

基于本体模型的锂矿知识图谱的构建研究

吴 卓

成都理工大学数学科学学院, 四川 成都

收稿日期: 2025年2月19日; 录用日期: 2025年3月11日; 发布日期: 2025年3月21日

摘 要

地球科学属于典型的数据密集型科学, 从非结构化数据中挖掘有效信息已逐渐发展为地球科学的热门研究方向之一。知识图谱作为一种非结构化信息的有效处理方案, 为地球科学领域的知识挖掘提供全新的研究思路与技术手段, 推动了地球科学知识的整合与共享。本体模型的提出进一步推动了知识图谱构建的规范性。但是, 目前尚且缺乏对锂矿本体模型的研究。本文从具体矿床名称出发, 构建了基于本体模型的锂矿知识图谱, 最终探讨了知识图谱与自然语言处理技术的协同工作方式与大型语言模型在地球科学中的应用前景。本文指出未来的研究工作应(1) 构建基于本体模型的地球科学知识图谱, (2) 考虑推进基于本体模型知识图谱的应用, 实现知识推理, (3) 将大型语言模型与知识图谱相结合, 以期为进一步推动地球科学知识图谱的发展提供参考。

关键词

知识图谱, 本体, 自然语言处理, 协同工作

Research on the Construction of Knowledge Graph of Lithium Ore Based on Ontology Model

Zhuo Wu

School of Mathematical Sciences, Chengdu University of Technology, Chengdu Sichuan

Received: Feb. 19th, 2025; accepted: Mar. 11th, 2025; published: Mar. 21st, 2025

Abstract

Earth science is a typical data-intensive science, mining effective information from unstructured data has gradually developed into one of the hot research directions of earth science. As an effective method of processing unstructured information, knowledge graph provides a new research idea

and technical means for knowledge mining in the field of earth science, and promotes the integration and sharing of earth science knowledge. The proposal of ontology model further promotes the normalization of knowledge graph construction. However, at present, there is a lack of research on lithium ore ontology model. Based on the name of specific ore deposit, this paper constructs the knowledge graph of lithium ore based on ontology model, and finally discusses the cooperative working mode of knowledge graph and natural language processing technology and the application prospect of large-scale language model on earth science. This paper points out that future research should (1) construct ontology-based knowledge graphs of Earth sciences, (2) consider promoting the application of ontology-based knowledge graphs to realize knowledge reasoning, and (3) combine large-scale language models with knowledge graphs, in order to provide references for further promoting the development of knowledge graphs of earth science.

Keywords

Knowledge Graph, Ontology, Natural Language Processing, Collaborative Work

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,大数据和人工智能技术的发展为地球科学数据处理带来了无限可能,机器学习(Machine Learning, ML)、深度学习(Deep Learning, DL)的方法在地球科学中的应用越来越广泛[1][2]。海量的报告、专著、论文等非结构化文本使得地球科学从传统的经验性科学成为典型的数据密集型科学,尽管海量非结构化文本向大数据、人工智能技术提供了必要数据,但难以直接套用。地球科学中数据密集、数据混杂、数据之间难以集成与共享等问题亟待进一步深度挖掘[3]。对非结构化数据进行逐字逐句分析是一项浩大繁琐的过程,获取特定的信息和知识需要耗费大量的时间精力[4],造成地球科学领域面临“海量信息但知识饥饿”的窘况[5]。如何实现不同领域之间知识的组织融合从而实现所需信息的快速检索是当前地球科学领域面临的核心问题之一。

知识图谱概念最早由谷歌公司于2012年提出[6],旨在以图形表示法提升用户的搜索体验[7]。当前地球科学的知识体系有必要从传统的百科全书式体系过渡到可理解的知识图谱体系,以图的形式表示知识,将实体、属性和关系用节点和边连接起来,整合地球科学领域的分散知识和信息,以人类能够理解的图形式表征非结构化信息。

知识图谱采用<实体1,关系,实体2>或<实体,属性,属性值>的三元组结构对数据进行存储,以圆内代表实体,以箭头代表关系,属性与属性值在实体内部。知识图谱因此能够将不同数据源中的相关实体、关系以及属性进行连接,以人类能够理解的图形式表征非结构化信息,具备扩展性强的典型特点。

知识图谱凭借结构化表示能够帮助研究人员从宏观上理解和分析地球科学数据中的关联和关系,同时对非结构化数据具备有效的推理能力,可以发掘地球科学数据中的潜在规律。通过多种知识图谱的构建方式并结合自然语言处理(Natural Language Processing, NLP)、DL等技术手段,可以有效应用于地球科学非结构化文本分析中,挖掘所需三元组信息,消除不同数据源之间的隔阂,确保不同来源和类型的知识能在知识图谱中融合,并实现知识之间的联接,为实现基于地球科学的智能化分析奠定基础。本文选取锂矿作为研究背景,构建锂矿知识图谱,尝试提出当前地球科学知识图谱应用的前沿方向。

2. 锂矿知识图谱的构建

知识图谱主要有两种构建方法：自上而下和自下而上的方法[8]。前者是通过定义规则、框架等方式生成知识三元组，需要有标准化、结构化、规范化的知识作为基础，需要设定框架性的模型约束。后者则是通过数据驱动的方式，从大规模的数据中提取信息，从而获得广泛的、多样性的知识。但目前自上而下和自下而上相结合构建地球科学知识图谱两种构建方式进行结合是目前较为常用和有效的策略[9][10]，因此现有研究多是将地球科学知识图谱从逻辑架构上分为两层：本体层(也称模式层)以及实体层(也称数据层或事实层)。本体源于哲学概念，但在信息科学中的本体定义得到了广泛认可，即本体是对于概念模型的一种标准明确的规范[11]。在自上而下和自下而上相结合构建领域知识图谱的过程中，首先通过自上而下的方式构建领域本体。领域本体作为领域知识图谱的核心和基础，旨在对该领域的共识知识以一种正式、系统的方式表达，对构建高质量的领域知识图谱起着至关重要的作用[12]。

现有研究主要以五元组的方式实现本体模型的表达[13]，如下式所示。

$$Onto = (Con, Rel, Prop, Rule, Ins)$$

其中，*Onto* 代表本体模型的总称；*Con* 指具备共同特性的事物集合总称；*Rel* 指概念之间的关系总称；*Prop* 指属性总称；*Rule* 代表对领域概念、实例取值等约束表达；*Ins* 代表基于具体领域概念下的具体实例。

2.1. 本体层构建

2.1.1. 本体建模流程

本研究采用七步法构建锂矿本体模型[14]，从具体矿床名称出发，根据《矿床学》[15]、《基础矿床学》[16]等官方教材选择特定的概念与关系，以覆盖地质领域中锂矿领域的特定知识。以数学表达式的形式对梳理锂矿本体模型的逻辑架构。

首先将锂矿知识图谱划分为矿床地质背景和区域地质背景两方面。

$$Lname = (A_1, A_2)$$

其中，*Lname* 代表具体锂矿床名称，*A₁* 代表区域地质背景，*A₂* 代表矿床地质背景。

在区域地质背景中，地层、岩浆岩、构造活动等情况是矿床学研究的基础[15]，大地构造为成矿是成矿物质沉淀富集的场所。将其添加至区域地质背景中。

$$A_1 = (B_1, B_2, B_3, B_4)$$

其中，*A₁* 代表区域地质背景，*B₁* 代表区域出露地层，*B₂* 代表构造，*B₃* 代表岩浆岩，*B₄* 代表大地构造。

不同的出露地层具有相应的地层岩性，以下式表示。

$$B_1 = (C_1)$$

其中，代表区域出露地层，*C₁* 代表区域地层岩性。

矿床地质背景较为复杂，需要形成一定的层次体系结构[17]。将矿床地质背景进行了细分，以下式进行表示。

$$B_2 = (E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9, E_{10}, E_{11}, E_{12}, E_{13})$$

其中，*B₂* 代表矿床地质背景，*E₁* 代表矿石，*E₂* 代表矿区出露地层，*E₃* 代表成矿作用，*E₄* 代表矿体，*E₅* 代表围岩蚀变类型，*E₆* 代表找矿标志，*E₇* 代表成因类型，*E₈* 代表矿床资源储量，*E₉* 代表矿床规模，*E₁₀* 代表矿床类型，*E₁₁* 代表矿区地层，*E₁₂* 代表地理位置，*E₁₃* 代表围岩。

本研究按照《矿床学》对矿石概念进行了划分,采用下式定义矿石概念集。

$$E_1 = (F_1, F_2, F_3, F_4, F_5, F_6)$$

其中, E_1 代表矿石, F_1 代表矿石品位, F_2 代表矿石结构, F_3 代表矿石矿物, F_4 代表脉石矿物, F_5 代表矿石构造, 由于构建的是锂矿, 因此选取 F_6 代表含锂矿物。

矿区出露地层概念包含区域出露地层。

$$E_2 = (G_1)$$

其中, E_i 代表矿区出露地层, G_i 代表矿区地层岩性。

本研究对成矿作用进行了划分,采用下式定义成矿作用概念集。

$$E_3 = (H_1, H_2, H_3, H_4, H_5)$$

其中, E_3 代表成矿作用, H_1 代表成矿区, H_2 代表成矿元素, H_3 代表伴生矿种, H_4 代表成矿带, H_5 成矿时期。

矿体可分为矿体产状与矿体形态，在矿体产状概念下继续细分，将其分为倾角、倾向、走向三个概念。

$$E_4 = (J_1, J_2)$$

$$J_1 = (K_1, K_2, K_3)$$

其中, E_4 代表矿体, J_1 代表矿体产状, J_2 代表矿体形态, K_1 代表倾角, K_2 代表倾向, K_3 代表走向。

2.1.2. 本体可视化

本体可视化是在上述语义关系表达式的基础上添加部分实例,完成本体建模。本研究采用的可视化软件是 Protégé (5.5.0 版本),在支持中文格式的输入的基础之上可以利用插件 Graphviz 显示本体模型的层级结构[18]。依次添加概念、对象属性、数据属性以及实例,共构建了 39 个类、19 条对象属性、19 条数据属性,36 条类公理。最终构建的锂矿本体模型如图 1 所示。从图中可以看出,本研究以具体矿床名称出发,逐层构建,最终形成结构化的知识体系,保证了模型的合理性和完备性。

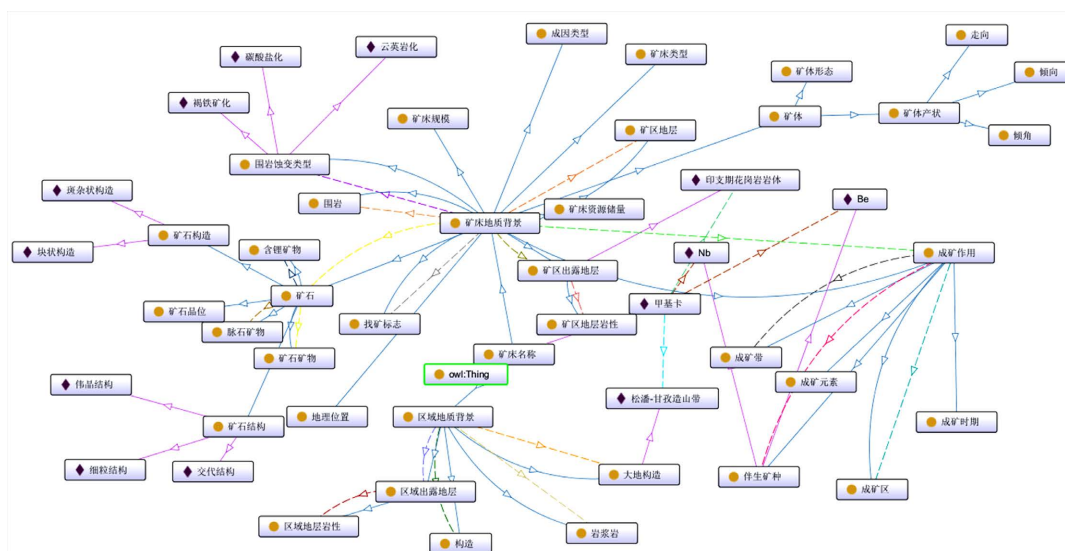


Figure 1. Lithium mining ontology model

图 1. 锂矿本体模型

2.2. 数据层构建

本体模型构建完毕后需要有抽象概念下对应的具体实体,本研究采用标注平台 Doccano,这是一个具有情感分析、文本分类等多种标注功能的文本标注平台。本研究采用开源数据集,在“中国知网”上将“伟晶岩型锂矿”、“川西伟晶岩型锂矿”进行主题搜索,检索到期刊与学位论文共计 186 篇,形成本研究所需数据集。剔除参考文献、地质图等与知识图谱构建无关的内容,将每篇文档整理成 txt 文本形式,上传到系统中,用于数据标注。标注过程中所有命名规范按照[15][16][19]等标准进行。标注过程如下:首先点击数据集,选中需要标注的文本后点击“标注”,进入标注页面。当遇到符合标注条件的实体时,选中该词,doccano 平台会自动弹出标签列表,从中选择对应的实体标签即可。以实体标注为例,若文本中出现“川西甲基卡花岗伟晶岩型稀有金属矿床”,可将“甲基卡”其标注为“矿床名称”;出现“青藏高原东部”,则标注为“地理位置”。进行关系标注时,需在页面右侧通过“Span”按钮将实体标签标注转为关系标注。然后依次点击两个实体标签,点击第二个实体标签后,平台会弹出关列表,从中选择对应的关系。比如,先点击甲基卡所在的矿床名称标签,再点击青藏高原东部的地理位置标签,从弹出的关列表中选择“位于”,就完成了实体与关系的创建和标注。一篇文章的标注完成后通过另一人的检验以及相关学者的通过后,最终将导出的数据转换成三元组的形式,为保证数据的一致性,需要进行实体消歧工作,如“九龙县打枪沟锂铍稀有金属矿床”与“川西九龙打枪沟锂铍矿”均指代打枪沟这一矿床,为此将表述不一致但实际内容一致的实体进行统一(此处命名为打枪沟),导入至图数据库 Neo4j 中。效果图如图 2 所示。

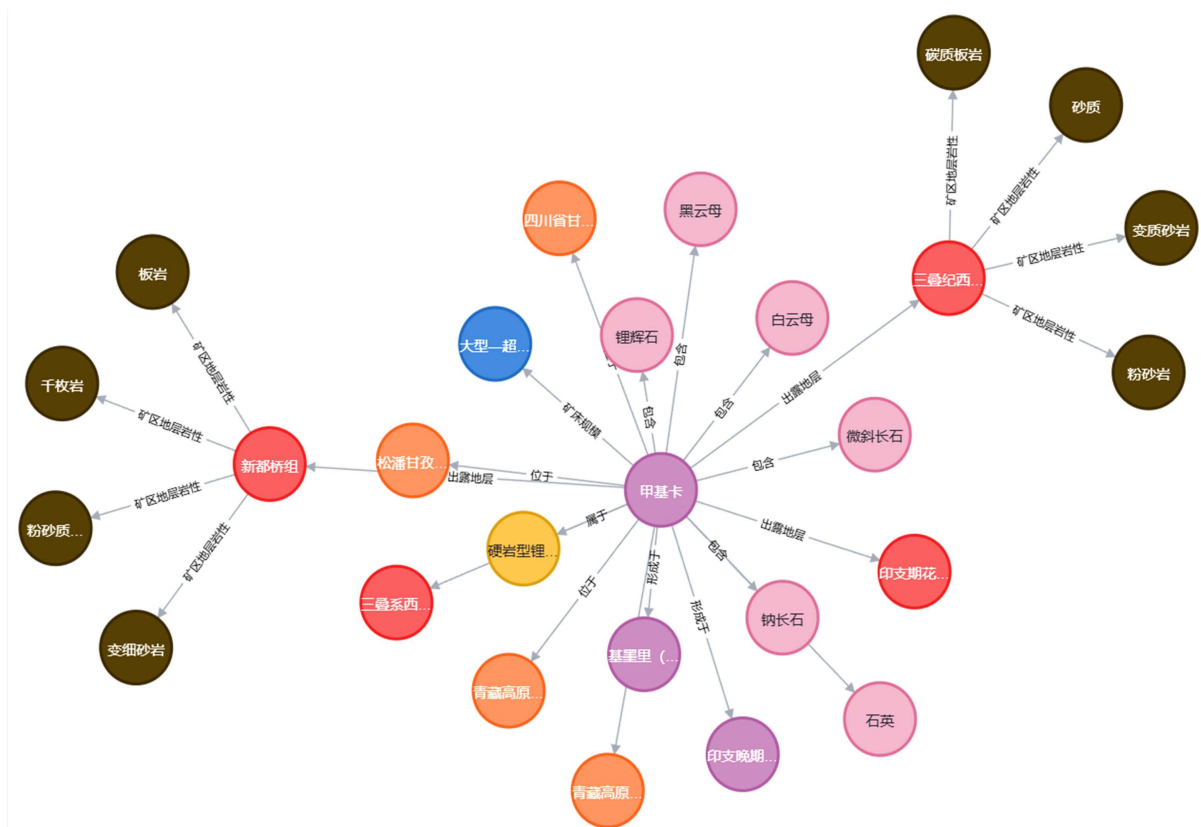


Figure 2. Final result diagram (part)

图 2. 最终结果图(部分)

由于自上而下构建知识图谱的构建过程局限于自我认知,所以需要 NLP 协助抽取非结构化数据中的实体以及关系,将非结构化数据转化为结构化的数据表示。如果只使用 NLP 会使知识图谱缺乏知识的规范性,所以需要实现知识图谱与 NLP 的协同工作,提高知识图谱构建的效率与质量,也为知识图谱的应用奠定基础。也正是知识图谱与 NLP 的协同应用得以实现更加高效、智能的地球科学信息处理和分析,为推动地球科学智能系统发展开辟全新的前景。将知识图谱与 NLP 技术相结合。知识图谱作为一种结构化的知识表示形式,可以帮助系统更好地理解和推理地球科学知识。NLP 则赋予系统处理文本的能力,使得系统能够从文本中提取有效特定的信息[20]。

在协同框架下,知识图谱能够为 NLP 提供有价值的地球科学领域专业知识,丰富的语义关联信息,加强对地球科学领域特定术语、概念以及实体的理解和关联,对文本中的含义和逻辑关系进行推理,以便于更好地理解和分析地球科学语义[21]。同时,NLP 技术能够为知识图谱注入更多的实体链接和关系抽取能力。二者紧密的融合使得智能系统在语义理解、信息检索、问答系统等领域展现更为强大的能力,为用户提供更加智能化、个性化的服务和体验,促进地球科学与计算机、大数据、人工智能的交叉研究和创新。

NLP 在自动处理和理解大量文本数据,帮助构建知识图谱的同时,也基于最新的研究成果将有价值的信息整合至知识图谱中,使其更加全面、准确和实用,发现本体模型中缺少的内容并进行补全,在提高知识图谱的构建效率和质量的同时,还拓展了其应用场景和功能。Qiu 等人[22]提出了将 NLP 和本体匹配信息提取进行组合的方法,通过构建本体、空间信息提取(创建空间地名词典、时态词典、地名词典)、语义信息提取(命名实体识别、本体匹配、时空关系提取)用于自动识别地质灾害报告中的语义和时空信息,最终构建出本体模型指导下的知识图谱。知识图谱的动态更新机制保持了知识图谱的实时性与实用性,为地球科学进一步研究和决策提供有力支持。

在知识图谱的应用中,知识图谱为 NLP 提供了丰富的语义信息和背景知识,能够增强文本理解和推理能力。NLP 能够支持地球科学知识图谱实现更智能化的搜索功能。以问答系统为例,目前需要构建大量的数据集以提高 NLP 模型的准确性,实现对用户问题的精准处理,利用 NLP 将文本中的实体、关系与知识图谱中的实体、关系进行链接,依照知识图谱中的结构化信息展示准确的答案,还可以通过知识图谱本体中的名称和定义构建自动问答数据集,提高工作效率,也为其他领域的问答系统构建提供一种思路[4]。知识图谱中的关联关系和 NLP 能够共同提高智能系统在文本理解、信息抽取和知识推理等任务中的表现,为地球科学领域提供更强大的语言理解能力,实现从“文档检索”到“知识检索”的范式转变[23]。图 3 展示了知识图谱与 NLP 协同构建的工作流程。

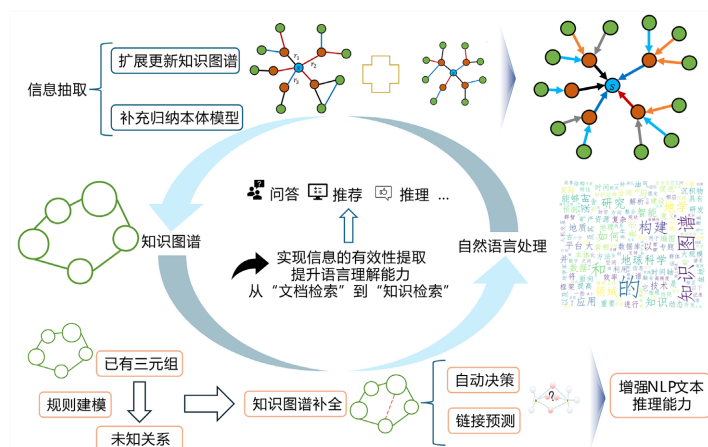


Figure 3. Workflow of collaborative construction of knowledge graph and natural language processing
图 3. 知识图谱与自然语言处理协同构建的工作流程

Table 1. Overview of collaborative enhancement of knowledge graph and natural language processing in geoscience
表 1. 在地球科学中知识图谱与自然语言处理协同增强概览

	自然语言处理	知识图谱
增强原因	整合多源异构地学数据 实现基于知识图谱的智能推理	
增强方式	提供事实层事实补充 发现隐藏的实体关系 扩展更新知识图谱	提供本体层指导约束 增强文本推理能力 修改本体模型
增强结果	处理复杂地学知识 保持知识图谱的时效性 从“文档检索”到“知识检索”	
后续展望	细化完善本体模型 实现知识图谱的补全更新	

地球科学知识图谱与 NLP 相互协同实现了信息的有效性提取和系统的功能提升。地球科学知识图谱通过结构化的语义信息增强了 NLP 的准确性，NLP 技术通过构建、更新知识图谱以丰富完善地球科学知识图谱。知识图谱与 NLP 在地球科学领域的协同应用有望促进地球科学领域的具体实践，为地球科学领域的发展带来更多可能性和机遇。表 1 总结了在地球科学工作中知识图谱与 NLP 各自的工作方向。未来应考虑构建基于地学知识图谱的 NLP 系统，用于生成有关地球科学事件的摘要或报告。系统利用知识图谱中的知识框架与 NLP 处理文本的能力，确保生成的内容既准确又全面。

3. 结论

近年来的研究证明，面对典型数据密集型的地球科学，地球科学知识图谱的构建是将多源异构的非结构化数据进行整合的有效手段，以数据可视化的形式帮助研究人员理解地球科学的复杂性。本文探讨了构建知识图谱时的 NLP 与知识图谱进行协同工作。本文的结论如下：

- a. 当前地学知识图谱的研究仍处于初期阶段，未来地球科学知识图谱的构建依旧需将自上而下与自下而上相结合的方式相结合，发挥本体模型的优势，进而形成规范的、标准的地球科学知识图谱。为此，需要发挥深时数字地球(Deep-time Digital Earth, DDE)大型项目带来的数据集成与共享的优势，形成一系列被广泛认可的本体模型。
- b. 未来在地球科学知识图谱的应用中，依旧需要围绕本体模型实现基于知识图谱的知识推理，将地球科学知识图谱更好地引入到资源勘查、灾害预警当中去，为研究人员提供准确的决策评估。还需要建立有效的机制来跟踪最新的研究成果和数据变化，及时将新知识纳入到知识图谱中，保持知识图谱的时效性。
- c. LLMs 的问世为地球科学的发展指明了前进的方向，目前已有地球科学的大型语言模型，但未来需继续将 LLMs 引入到三元组抽取、知识推理等过程中去，将 LLMs 与知识图谱、NLP 相结合，更好地构建知识图谱、理解地球科学中的问题，促进知识图谱与地球科学的发展。

基金项目

国家重点研发计划课题(编号：2023YFC2906403)。

参考文献

[1] Bergen, K.J., Johnson, P.A., de Hoop, M.V. and Beroza, G.C. (2019) Machine Learning for Data-Driven Discovery in

- Solid Earth Geoscience. *Science*, **363**, eaau0323. <https://doi.org/10.1126/science.aau0323>
- [2] Ma, X. (2022) Knowledge Graph Construction and Application in Geosciences: A Review. *Computers & Geosciences*, **161**, Article ID: 105082. <https://doi.org/10.1016/j.cageo.2022.105082>
 - [3] 牛凤桂, 张贝, 陈石. 大数据时代的地球科学知识图谱研究现状与展望[J]. 地震学报, 2024, 46(3): 353-376.
 - [4] Qiu, Q., Tian, M., Ma, K., Tan, Y.J., Tao, L. and Xie, Z. (2023) A Question Answering System Based on Mineral Exploration Ontology Generation: A Deep Learning Methodology. *Ore Geology Reviews*, **153**, Article ID: 105294. <https://doi.org/10.1016/j.oregeorev.2023.105294>
 - [5] Qiu, Q., Wang, B., Ma, K., Lü, H., Tao, L. and Xie, Z. (2023) A Practical Approach to Constructing a Geological Knowledge Graph: A Case Study of Mineral Exploration Data. *Journal of Earth Science*, **34**, 1374-1389. <https://doi.org/10.1007/s12583-023-1809-3>
 - [6] Singhal, A. (2012) Introducing the Knowledge Graph: Things, Not Strings. <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>
 - [7] Fensel, D., Şimşek, U., Angele, K., *et al.* (2020) Knowledge Graphs: Methodology, Tools and Selected Use Cases. Springer. <http://link.springer.com/10.1007/978-3-030-37439-6>
 - [8] Hao, X., Ji, Z., Li, X., Yin, L., Liu, L., Sun, M., *et al.* (2021) Construction and Application of a Knowledge Graph. *Remote Sensing*, **13**, Article 2511. <https://doi.org/10.3390/rs13132511>
 - [9] 王益鹏, 张雪英, 党玉龙, 等. 顾及时空过程的台风灾害事件知识图谱表示方法[J]. 地球信息科学学报, 2023, 25(6): 1228-1239.
 - [10] 冉一早, 董少春, 王汝成, 等. 铌钽矿床知识图谱的构建及应用实践[J]. 高校地质学报, 2023, 29(3): 359-371.
 - [11] Gruber, T.R. (1993) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**, 199-220. <https://doi.org/10.1006/knac.1993.1008>
 - [12] AL-Aswadi, F.N., Chan, H.Y. and Gan, K.H. (2022) From Ontology to Knowledge Graph Trend: Ontology as Foundation Layer for Knowledge Graph. In: Villazón-Terrazas, B., Ortiz-Rodriguez, F., Tiwari, S., Sicilia, M.A. and Martín-Moncunill, D., Eds., *Knowledge Graphs and Semantic Web*, Springer, 330-340. https://doi.org/10.1007/978-3-031-21422-6_25
 - [13] 邱芹军, 吴亮, 马凯, 等. 面向灾害应急响应的地质灾害链知识图谱构建方法[J]. 地球科学, 2023, 48(5): 1875-1891.
 - [14] Noy, N.F., Ferguson, R.W. and Musen, M.A. (2000) The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility. In: Dieng, R. and Corby, O., Eds., *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, Springer, 17-32. https://doi.org/10.1007/3-540-39967-4_2
 - [15] 翟裕生, 姚书振, 蔡克勤. 矿床学[M]. 北京: 地质出版社, 2011. <https://book.douban.com/subject/20315635/>
 - [16] 薛春纪. 基础矿床学[M]. 北京: 地质出版社, 2006. <https://product.dangdang.com/20057755.html>
 - [17] 张前龙, 周永章, 虞鹏鹏, 等. 多层次矿床本体的构建及在知识图谱中的应用[J]. 矿物岩石地球化学通报, 2024, 43(1): 211-217.
 - [18] 廖振鑫. 面向地质领域的知识图谱构建研究及应用[D]: [硕士学位论文]. 成都: 电子科技大学, 2021.
 - [19] 中国标准化研究院. GB/T 9649.16-2009 地质矿产术语分类代码第 16 部分: 矿床学[S]. 北京: 中国标准出版社, 2009. <https://std.samr.gov.cn/gb/search/gbDetailed?id=71F772D7D343D3A7E05397BE0A0AB82A>
 - [20] Wang, B., Ma, K., Wu, L., Qiu, Q., Xie, Z. and Tao, L. (2022) Visual Analytics and Information Extraction of Geological Content for Text-Based Mineral Exploration Reports. *Ore Geology Reviews*, **144**, Article ID: 104818. <https://doi.org/10.1016/j.oregeorev.2022.104818>
 - [21] Wang, B., Wu, L., Xie, Z., Qiu, Q., Zhou, Y., Ma, K., *et al.* (2022) Understanding Geological Reports Based on Knowledge Graphs Using a Deep Learning Approach. *Computers & Geosciences*, **168**, Article ID: 105229. <https://doi.org/10.1016/j.cageo.2022.105229>
 - [22] Qiu, Q., Huang, Z., Xu, D., Ma, K., Tao, L., Wang, R., *et al.* (2023) Integrating NLP and Ontology Matching into a Unified System for Automated Information Extraction from Geological Hazard Reports. *Journal of Earth Science*, **34**, 1433-1446. <https://doi.org/10.1007/s12583-022-1716-z>
 - [23] Enkhsaikhan, M., Liu, W., Holden, E. and Duuring, P. (2018) Towards Geological Knowledge Discovery Using Vector-Based Semantic Similarity. In: Gan, G., Li, B., Li, X. and Wang, S., Eds., *Advanced Data Mining and Applications*, Springer, 224-237. https://doi.org/10.1007/978-3-030-05090-0_20