
Wikidata	https://www.wikidata.org/wiki/Wikidata:Main_Page	创建一个可自由协作编辑的结构化知识库, 为Wikimedia项目提供支撑	2012	Wikimedia基金会, 美国	超过12亿个三元组, 超过9500万个实体。	是	2021*
Knowledge Vault	https://developers.google.com/knowledge-graph	从互联网数据中抽取知识并构建知识图谱。	2014	谷歌公司, 美国	16亿个三元组, 4500个概念, 4469种关系。	否	—
Microsoft Concept Graph	https://concept.research.microsoft.com/Home/Introduction	以数以亿计的网页和数年积累的搜索日志为数据源, 创建知识图谱。	2016	微软, 美国	超过1255万个实体, 540万个概念以及8760万个关系。	是	v1

* 表示该知识图谱未提供版本信息, 因而以最新版本的更新年份指代版本信息; —表示未发现相关信息。

完全不再维护或更新的知识图谱将被剔除; (5) 包含丰富的地学知识: 指知识图谱需包含大量地学知

识, 因此, 完全不包含或者只包含少量地学知识知识图谱将被剔除。

根据上述原则,按序剔除不符合要求的知识图谱,剔除过程如图2所示。第一步,考虑通用性原则。LinkedEarth 为面向古气候数据的任务型知识图谱,通用性低,因此被剔除。第二步,考虑数据不重复原则。LinkedGeoData 以 OpenStreetMap 为数据源 (Stadler et al., 2012; Auer et al., 2009),通过建立数据与 RDF 的映射关系,将其进行 RDF 转换,因此它与 OSM Semantic Network 数据相同,本文保留 OSM Semantic Network,剔除 LinkedGeoData。第三步,根据开源获取原则,表1中 Knowledge Vault (Dong et al., 2014) 和 CYC (Matuszek et al., 2006) 无法开源获取,因此被剔除。第四步,考虑对于活跃度的要求,GeoWordNet 已不再维护和更新 (Giunchiglia et al., 2009; Giunchiglia et al., 2010),因此被剔除;Freebase 目前也已不再维护和更新,但考虑到 Freebase 在开放知识库发展中的重要性,予以保留。第五步,Nell 和 Microsoft Concept Graph 中未发现大量地学相关的实例以及时空关系等知识 (Mitchell et al., 2018),因此被剔除。经过上述过程后仍然保留的5个地学知识图谱 GeoSciML、OSM Semantic Network、GeoNames Ontology、SWEET 和 GCIS Ontology 以及4个通用知识图谱 DBpedia、YAGO、Freebase 和 Wikidata,共同构成本文地学知识图谱比较分析的目标对象。

3 比较分析

构建方法和内容是地学知识图谱研究的核心,

因此本节对现有地学相关知识图谱的比较分析主要聚焦在这两个方面。

3.1 构建方法对比分析

本节分别从知识来源、构建方式、自动化程度和序列化格式四个方面比较分析现有地学知识图谱的构建技术方法及其实现模式,并总结其特点和优劣。

3.1.1 知识来源

地学知识图谱的知识源可大体分为两类:(1)人工知识源:指可以为知识图谱贡献知识的人,主要包含领域专家和众包社区的志愿者;(2)数据知识源:指蕴含知识的数据,按其结构化程度可分为:结构化的已有知识库、表格类数据,半结构化的百科类网页以及非结构化或结构化程度较低的普通网页、语料库等 (黄恒琪等, 2019)。

9个地学相关知识图谱的知识来源如表2所示。针对5个地学知识图谱,从模式层看,GeoSciML、GeoNames Ontology 和 GCIS Ontology 的主要知识来源均为领域专家 (魏勇等, 2016);OSM Semantic Network 在领域专家贡献知识的基础上,允许众包社区用户视需求在模式层中添加知识;SWEET 则进一步增加了地球科学领域词典作为知识来源 (马胜男等, 2010)。从实例层看,GeoSciML 和 SWEET 不包含实例知识;OSM Semantic Network 实例层知识来源于众包社区用户和部分国家政府数据;GeoNames Ontology 实例层结合了众包方法和各国官方结构化地名库的直接转换 (Acheson et al., 2020);GCIS Ontology 实例层知识则仅来源于领域专家 (Ma, 2014)。相比于地学知识图谱仅通

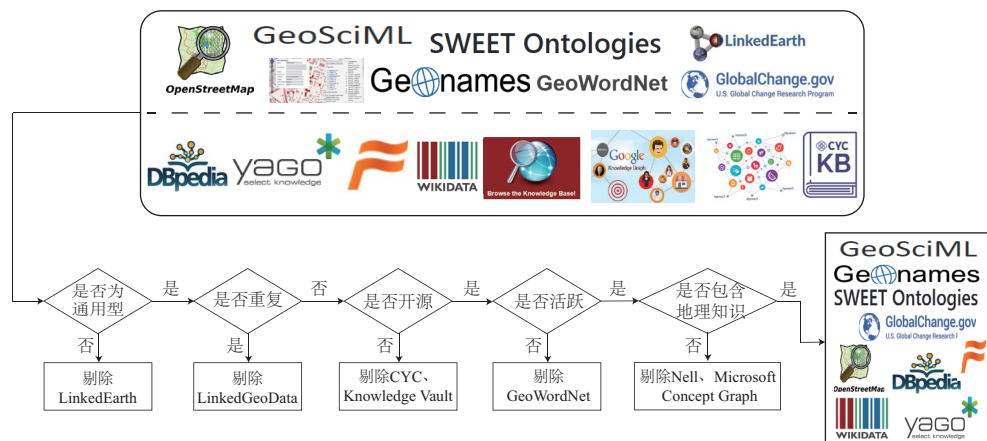


图2 典型地学相关知识图谱确定过程

Fig. 2 The process of determining typical geoscience knowledge graphs

表2 地质相关知识图谱的知识来源
Table 2 Knowledge sources of the related geoscience knowledge graphs

地质相关知识图谱	模式层知识来源	实例层知识来源
GeoSciML	领域专家	—
OSM Semantic Network	领域专家/众包社区用户	众包社区用户/政府数据
GeoNames Ontology	领域专家	众包社区用户/结构化地名库
SWEET	领域专家/众包社区用户/领域词典	—
GCIS Ontology	领域专家	领域专家
DBpedia	领域专家	维基百科
YAGO	已有知识库	维基百科
Freebase	领域专家/众包社区用户	维基百科/普通网页/众包社区用户
Wikidata	领域专家/众包社区用户	维基百科/众包社区用户

过领域专家、众包社区用户和领域词典等获取知识, DBpedia、YAGO、Freebase 和 Wikidata 作为基于开放网络资源的大型知识库, 知识来源更为丰富 (Mendes et al., 2012)。在模式层上, DBpedia 主要由领域专家根据维基百科最常用的信息框中的数据而创建 (Lehmann et al., 2015), YAGO 主要为 WordNet 等已有本体的融合, 而 Freebase 和 Wikidata 则由领域专家和众包社区用户共同构建。在实例层上, DBpedia 和 YAGO 的主要知识源为维基百科; Freebase 既集成了维基百科以及普通网页 (例如, 美国国会图书馆和斯坦福大学图书馆网站的网页) 等网络资源, 还包含了众包社区成员的贡献; Wikidata 则主要来源于维基百科和众包社区用户 (Vrandečić and Krötzsch, 2014)。

3.1.2 构建方式

知识图谱的构建方式主要有自顶向下、自底向上和二者混合三种 (黄恒琪等, 2019)。其中, 自顶向下指先通过专家知识等方式定义知识图谱的模式层, 再在模式层的约束下填充实例层知识; 自底向上指在模式层缺失的情况下直接提取实例层知识, 并根据实例层知识自下而上抽象出模式层 (刘峤等, 2016)。混合法指预先定义模式层, 在模式层的约束下填充实例层, 填充过程中可通过实例层知识的归纳总结迭代更新模式层。

GeoSciML、SWEET、GCIS Ontology、GeoNames、DBpedia 和 YAGO 均采用自顶向下的构建方式, 但具体实现有所不同。前三者通过自顶向下的方式创建模式层, 且由于其目的均为建立相关领域的概念模型, 主要包含概念性知识, 缺乏实例知识或实例知识较少, 因此可理解为只包含模式层的地学知识

图谱 (Raskin and Pan, 2005; Ma, 2014; Emile-Geay et al., 2019)。GeoNames 构建了模式层, 并以众包的方式获取实例知识。DBpedia 和 YAGO 二者比较相近, 均采用自顶向下方式先创建模式层, 在此约束下从网络资源中提取实例层知识。

OSM Semantic Network、Freebase 和 Wikidata 均采用混合构建方式, 即其模式层是可以迭代更新的。OSM Semantic Network 的模式层对于几何数据规定了节点、路和关系的数据结构, 对于属性数据则采用键值对的方式实现 (Girres and Touya, 2010)。在模式层的约束下, 该知识图谱采用众包方法由志愿者协同编辑的方式获取知识。尽管已有模式层, 但用户仍然可以视需求在模式层上添加知识。例如, 在该知识图谱中, 一级类 building 被划分为住宿、商业、宗教等共 78 个子类, 社区用户添加 building 的实例时, 需为其选择二级类, 用户可首先从已有的子类中选择, 但当发现缺乏合适的类型时, 用户可自行创建要素类型, 经过审核后即可作为模式层的新知识。Freebase 和 Wikidata 二者均定义了顶层架构 (例如, Freebase 的“域-类型-主题”的三层结构), 但对于模式层并没有严格的限制, 社区用户可自由编辑。

3.1.3 自动化程度

自动化程度指知识图谱构建过程中的人工参与程度, 可分为人工、半自动和自动 (徐增林等, 2016)。自动化程度与构建方式相关, 采用自底向上构建方式的知识图谱通常均为全自动化; 而采用自顶向下构建方式时则有不同程度的人工参与。自动化程度还与知识图谱的不同构建阶段有关。知识图谱构建可分为模式层构建阶段和实例层构建阶

段。模式层构建阶段的目标是形成知识图谱的概念模型，由于其对于专业知识要求较高（尤其对于地学等领域知识图谱，更是如此），因此通常以学科领域专家手工构建为主，计算机自动构建或复用已有本体为辅。实例层构建阶段的目标是获取概念模型的实例数据，由于其数据量庞大，有限人力无法穷尽，可采用两种方式构建：第一种以自动化知识抽取手段为主，以人工构建等方式为辅；第二种采用众包方法借助于大量志愿者用户获取知识。

地学相关知识图谱两个阶段的构建技术以及其总体的自动化程度如表 3 所示。总体上，通用知识图谱比地学知识图谱具有更高的自动化程度。GeoSciML 仅包含模式层，且以人工方式构建。OSM Semantic Network 和 GeoNames 均以人工方式构建模式层 (Girres and Touya, 2010; Frontini et al., 2013)，以众包方式构建实例层。与依赖于有限人力的完全人工构建不同，众包方式由大量志愿者的共同协作完成；但同时，其本质上仍然依赖于人工输入，与自动知识抽取有显著差别。因此，本文中将其视为与完全的人工构建和自动知识抽取不同的方式。众包方式的优点在于可借助于大量贡献者快速获取大量知识，而且内容更新快，具有更好的现势性。然而，该方式也有以下缺点。首先，它对于众包社区的活跃度具有较高的要求，当缺乏足够的贡献者自愿贡献知识时，将无法实现知识图谱的构建。因此，当通过众包方式构建知识图谱时，需通过积分奖励等机制鼓励社区成员贡献更多知识，以保持社区活跃，实现“共建共享”。其次，众包方式获取数据的质量无法保证。例如，OpenStreetMap 采用众包方式获得的矢量数据的精度可能存在一定

的人为误差。因此，需通过贡献者声誉排名、知识审核、加强知识质量评估等方式提升数据可靠性 (Barron et al., 2014)。SWEET 在初始阶段为完全人工方式构建 (马胜男等, 2010; Emile-Geay et al., 2019)，目前已实现通过群智协同和专家审核的方式进行构建更新，而 GCIS Ontology 则主要依赖于人工构建 (Ma, 2014)。

通用知识图谱中，DBpedia 在模式层阶段通过人工构建 DBpedia Ontology，而在实例层构建阶段则是通过自动化地解析、提取等步骤建立维基百科文本和 DBpedia Ontology 间的映射，从而将其转换为结构化的三元组。同时，DBpedia 通过海量本体的对齐有效提升了知识图谱质量 (Brümmer et al., 2016)。YAGO 通过已有本体的自动融合形成模式层，并在此约束下，从多语言的维基百科文本中自动抽取海量的三元组，因此 YAGO 是完全的自动化构建 (Mahdisoltani et al., 2014)。YAGO 的特点在于通过内容准确性的评测实验评估知识的准确性，以减少由于自动知识抽取带来的潜在错误。Freebase 和 Wikidata 总体上为半自动构建，其模式层构建结合了人工和众包两种方式。二者的特点在于实例层使用了自动抽取和众包两种构建方式，既整合大量网络资源并从中自动提取知识，也接受社区贡献者通过人工的协同编辑贡献知识，极大提升了知识图谱知识来源的丰富性 (Bollacker et al., 2008; Vrandečić and Krötzsch, 2014)。

3.1.4 序列化方式

序列化是指将地学知识转换为可以持久性存储的形式过程。知识图谱的序列化存储格式主要有：RDF/XML、OWL (Web Ontology Language)、

表3 地学相关知识图谱构建的自动化程度

Table 3 Degree of automation for constructing the related geoscience knowledge graphs

地学相关知识图谱	模式层构建技术	实例层构建技术	自动化程度
GeoSciML	人工	—	人工
OSM Semantic Network	人工	众包	人工
GeoNames Ontology	人工	众包	人工
SWEET	人工/众包	—	人工
GCIS Ontology	人工	人工	人工
DBpedia	人工	自动抽取	半自动
YAGO	本体融合	自动抽取	自动
Freebase	人工/众包	自动抽取/众包	半自动
Wikidata	人工/众包	自动抽取/众包	半自动

NT (N-Triples)、N3 (Notation3)、TTL (Turtle)、JSON 等。9 个地学相关知识图谱所提供的序列化方式如表 4 所示。由表可知, RDF/XML 和 NT 是两种最为常用的格式, 前者是 RDF 数据序列化的基于 XML 的一种标准格式, 后者则是一种轻便且易于解析的基于行的格式。开源地学知识图谱通常会提供相应格式的文件下载, 其中部分还会以 SPARQL Endpoint 的形式对外共享。

3.2 内容对比分析

本节通过知识图谱的文件解析统计和文献资料概括总结相结合的方式, 从知识图谱内容的角度, 对地学知识图谱的主要概念、属性及关系进行比较分析; 同时, 由于地学知识具有显著且复杂的时空特征, 这是地学知识区别于其他领域知识的最主要特点, 因此, 本文还对知识图谱中时空特征及其表示进行了详细分析。

3.2.1 主要概念及属性

GeoSciML 包含了地球物质、地质单元和地层学、地质年代、地质构造、地貌学和地球化学等固体地球科学的 1772 个主要术语, 并将其进行二级划分, 其中一级类包含岩性、断层、钻孔等概念, 二级类则是一级类的细化, 例如岩性包含石灰石、大理石等子类。OSM Semantic Network 包含了覆盖全球的地理实体的矢量数据, 尤其是城市路网数据非常完整。它将地理实体划分为地点、边界、建筑、水系等 29 个大类, 地理实体的属性信息以 Tag 标签的形式添加, 所有实体均包含地名、别名、要素类型等基本的属性信息。在此基础上, 每个实体可以任意添加其他属性。GeoNames Ontology 包含了

全球所有国家和地区约 1200 万个地理实体的地名及其属性信息 (魏勇等, 2016), 并将地理实体划分为行政区域、聚居地、道路、水文等 9 大类。与 OpenStreetMap 不同, GeoNames 中实体具有统一的属性, 包含唯一标识、地名、别名、经纬度、高程、实体类型等共 19 项。GeoNames 地名数据比较完整, 既包含实体的地名, 还包含实体的别名以及多语言地名。然而, 它提供的多语言地名存在翻译和音译的不一致问题。

SWEET 包含地球科学领域的人类活动、自然现象、物理过程、物质等顶层概念, 共 12 层分类体系、4533 个概念以及 32 种属性 (马胜男等, 2010; Raskin and Pan, 2003)。SWEET 的特点在于它以“概念空间正交化”(例如, 概念“空气温度”为概念“空气”和“温度”的正交化组成)的方式分解概念, 极大提升了知识表示的可扩展性。与上述以地学实体相关信息为核心的知识图谱不同, GCIS Ontology 以数据或资源的共享为目标, 因此, 它主要包含与数据和资源相关的概念。GCIS Ontology 对“全球变化”领域的相关资源分类十分详细, 包含报告、章节、图表或数据、出版物、软件等 69 个概念以及唯一标识、出版年份、出版状态等 26 种属性。从上述概念属性内容的学科领域上看, 现有地学知识图谱主要聚焦在地理学, 在固体地球科学等领域则很少 (齐浩等, 2020)。在本文初步选定的 8 个地学知识图谱中, 只有 GeoSciML 和 SWEET 包含固体地球科学的相关知识, 但也仅仅涉及顶层本体, 不包含具体的实例知识。

表4 地学相关知识图谱的序列化存储格式
Table 4 Formats for serializing the related geoscience knowledge graphs

地学相关知识图谱	序列化格式					
	RDF/XML	OWL	NT	N3	TTL	JSON
GeoSciML	✓					
OSM Semantic Network	✓					
GeoNames Ontology	✓					
SWEET	✓	✓				
GCIS Ontology		✓				
DBpedia	✓		✓	✓	✓	✓
YAGO			✓			
Freebase			✓			
Wikidata	✓		✓		✓	✓

通用知识图谱除包含地学知识外, 还包含很多其他知识。DBpedia 包含地点、资源、人、物种、工作、组织等 768 个类, 每个类具有不同的属性, 共有超过 3000 个属性 (Vaidya et al., 2015)。DBpedia 实现了自动监测机制能够根据新的网络资源定时更新知识图谱的内容。YAGO 包含国家、名人、城市、电影、艺术作品和组织机构等概念以及标记、别名、注释等基本属性 (Suchanek et al., 2007; Hoffart et al., 2013; Suchanek et al., 2013; Mahdisoltani et al., 2014; Tanon et al., 2020)。由于采用了内容准确性评测实验, 使得 YAGO 的内容具有高精度的优势。Freebase 则包含地点、人、事物等超过 4000 个概念以及超过 7000 个属性 (Bollacker et al., 2008; 李俊, 2011; 周鹏程等, 2016)。Wikidata 包含超过 74000 个概念和超过 9000 个属性 (Pellissier Tanon et al., 2016)。

3.2.2 主要关系

地学知识图谱中概念或实体间的关系, 尤其是时空关系, 是知识计算和推理的基础, 对于支撑下游应用极为重要。然而, 现有地学知识图谱所实现的关系均较为简单。GeoSciML 包含了同义近义等语义关系以及开始于、结束于等时间关系。OSM Semantic Network 主要为矢量数据, 因此包含实体间的空间拓扑关系。GeoNames Ontology 仅构建了地名实体间的包含和邻近关系。SWEET 包含了之前、之后等时间关系, 上、下、左、右等空间关系, 大于、最大等数值关系以及 part of、related to 等共 327 种关系。GCIS Ontology 包含开始于、结束于等时间关系, 位于、工作于等空间关系以及其他语义关系等共 78 种。

相比于地学知识图谱, 通用知识图谱中的语义关系更为丰富, 但时空关系则更为简单。DBpedia、YAGO、Freebase 和 Wikidata 中有大量的 isA (或 instance of)、isPartOf、subClassOf、sameAs 等语义关系; 时空关系上, 则主要有包含、邻近等时空拓扑关系以及出生地位于、死亡地位于、出生日期、死亡日期等时空描述关系。

3.2.3 时空特征及表示

相比于通用知识, 地学知识的特点在于其显著的时空特性, 在形式化的地学知识图谱中主要表现为时空谓词。GeoSciML 包含 “during (在…期间)、

interval starts (开始于)、interval finishes (结束于)”等时间谓词。OSM Semantic Network 和 GeoNames Ontology 是典型的地学知识图谱, 包含大量地学知识的空间谓词, 例如 “gn:parentFeature (包含于)、gn:nearbyFeatures (邻近)、wgs84_pos:long (经度)、wgs84_pos:lat (纬度)”等, 但二者均缺乏关于时间特征的表达。SWEET 包含 “has start time (开始时间)、has end time (结束时间)、has time reference (时间参考)”等时间谓词以及 “has location (位于)、has coordinates (空间坐标)、has spatial extent (空间范围)、has spatial projection (空间投影)”等空间谓词。GCIS Ontology 中包含 “started at (开始于)、ended at (结束于)、at location (位于)、<east bound longitude, west bound longitude, south bound latitude, north bound latitude> (外接矩形经纬度)”等时空谓词。

通用知识图谱通常包含与人物相关的出生地、死亡地、出生日期、死亡日期等时空信息, 但使用不同的谓词表达。例如, DBpedia 中表示为 “dbo:birthPlace、dbo:deathPlace、dbo: birthDate、dbo:deathDate”, YAGO 中表示为 “schema:birthPlace、schema:deathPlace、schema:birthDate、schema:deathDate”, 而 Wikidata 中则表示为 “place of birth”、“place of death”、“date of birth”、“date of death”。DBpedia 还包含 “geo:geometry (几何类型)、geo:long (经度)、geo:lat (纬度)”等描述地理实体空间信息的谓词。综上, 不同知识图谱对于相同的时空特征, 使用不同的时空谓词, 带来了地学知识图谱在模式层上的异构。

4 讨论与启示

在对现有地学知识图谱构建方法和内容比较分析的基础上, 本文发现: 构建方法上, 现有地学知识图谱的知识来源比起通用知识图谱还显得过于单一, 构建的自动化程度也较低; 内容上, 现有地学知识图谱对于时空特征, 尤其是复杂时空计算和推理规则等缺乏考虑, 同时, 时空谓词表达也存在异构。基于上述比较分析结果, 本文从构建方法、内容和应用推广三个方面对于未来地学知识图谱研究提出以下七点启示。

(1) 构建地学知识图谱统一表达模型

不同地学知识图谱的构建均基于其自身的模式层,相互间存在严重的异构,例如地学实体的分类体系、属性的定义与描述、空间关系或者语义关系的谓词表示等均存在不一致。同时,由于实例层是在模式层的约束下构建,因此,模式层的异构也将间接导致实例层的异构。鉴于此,应考虑建立综合地学实体、现象、机理、状态、活动、过程、事件等要素的地学知识统一表达模型(Zheng et al., 2021; 张雪英等, 2020; Wang et al., 2019)。在表达模型的指导下,通过已有地学本体的共享复用和对齐融合以及众包和专家审核相结合的方式构建统一的模式层,为实例层知识的自动抽取奠定统一的顶层约束和规范。

(2) 建立融合多源、多模态数据的知识源

在模式层的约束下,从大规模知识源中自动抽取知识,是地学知识图谱实例层构建的一种重要方法。然而,目前地学知识图谱的知识来源还过于单一,因此,需构建融合网络自由文本、结构化知识库或数据库、矢量数据、栅格影像等多源、多模态数据的知识源,为实例层知识的自动抽取奠定数据基础,以提升地学知识图谱的知识完整性。

(3) 研究地学知识表示与计算方法

通过嵌入式表示将知识图谱中的实体及关系表示在低维实数向量空间中,是知识图谱推理、预测和补全等方法以及语义检索、问答和推荐等应用的基础,也是目前知识图谱领域的研究热点。借助于循环神经网络、图神经网络等深度学习方法,通用领域的知识表示研究已有系列成果(Dai et al., 2020)。然而,目前地学知识的嵌入式表示研究还较少,主要集中在地学实体的地名和空间坐标上(Mai et al., 2020; Santos et al., 2018; Gao and Yan, 2018)。因此,需加强利用人工智能方法实现地学知识的嵌入式表示,为地学知识的计算和推理奠定基础。

(4) 加强地学知识时空特征描述

现有地学知识图谱在模式层上缺乏对于地学知识复杂时空特征的考虑,导致实体时空特征的描述过于简单。因此,知识图谱表达模型的建立以及模式层的构建应从时空尺度、粒度、频度以及类型、格式等多方面全面考虑时空知识的表达。针对时间特征,应充分考虑时间点、时间段、时间基准与参考系、时间不确定度等内容;针对空间特征,应充

分考虑空间位置、空间形态、几何形状、空间基准与参考系、空间不确定性等,以使得地学知识图谱既可以表达简单时空特征(例如,现代时间),也可以支持复杂时空特征的表达(例如,地质时间)。

(5) 考虑地学知识复杂关系及推理规则

现有地学知识图谱,在概念层通常仅具有“父子类”等简单的语义关系,在实例层则仅存在“相同”等简单的语义关系以及“包含”、“邻接”等空间拓扑关系,缺乏对于地学知识中实体间复杂的计算、演化等非线性关系的表达,由此而导致的基于关系的推理规则也同样过于简单,无法支撑地学知识的复杂推理。因此,未来地学知识图谱构建中,应充分考虑地学知识的复杂关系和规则,以增强其在复杂关系上的推理能力。同时,已有地学知识图谱主要集中在地理学,应加强对于包含固体地球科学在内的其他地学学科的知识图谱研究。

(6) 发展地学知识质量评估及修正方法

目前,大规模地学知识图谱的构建主要依赖于众包或者自动知识抽取,但这两种方式均可能带来错误。众包方式可能由于社区贡献者的操作错误或者理解不一致导致知识的错误、异构等问题,尽管通过权限控制和质量审核等手段可一定程度上缓解该问题,但部分错误仍将存在。自动知识抽取通常以网络资源为主要数据源,而网络资源自身就极有可能存在知识的偏差甚至错误。因而,需发展地学知识质量评估及修正方法,从地学知识的准确性、全面性、现势性等方面评估地学知识,检测并修正错误,从而不断提升地学知识图谱的质量。

(7) 提升地学知识图谱应用成效

与通用知识图谱在智能问答、个性化推荐、智能生活助手等实际应用中已取得的应用成效相比,地学知识图谱仍然处于理论探索和实验阶段,缺乏实际应用。GeoNames Ontology 和 OSM Semantic Network 等地学知识图谱仍然仅作为一个地名词典或者空间数据库使用,未能真正发挥知识图谱知识推理和预测的作用,因此应在地学研究中选择典型案例开展实际应用研究,提升地学知识图谱的应用成效。

致谢: 感谢中国地质大学(北京)王成善院士、中国科学院地理科学与资源研究所周成虎院士、中国地质大学(北京)成秋明院士等的指导,感谢深时

数字地球国际大科学计划（DDE）大知识工作组
南京大学胡修棉教授、清华大学闫海荣副教授、上海交通大学王新兵教授、成都理工大学马超教授等的讨论交流，感谢匿名审稿人和编辑对论文提出的修改意见和建议。

参考文献(References):

陈晓慧, 刘俊楠, 徐立, 等. 2020. COVID-19 病例活动知识图谱构建——以郑州市为例[J]. 武汉大学学报 (信息科学版), 45(6): 816–825.

杜志强, 李钰, 张叶廷, 等. 2020. 自然灾害应急知识图谱构建方法研究[J]. 武汉大学学报 (信息科学版), 45(9): 1344–1355.

高松. 2020. 地理空间人工智能的近期研究总结与思考[J]. 武汉大学学报 (信息科学版), 45(12): 1865–1874.

黄恒琪, 于娟, 廖晓, 等. 2019. 知识图谱研究综述[J]. 计算机系统应用, 28(6): 1–12.

蒋秉川, 游雄, 李科, 等. 2020. 利用地理知识图谱的 COVID-19 疫情态势交互式可视分析[J]. 武汉大学学报 (信息科学版), 45(6): 836–845.

李俊. 2011. 语义数据库 Freebase 研究[J]. 现代图书情报技术, 27(10): 18–23.

刘峤, 李杨, 段宏, 等. 2016. 知识图谱构建技术综述[J]. 计算机研究与发展, 53(3): 582–600.

刘俊楠, 刘海砚, 陈晓慧, 等. 2020a. 面向多源地理空间数据的知识图谱构建[J]. 地球信息科学学报, 22(7): 1476–1486.

刘俊楠, 刘海砚, 陈晓慧, 等. 2020b. 基于地理空间数据的知识图谱构建技术研究[J]. 中文信息学报, 34(11): 29–36.

陆锋, 余丽, 仇培元. 2017. 论地理知识图谱[J]. 地球信息科学学报, 19(6): 723–734.

马胜男, 孙翊, 郭明明. 2010. SWEET 本体研究述评[J]. 标准科学, (9): 38–43.

裴韬, 刘亚溪, 郭思慧, 等. 2019. 地理大数据挖掘的本质[J]. 地理学报, 74(3): 586–598.

齐浩, 董少春, 张丽丽, 等. 2020. 地球科学知识图谱的构建与展望[J]. 高校地质学报, 26(1): 2–10.

孙鸿烈. 2017. 地学大辞典[M]. 北京: 科学出版社.

陶坤旺, 赵阳阳, 朱鹏, 等. 2020. 面向一体化综合减灾的知识图谱构建方法[J]. 武汉大学学报 (信息科学版), 45(8): 1296–1302.

王志华, 杨晓梅, 周成虎. 2021. 面向遥感大数据的地学知识图谱构想[J]. 地球信息科学学报, 23(1): 16–28.

魏勇, 胡丹露, 李响, 等. 2016. 基于 GeoNames 和 Solr 的地名数据全文检索[J]. 测绘工程, 25(2): 28–32.

徐增林, 盛泳潘, 贺丽荣, 等. 2016. 知识图谱技术综述[J]. 电子科技大学学报, 45(4): 589–606.

余丽, 陆锋, 刘希亮. 2016a. 开放式地理实体关系抽取的 Bootstrapping 方法[J]. 测绘学报, 45(5): 616–622.

余丽, 陆锋, 刘希亮, 等. 2016b. 稀疏地理实体关系的关键词提取方法[J]. 地球信息科学学报, 18(11): 1465–1475.

张雪英, 张春菊, 吴明光, 等. 2020. 顾及时空特征的地理知识图谱构建方法[J]. 中国科学: 信息科学, 50(7): 1019–1032.

赵红伟, 诸云强, 侯志伟, 等. 2016. 地理空间元数据关联网络的构建[J]. 地理科学, 36(8): 1180–1189.

周成虎, 王华, 王成善, 等. 2021. 大数据时代的地学知识图谱研

究[J]. 中国科学: 地球科学, 51(7): 1070–1079.

周鹏程, 武川, 陆伟. 2016. 基于多知识库的短文本实体链接方法研究——以 Wikipedia 和 Freebase 为例[J]. 现代图书情报技术, 6(1): 105–112.

Acheson E, Volpi M and Purves R S. 2020. Machine learning for cross-gazetteer matching of natural features [J]. International Journal of Geographical Information Science, 34(4): 708–734.

Auer S, Lehmann J and Hellmann S. 2020. Linkedgeodata: Adding a spatial dimension to the web of data [C] // International Semantic Web Conference, Springer: 731–746.

Barron C, Neis P and Zipf A. 2014. A comprehensive framework for intrinsic OpenStreetMap quality analysis [J]. Transactions in GIS, 18(6): 877–895.

Bollacker K, Evans C, Paritosh P, et al. 2008. Freebase: a collaboratively created graph database for structuring human knowledge [C] // Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data: 1247–1250.

Brümmer M, Dojchinovski M and Hellmann S. 2016. Dbpedia abstracts: A large-scale, open, multilingual NLP training corpus [C] // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16): 3339–3343.

Dai Y, Wang S, Xiong N N, et al. 2020. A survey on knowledge graph embedding: Approaches, applications and benchmarks [J]. Electronics, 9(5): 750–778.

Degrossi L C, Porto De Albuquerque J, Santos Rocha R D, et al. 2018. A taxonomy of quality assessment methods for volunteered and crowdsourced geographic information [J]. Transactions in GIS, 22(2): 542–560.

Dong X, Gabrilovich E, Heitz G, et al. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion [C] // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 601–610.

Emile-Geay J, Khider D, Garijo D, et al. 2019. The linked earth ontology: A modular, extensible representation of open paleoclimate data [DS]/ Zenodo.

Frontini F, Del Gratta R and Monachini M. 2013. GeoDomainWordNet: Linking the geonames ontology to WordNet [C] // Proceedings of the 2013 Language and Technology Conference. Springer: 229–242.

Gao S and Yan B. 2018. Place2vec: Visualizing and reasoning about place type similarity and relatedness by learning context embeddings [C] // Adjunct Proceedings of the 14th International Conference on Location Based Services. ETH Zurich: 225–226.

Girres J F and Touya G. 2010. Quality assessment of the French OpenStreetMap dataset [J]. Transactions in GIS, 14(4): 435–459.

Giunchiglia F, Dutta B and Maltese V. 2009. Faceted lightweight ontologies [C] // Conceptual Modeling: Foundations and Applications, Springer: 36–51.

Giunchiglia F, Maltese V, Farazi F, et al. 2010. GeoWordNet: a resource for geo-spatial applications [C] // Extended Semantic Web Conference, Springer: 121–136.

Hey A J, Tansley S and Tolle K M. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery [M] // Microsoft Research Redmond, WA.

Hoffart J, Suchanek F M, Berberich K, et al. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial Intelligence, 194: 28–61.

- Kitchenham B and Charters S. 2007. Guidelines for performing systematic literature reviews in software engineering [R]. Technical Report EBSE, School of Computer Science and Mathematics, Keele University: Keele, UK.
- Lehmann J, Isele R, Jakob M, et al. 2015. Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia [J]. *Semantic Web*, 6(2): 167–195.
- Li L, Liu Y, Zhu H, et al. 2017. A bibliometric and visual analysis of global geo-ontology research [J]. *Computers & Geosciences*, 99: 1–8.
- Ma X M. 2014. Experience in ontology engineering with the Global Change Information System [C] // *Proceedings of ESIP Semantic Web Cluster Webinar*.
- Mahdisoltani F, Biega J and Suchanek F. 2014. Yago3: A knowledge base from multilingual wikipeas [C] // 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference.
- Mai G, Janowicz K, Cai L, et al. 2020. SE-KGE: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting [J]. *Transactions in GIS*, 24(3): 623–655.
- Matuszek C, Witbrock M, Cabral J, et al. 2006. An introduction to the syntax and content of Cyc [C] // *UMBC Computer Science and Electrical Engineering Department Collection*.
- Mendes P N, Jakob M and Bizer C. 2012. DBpedia: A multilingual cross-domain knowledge base [C] // *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association (ELRA): 1813–1817.
- Mitchell T, Cohen W, Hruschka E, et al. 2018. Never-ending learning [J]. *Communications of the ACM*, 61(5): 103–115.
- Pellissier T T, Vrandečić D, Schaffert S, et al. 2016. From freebase to wikidata: The great migration [C] // *Proceedings of the 25th International Conference on World Wide Web*: 1419–1428.
- Qiu P, Gao J, Yu L, et al. 2019a. Knowledge embedding with geospatial distance restriction for geographic knowledge graph completion [J]. *ISPRS International Journal of Geo-Information*, 8(6): 254–276.
- Qiu Q, Xie Z, Wu L, et al. 2019b. BiLSTM-CRF for geological named entity recognition from the geoscience literature [J]. *Earth Science Informatics*, 12(4): 565–579.
- Raskin R and Pan M. 2003. Semantic web for earth and environmental terminology (sweet) [C] // *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*.
- Raskin R and Pan M. 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET) [J]. *Computers & Geosciences*, 31(9): 1119–1125.
- Santos R, Murrieta-Flores P, Calado P, et al. 2018. Toponym matching through deep neural networks [J]. *International Journal of Geographical Information Science*, 32(2): 324–348.
- Senaratne H, Mobasheri A, Ali A L, et al. 2017. A review of volunteered geographic information quality assessment methods [J]. *International Journal of Geographical Information Science*, 31(1): 139–167.
- Shao B, Li X and Bian G. 2021. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph [J]. *Expert Systems with Applications*, 165, 113764.
- Shin S, Jin X, Jung J, et al. 2019. Predicate constraints based question answering over knowledge graph [J]. *Information Processing & Management*, 56(3): 445–462.
- Stadler C, Lehmann J, Höffner K, et al. 2012. Linkedgeodata: A core for a web of spatial open data [J]. *Semantic Web*, 3(4): 333–354.
- Suchanek F M, Hoffart J, Kuzey E, et al. 2013. Yago2s: Modular high-quality information extraction with an application to flight planning [C] // *Datenbanksysteme für Business, Technologie und Web (BTW)* 2048: 515–518.
- Suchanek F M, Kasneci G and Weikum G. 2007. Yago: a core of semantic knowledge [C] // *Proceedings of the 16th International Conference on World Wide Web*: 697–706.
- Sun K, Hu Y, Song J, et al. 2020. Aligning geographic entities from historical maps for building knowledge graphs [J]. *International Journal of Geographical Information Science*, 35(10): 2078–2107.
- Sun K, Zhu Y and Song J. 2019. Progress and challenges on entity alignment of geographic knowledge bases [J]. *ISPRS International Journal of Geo-Information*, 8(2): 77–101.
- Tanon T P, Weikum G and Suchanek F. 2020. Yago 4: A reasonable knowledge base [C] // *European Semantic Web Conference*. Springer: 583–596.
- Vaidya G, Kontokostas D, Knuth M, et al. 2015. DBpedia commons: structured multimedia metadata from the wikimedia commons [C] // *International Semantic Web Conference*, Springe: 281–289.
- Vrandečić D and Krötzsch M. 2014. Wikidata: a free collaborative knowledgebase [J]. *Communications of the ACM*, 57(10): 78–85.
- Wang C, Ma X, Chen J, et al. 2018. Information extraction and knowledge graph construction from geoscience literature [J]. *Computers & Geosciences*, 112: 112–120.
- Wang J, Hu Y and Joseph K. 2020. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages [J]. *Transactions in GIS*, 24(3): 719–735.
- Wang S, Zhang X, Ye P, et al. 2019. Geographic knowledge graph (GeoKG): A formalized geographic knowledge representation [J]. *ISPRS International Journal of Geo-Information*, 8(4): 184–207.
- Zhang N, Deng S, Chen H, et al. 2018. Structured knowledge base as prior knowledge to improve urban data analysis [J]. *ISPRS International Journal of Geo-Information*, 7(7): 264–279.
- Zheng K, Xie M H, Zhang J B, et al. 2021. A knowledge representation model based on the geographic spatiotemporal process [J]. *International Journal of Geographical Information Science*: 36(4): 674–691.
- Zhu Y, Zhu A X, Feng M, et al. 2017. A similarity-based automatic data recommendation approach for geographic models [J]. *International Journal of Geographical Information Science*, 31(7): 1403–1424.