2020年2月，第26卷，第1期，002-010页
February 2020，Vol. 26，No.1， pp. 002-010

高 校 地 质 学 报
Geological Journal of China Universities

DOI: 10.16108/j.issn1006-7493.2019099
引用格式：齐浩，董少春，张丽丽，胡欢，樊隽轩. 2020. 地球科学知识图谱的构建与展望[J]. 高校地质学报，26（1）：002-010

地球科学知识图谱的构建与展望

齐 浩[1]，董少春[1*]，张丽丽[2]，胡 欢[1]，樊隽轩[1]

1. 南京大学 地球科学与工程学院，南京 210023；
2. 中国科学院 计算机网络信息中心，北京 100190

摘要：大数据为地球科学研究带来了新的思路和挑战。但由于存在描述规范不统一、共享机制不明、语义异构等问题，在数据集成、共享与复用等方面存在较大困难，使得大数据的众多优势在地球科学相关研究中难以充分发挥。知识图谱能够准确、清晰地表达概念及其相互之间的复杂语义关系，为机器所理解，是实现语义翻译、数据融合和复用的关键技术。文章对地球科学知识图谱的内涵和特点进行了深入的分析，归纳了地球科学知识图谱的主要构建方法，梳理了数据字典、知识体系和知识图谱之间的关系，对与地球科学知识图谱构建相关的专题数据库和领域本体的建设现状进行了回顾，指出了地球科学知识图谱构建中存在的主要问题，并阐述了地球科学知识图谱的应用前景，以期推动和完善地球科学知识图谱的建设和应用。

关键词：地球科学；知识图谱；知识体系；大数据；人工智能
中图分类号：P5; TP391　　**文献标识码：**A　　**文章编号：**1006-7493（2020）01-002-09

Construction of Earth Science Knowledge Graph and Its Future Perspectives

QI Hao[1]，DONG Shaochun[1*]，ZHANG Lili[2]，HU Huan[1]，FAN Junxuan[1]

1. School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, China；
2. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Big data have brought innovations as well as challenges to Earth science research. However, due to inconsistent data description standards, unclear data sharing mechanism and significant semantic heterogeneity, there are significant difficulties existed in big data integration, sharing and reuse in Earth science research. The knowledge graph can be used to explicitly represent concepts and their complex semantic relationships in a machine-understandable way. Therefore, it has been widely applied for semantic translation, data integration and data reuse. In order to establish Earth science knowledge graph, the current paper analyzes the characteristics of the existing knowledge graphs and investigates the main construction methods. Relationships of data dictionary, knowledge system and knowledge graph are also illustrated and analyzed. Current Earth science thematic databases and domain ontologies have also been reviewed, and future perspectives on Earth science knowledge graph applications are provided.

Key words: Earth science; knowledge graph; knowledge system; big data; artificial intelligence

Corresponding author: DONG Shaochun, Assciate Professor; E-mail: dsc@nju.edu.cn

经过不断的积累和持续建设，地球科学大数据已具备相当的规模，形成了种类繁多、内容丰富、覆盖面广的各类结构化专题数据库和庞大的非结构化文献资料库，在数据共享、交流和应用

收稿日期：2019-11-25；**修回日期：**2019-12-25
基金项目：国家重点研发计划（2018YFE0204201）；国家自然科学基金（40802080）联合资助
作者简介：齐浩，女，1995年生，硕士研究生，地球探测与信息技术专业；E-mail: MG1729089@smail.nju.edu.cn
***通讯作者：**董少春，女，1976年生，博士，副教授，主要从事遥感与地理信息系统应用研究；E-mail: dsc@nju.edu.cn

中发挥重要作用，为深入开展地球科学研究奠定坚实的数据基础。地球科学大数据除了具有数据量大（Volume）、类型繁多（Variety）、处理速度快（Velocity）、真实性（Veracity）和数据价值（Value）等5V特性之外（程学旗等，2014；李学龙和龚海刚，2015），还有高时空性、高度可视化、高相关性和高（多）维度等特征（郭华东等，2014；董少春等，2019）。数据的爆发式增长，改变了地球科学的传统研究方式，为地球科学领域带来新的机遇和挑战。但是，由于缺乏统一的描述规范、共享机制不明、语义异构现象显著等问题，使得多源异构的地球科学大数据在数据共享、融合和复用等方面存在诸多困难，限制了数据的大规模集成和深层次应用。

知识图谱（Knowledge Graph）最初的概念和雏形可以追溯到20世纪60年代，是随着语义网的出现不断发展成熟起来的，现在大家广泛认可的概念是谷歌于2012年提出的（Amit，2012）。知识图谱旨在提高搜索引擎的能力，增强用户的搜索质量和搜索体验，在智能检索、机器回答等领域已经得到了广泛应用。知识图谱通过图的方式揭示客观世界中的事物及其相互之间的关系，并进行形式化的描述，形成可以被人和机器理解的大规模知识库（曹倩等，2015；徐增林等，2016；黄恒琪等，2019），是实现语义翻译和数据融合的关键技术之一。因此，为整合全球地球演化数据，共享全球地学知识，推动地球科学研究范式的变革，有必要引入知识图谱的概念和思路，建立地球科学知识图谱，消除地球科学数据的语义异构瓶颈，充分挖掘地球科学数据的价值，推动大数据驱动下的知识发现，真正实现数据共享、复用和融合，深化地球科学基础研究和应用研究的发展。

本文对地球科学知识图谱的内涵、特点以及构建方法进行了全面的梳理，对地球科学知识图谱的应用进行了分析。回顾了与地球科学知识图谱建设有关的专题数据库和领域本体的国内外建设现状，对知识图谱建构中存在的问题进行了总结，以期对推动和完善地球科学知识图谱的建设工作和应用提供帮助。

1　地球科学知识图谱的特点及功能

地球科学知识图谱是对地球科学知识的全面梳理。它以科学家共同认可的知识体系为基础，利用标准化编码对地球科学领域内的所有知识点（包括基本概念、对象、现象、过程、标准、方法等）以及这些知识点之间的相互关系进行清晰、明确的阐释，并为区分各种类型的对象以及它们之间的联系提供标准，形成可以为机器所理解的地球科学知识库，具有灵活多样的可视化方式，为机器学习提供了语义翻译的基础，是跨领域数据融合和数据挖掘的基础。

1.1　地球科学知识图谱的特点

地球科学知识图谱的特点主要体现在以下三个方面。

1.1.1　准确、清晰的知识表达

地球科学知识图谱包含对知识及其相互关系的全面、清晰、明确的描述，而且采用国际标准化的编码对知识及其相互关系进行形式化的表达，具有科学性、系统性和规范性，提供与其他描述规范进行互操作的基础。知识图谱的架构具有开放性的特点，便于修改和扩充，能够在不同层面上满足对地球科学知识的需求，是进行科普、教学和科研的知识库，是领域科学家进行学术交流的通用语言和基石，更是计算机可理解的数字化、结构化的知识体系。

1.1.2　丰富的语义表达能力

地球科学知识图谱充分表达了知识点之间的对等关系（例如同义词等）、包含关系（或称为从属关系）、继承关系、实例关系和属性归属关系等丰富的语义信息，可非常清晰和方便的表示为层次化或网状化的知识体系。

1.1.3　语义推理能力

除了充分表达知识点之间丰富的原生语义关系以外，知识图谱还具有强大的推理能力，能够从原生知识关联中通过推理产生新的知识，即可将隐性知识显性化，从而为数据挖掘和知识发现提供语义推理服务。

1.2　地球科学知识图谱的功能

地球科学知识图谱提供了地球科学领域内最全面的知识体系和内容，可以应用于多样化的知识展示，满足不同层面知识获取、智能检索和智能回答、机器学习等方面的需求。

1.2.1　为不同层次的用户提供多样化知识展示

在知识图谱的构建过程中，知识体系的规范化

描述和存储与显示方式是分离的。这就使得知识图谱的展示方式具有高度灵活性和多样性，可以根据不同层次不同需求的用户定制不同的显示样式，满足不同场景的需求。知识图谱的“图”更多的是强调知识点相互之间关系的表达，而并非特指其呈现方式只局限于“图”的形式。例如：(1) 可以按照知识点的英文名称/中文名称的音序或者首字母进行排序，形成在线地球科学术语的数据字典，随时进行浏览和查询，为专业人士或非专业人士提供术语解释等（图1）。(2) 可以按照学科分类的方式以目录树的形式进行展示，类似教科书的方式，满足系统学习地球科学知识的需求，为高等院校专业学生提供丰富的教学资源库（图2）。(3) 可以方便的生成百科全书的样式，在知识点的描述中建立超链接，使得知识的展现方式不再是线性的而是网状的，利用层层开启的方式全面展示地球科学的内容，为公众普及地球科学知识提供学习途径，类似于维基百科的展示方式（图3）。(4) 知识图谱还可以方便的导出成多种格式，例如：RDF格式，excel格式，文本格式或图的格式等，满足离线学习的需求。

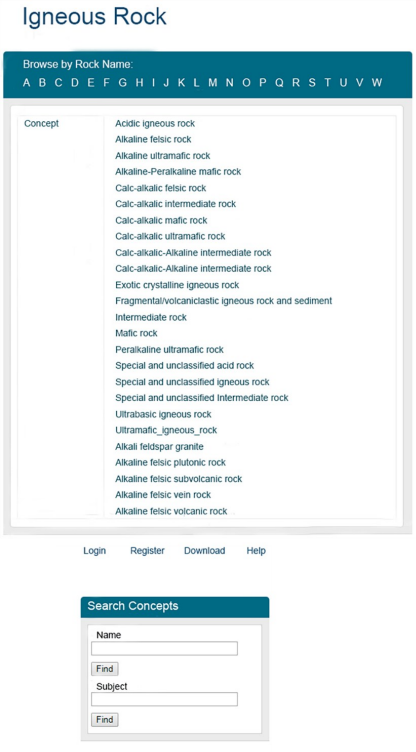

图1 数据字典展示方式示意图

Fig. 1 Data dictionary style of knowledge graph presentation

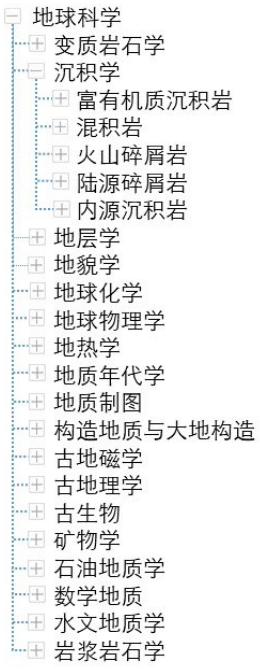

图2 目录树展示方式示意图

Fig. 2 Catalogue style of knowledge graph presentation

1.2.2 实现语义翻译，为用户提供智能检索和智能回答服务

传统的数据检索机制都是基于关键字的语法匹配和全文检索技术，主要借助于目录、索引和关键词等方法实现，而多源异构数据之间隐含的各种联系需要通过语义翻译才能体现。由于知识图谱利用规范的形式化语言诠释了地球科学领域内的概念和关系，因此可以根据知识图谱对检索请求或提问进行翻译，通过逻辑推理规则，挖掘隐含语义联系，从而实现语义翻译的功能。用户的检索请求或基于自然语言的提问可以通过语义翻译映射到知识图谱的某一个或几个节点上，使得机器能够解析出用户问题关键词的上下文语义关系，建立数据请求与基本数据集之间的映射关系，从而进行语义推理，获取最相关的信息，组成最符合用户需求的回答或检索结果集。

1.2.3 为多源异构数据库之间提供语义互联互通的服务，实现数据融合

地球科学大数据经过几十年的积累，已经形成了庞大的多源、异构的专题数据库。由于缺乏统一的建库标准，难免存在不同数据集利用不同的标识符表达相同概念，或者相同的标识符表达不同概念的情况。知识图谱可以帮助应用系统在

图3　百科全书式的展示方式（来自维基百科）

Fig. 3　Encyclopedia style of knowledge graph presentation (from Wikipedia)

不同数据集之间建立联系，识别同义异名、同名异义、包含关系、部分与整体的关系等语义联系，使得异构、多源地球科学数据之间的隐性知识显性化，使不同数据集之间的各种联系能够为应用系统所识别。在数据获取、交换以及融合过程中可以利用知识图谱进行语义翻译预处理，达到消除语义瓶颈、解决语义异构的目的，使得异构数据的访问、获取、解释和复用成为现实，让机器理解不同数据集的内在含义，帮助机器进行学习，从而在人与机器以及机器与机器之间达成领域知识共享，为数据融合和数据挖掘提供基础。

1.2.4　*为新数据库的建设提供标准，规范数据结构和数据字典*

对于新建设的数据库，可以根据知识图谱在数据库需求分析、概念模式和应用模式等不同环节建立起统一、科学、规范的标准，避免或减少新建数据库之间的语义异构问题，使得机器能够更加高效、智能的收集、处理和分析数据。

1.2.5　*为知识发现提供推理规则*

知识图谱不仅包含对原生知识体系的全面描述，同时也具备进行推理，产生新知识的能力。基于知识图谱的推理规则集和人工智能技术，应用系统能够根据其丰富的语义关系进行推理和演算，从大数据集中挖掘出新的规律和特征，为科学家认识地球提供新的视角和思路。

2　地球科学知识图谱的构建方法

2.1　自动建构方法

知识图谱的建立主要通过自动抽取技术，从结构化、半结构化和非结构化的数据中抽取出实体或概念，分析这些实体或概念之间的联系，以形式化的语言描述相互之间的语义关系，并以图的方式进行表达，形成某一领域的知识图谱（漆桂林等，2017；边慧珍和哈斯，2018；王颖等，2019）。这种方式能够充分利用大数据的特点，从大规模自然语言描述的文献库和结构化存储的关系型数据库中抽取出不同的知识点，形成数据字典，通过上下文分析这些知识点之间的联系，利用标准化的编码描述知识点和他们之间的关系，建立数字化的知识体系，具有建构速度快、自动化程度高、人工参与少的特点。为了确保准确、高效地自动抽取重要概念和对象，建立和表达概念间的关系，很多关键技术，如自动分词、知识提取、语义相似度计算、本体构建等都被应用于知识图谱的构建中（刘峤等，2016；李涛等，2017；朱月琴等，2017；王颖等，2019），并不断发展成熟起来。这种方法常见于快速建构知识内容体量较小的知识图谱，如某一本书或某个朝代

的人物关系图（Bonato et al.，2016；杨海慈和王军，2019），某种疾病的知识图谱（王淑斌，2014）以及矿产资源知识图谱（朱月琴等，2017）等。

2.2 人工建构方法

对于门类齐全、覆盖面广、知识点众多的地球科学知识图谱来说，通过自动提取方式建立的知识体系难以保证知识的准确性、系统性和完整性，无法全面、清晰的表达地球科学领域的全部内容及相互关系，需要系统、完整的地球科学知识体系作为基础进行地球科学知识图谱建设。知识体系不仅要覆盖该学科领域内所有重要的基本概念或对象（或称为知识点），而且要对这些概念或基本对象之间的关系（包含关系/从属关系、对等关系、引用关系/调用关系）进行描述。因此，要建立地球科学领域系统性、全面性、科学性的知识图谱，现阶段只能通过专家人工建设的方式进行，但是在由专家主导和控制的建设过程中可以使用一些自动和半自动的数据处理技术。

由专家建立的知识体系可以认为是原生知识体系，以自然语言进行描述，具有系统性、科学性和严谨性的特点，是科学家进行专业交流的通用语言，也是知识图谱构建的核心内容，但是无法为机器所识别，因此必须进行建模后才能成为可以被机器理解的知识体系。本体是共享概念模型的显式说明，描述概念与概念之间的关系，是语义Web的关键技术之一（Gruber，1993；Guarino，1998；Studer，2008；黄恒琪等，2019）。本体可以清晰的表达基本概念，揭示基本概念之间丰富的关系，阐明复杂的语义，建立以“概念–关系”为中心的信息描述框架。本体建模语言具有机器可理解、可处理、可扩展等诸多优点，可以作为在特定领域内有效表现概念层次结构和相互关系的模型（李曼等，2005；汪方胜等，2005；杨俊柯等，2005），能够促进该领域内不同主体（人、机器、软件系统等）之间的语义交流（对话、互操作、共享等）（杜小勇等，2004）。因此，通过本体建模语言，例如Simple Knowledge Organization System（SKOS）、Resource Description Framework（RDF）、Web Ontology Language（OWL）等，将科学家进行交流的自然语言（知识体系）转化为机器可以交流的形式化语言（知识图谱），是构建地球科学知识图谱的关键步骤。

在专家建立的知识体系的基础上，遵循W3C和相关地学国际标准，利用标准规范的本体建模语言将由自然语言表达的知识体系描述为机器可理解、可操作的地球科学领域本体，最终实现地球科学知识图谱的构建。为了方便知识体系向知识图谱的转变，在进行知识体系描述的时候需要遵循知识点尽量最小化、知识网络覆盖全面化、语义关系尽量明确化、质量评价尽量标准化的原则。通过这种方式建立的知识图谱不仅体现了最完整、最系统、最科学的知识体系，同时由于知识体系对语义关系进行了详细、清晰的表达，因此也具备从原生知识体系中通过推理获得新知识的能力。此外，知识图谱的知识体系不是封闭的，而是具有开放性、互连接的知识体系，可以随着认识的不断演进进行更新和维护。

此外，由于地球科学知识图谱对地球科学不同分支的专业词汇（基本概念或对象）进行了系统性、科学性的梳理，形成了机器可理解的领域本体，因此可以很方便地自动抽取并建立内容完整、覆盖面齐全的地球科学数据字典。

图4阐述了自动构建和人工构建两种方式下数据字典、知识体系和知识图谱之间的关系。从中可以总结出以下几点：（1）数据字典只包含知识点（概念或对象），可方便为专业人士和非专业人士提供术语解释的服务，但不包含相互之间的语义关系（例如概念之间的包含关系，对等关系，

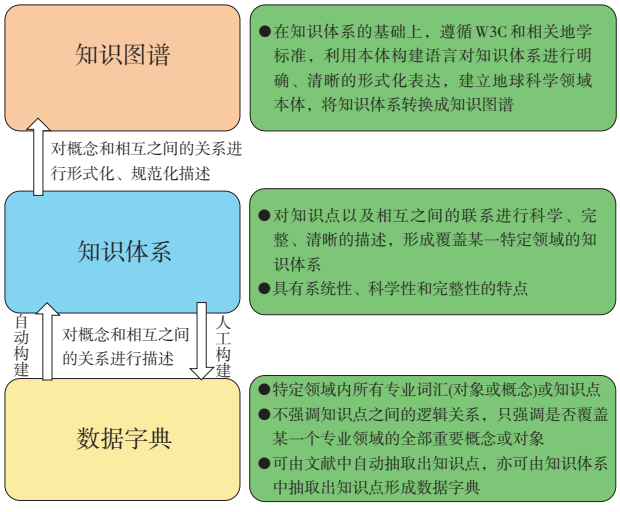

图4 数据字典、知识体系与知识图谱关系示意图

Fig. 4 Relationships of data dictionary, knowledge system and knowledge graph

属性之间的传递关系，推理等），因此无法进行语义翻译和推理，不能被机器所理解。（2）知识体系强调覆盖某一领域内全部的知识点，既包含概念或对象的描述，也包括概念间相互关系的描述。因此可以进行语义翻译，但知识体系的描述形式没有特定的要求，通常是科学家用自然语言进行描述，因此无法为机器所理解。（3）知识图谱以知识体系为基础，采用本体建模语言进行编码，以本体为表现方式，形成机器可理解的模式，是机器进行语义翻译和基于语义的数据融合的基础。

由此可见，知识图谱不仅仅是一张图谱，而是涵盖了完整的知识体系、具有形式化和规范化的知识表达，能够揭示丰富的语义内涵，为机器和人共同理解，在人与机器以及机器与机器之间达成知识共享，并提供灵活多样的可视化方式，服务于不同需求的用户。因此地球科学知识图谱的构建实际上是一项系统知识工程。

3　国内外地球科学数据库和领域本体建设现状

地球科学知识图谱是一个新的概念，还没有形成成熟的建设规范和标准体系。国内外目前并没有完整的地球科学知识图谱的建设先例。但是与知识图谱建设密切相关的地球科学专题数据库和地球科学领域本体都已经积累了相当长时间，这里主要回顾了地球科学专题数据库和领域本体建设的国内外现状，作为地球科学知识图谱建立的基础和参考。

3.1　地球科学数据库建设现状

经过几十年的不断建设，目前国际上已建成了很多以数据为中心的地球科学专业数据库，包括综合性的数据中心，也包括具体学科领域的专题数据库（参见本专辑其他文章的介绍）。综合性的数据中心主要以各国地质调查局、国家级研究/信息中心为建设主体，数据覆盖面广，数据量大。例如美国地质调查局（USGS，https://www.usgs.gov/）收集并共享了各类地质、环境、资源、灾害相关的数据资料，提供数据分类查询、元数据描述、数据下载以及在线数据浏览等服务。美国国家环境信息中心（NCEI，https://www.ngdc.noaa.gov/）主要提供海洋测深、地球空间观测、古地磁、海洋地球物理（重、磁、地震）、自然灾害等数据，支持数据在线浏览、数据下载及数据成图等功能。英国地质调查局（BGS，http://www.bgs.ac.uk/）主要提供本国地质、钻孔、地磁、地下水、岩石等数据和资料，支持数据的在线浏览和数据下载，且提供专门的应用程序来实现数据的可视化。德国地球科学研究中心（GFZ，https://www.gfz-potsdam.de/startseite/）提供地球化学、测量学、地质学、地磁学、地球物理、水文地质、古生物等地球科学各个领域的信息，用户可根据学科进行数据检索。加拿大地质调查局（GSC，https://www.nrcan.gc.ca/）存储和管理矿产资源、地球化学、地质、地热、地下水、岩石、土地测量等不同门类的数据资料，提供数据的在线浏览和下载服务。中国地质科学院地质科学数据共享网（http://www.geoscience.cn/）整合了中国基础地质、矿产地质、构造地质、物化探地质、水文地质、岩溶地质、岩矿测试以及环境地质数据，为国家和社会公众提供地质基础数据服务。

各专业领域的专题数据库建设也具备了相当大的规模和影响力。例如美国地层学数据库（Macrostrat，https://macrostrat.org/）以发育相同沉积格架的区域作为基本单位数据，目前数据覆盖北美、南美局部、新西兰等地区，内容包括岩性、沉积相、古生物和岩石地层单位数据等。古生物学数据库（PBDB，https://paleobiodb.org/）提供化石收集记录、生物化石信息、已发表的参考文献、类群分类信息、地层单位以及古地理位置信息等。古生物学和地层学专业数据库（GBDB，http://www.geobiodiversity.com/）收集生物地层学和岩石地层学数据，以剖面为基本单位，支持多种定量地层学方法的后续分析，剖面数量大，分布广，提供数据在线浏览和共享下载服务。沉积岩数据库（SedDB，http://www.earthchem.org/seddb）整合全球沉积岩石地球化学数据。磁学信息联盟数据库（MagIC，https://www2.earthref.org/MagIC）提供不同时期的古地磁和岩石地磁数据。全球火成岩数据库（GEOROC，http://georoc.mpch-mainz.gwdg.de/georoc/）收集发表的岩浆岩数据（火山岩、侵入岩、地幔包体），包括主量、微量元素、放射性和非放射性同位素，以及全岩、玻璃、矿物和包裹体的年龄分析数据等，元数据包括经纬

度、岩石分类和岩性、蚀变程度、分析方法、实验室和参考文献等。全球海底岩石学数据库（PetDB，https://search.earthchem.org/）收集全球海底岩浆岩、变质岩、矿物和包裹体等的元素化学数据、同位素数据和矿物学数据。全球地质年代学数据库（GeoChron，http://www.geochron.org/）收集全球岩浆岩、变质岩和沉积岩碎屑矿物年代学数据，年龄数据信息齐全，可视化强。全球重力场数据库（BGI，http://bgi.omp.obs-mip.fr/data-products/Gravity-Databases）提供陆地重力数据、海洋重力数据、重力参考台以及绝对重力数据。全球地下水信息系统（GGIS，https://www.un-igrac.org/global-groundwater-information-system-ggis）提供全球地下水信息，包含跨界含水层、地下水应力、地下水水质及地下水资源信息，用户可根据国家或区域进行数据检索。国际开源拉曼光谱数据库（RRUFF，http://rruff.info/）提供5477种国际矿物学会（IMA）认可矿物的光谱数据，支持全球范围内的数据共享。全球地质图数据库（OneGeology，http://www.onegeology.org/）提供全球范围内地质图网络数据服务，用户可自主选择需要在地质图上叠加的专题数据（例如地质、地热、地磁、地球物理、岩石、矿物、构造、年代地层、地质灾害及海洋观测等），支持元数据查看和数据共享下载。全球热流数据库（GHFD，http://www.heatflow.org/）提供全球热流数据，用户可根据国家进行数据检索。此外还有北美火山和侵入岩石数据库（NAVDAT，http://www.navdat.org/）、澳大利亚地层单位数据库（ASUD，https://asud.ga.gov.au/）、古生物数据库（Fossilworks，http://fossilworks.org/）、中国地球物理科学数据中心（http://geospace.geodata.cn/）等地球科学专题数据库。

这些数据库的建立推动了地球科学数据在全球范围内的共享，极大地促进了地球科学研究的发展。但是由于这些数据库建设标准规范不统一、共享机制不同、语义异构问题严重，因此在大规模数据集成、共享和融合时难以快速直接利用，限制了大数据驱动下的地球科学研究新发展。

此外，还有大量的数据以表格或其他表现形式发表于地球科学期刊论文中，也形成了庞大的专题数据集，但是由于这类数据需要从论文的文字、表格和图片中进行提取，无法直接使用，因此在数据共享、集成和复用方面存在较大困难。

3.2 地球科学领域本体建设现状

为解决数据语义异构问题，实现地球科学数据的智能化查询、共享与管理，国内外积极开展地球科学领域本体的研究。早在21世纪初，美国地质调查局（USGS）、弗吉尼亚理工学院、圣地亚哥超级计算中心等十多家科研机构、高等院校参与的GEON（Geoscience Network）就在元素和同位素、岩石学以及地质环境等方面建立了专门的本体（Seber et al., 2003；Raskin and Pan, 2005；Sinha, 2006b），用于协调异构地质图的概念模式，解决不同地质图的内容异构问题。基于本体中数据项与术语之间的映射实现了数据的自动化集成（Ludascher et al., 2003；Baru et al., 2009；Ma et al., 2012），并在专题地图的Web语义集成（Lin and Ludascher, 2003）等方面取得了一定的进展。地球科学信息管理和应用委员会（CGI）主导了基于GML的GeoSciML和EarthResourceML数据交换规范，即从基础地图数据到复杂关系地质数据库的地质数据模型和数据传输标准，建立了地球科学专业术语库。中国地质调查局还专门为此翻译出版了中文版本，并发布于CGI的官方网站上（http://www.geosciml.org/）。加利福利亚大学圣地亚哥分校对其负责的MMI项目（Marine Metadata Interoperability）进行了海洋本体的开发（Graybeal et al., 2005），用户可使用基于语义Web的查询机制查找本体概念和相关注释，进行语义映射，增强数据的互操作性，实现海洋元数据的科学管理和共享（Rueda et al., 2009；Graybeal et al., 2012）。联合国粮食及农业组织（FAO）和欧洲共同体委员会（CEC）联合编制了多语种农业叙词表AGROVOC，涵盖粮食、农业、林业、渔业和其他相关领域（Rajbhandari and Keizer, 2012），基于AGROVOC可将信息标引标准化，提高查全率和查准率，实现多语种检索和智能化检索。美国地质调查局（USGS）、加拿大地质调查局（GSC）、联邦地理数据委员会（FGDC）和地球科学信息管理和应用委员会（CGI）共同负责的NGMDB（National Geological Map Database）受控词表的开发（Richard et al., 2003；Soller et al., 2005），旨在使用统一的描述性术语或科学语言，以一致的数据结构为全球用户提供地质数据（Soller and Berg, 2005）。国内学者在石油地质本体

（潘懋等，2014）、地质灾害本体（王艳妮和刘刚，2011）、海洋生态领域本体（熊晶等，2012）、矿床领域本体（姚健鹏等，2017）等方面也进行了相关研究，建立了一批领域本体。

特别值得一提的是，由美国宇航局地球科学技术办公室（NASA Earth Science Technology Office）主导，大量志愿者参与建设的SWEET本体（Semantic Web for Earth and Environmental Terminology）是目前规模最大的地球科学本体，它基本涵盖了地球系统科学的主要研究范畴，包括7000多个基本概念。但该本体只定义了整体概念框架，没有对具体概念及其相互关系进行详细的语义描述。

此外，作为地球科学研究基础的地质时间领域本体的构建也得到了国内外学者的广泛关注。Cox 和 Richard（2005）建立了地质年代本体GTS（Geologic Time Scale）。Ma等（2012）在此基础上构建了基于SKOS的地质时间本体，并利用flash动画实现了GTS本体的可视化，用于在地质图中标注地质年代信息。董少春等（2010）提出了地质年代本体构建的编码表达，阐述了本体与关联数据库之间的映射原理，并探讨了地质年代本体在异构数据检索中的应用；侯志伟等（2015）根据地学数据中的时间概念及其特征进行时间本体建模，分析了时间拓扑关系与时态信息的确定与表达，并将时间本体应用于地学数据检索中，结果表明地学数据时间本体能够明显优化数据检索质量，在此基础上侯志伟等（2018）提出将中国地质年代与地层概念相结合构建地质年代本体，不仅能够为用户提供多样化的知识查询服务，且能够更好的解决检索中存在的语义异构问题。

地球科学领域本体的建设为解决数据语义异构、利用率低、共享困难、检索效率低等问题奠定了坚实的基础。但目前还没有一个完整覆盖整个地球科学领域的本体，无法系统、完整、清晰地表达地球科学领域内全部基本概念及其相互关系，真正成为异构数据语义翻译的基础，为大数据驱动下全球数据复用、融合提供服务。

4　讨论和展望

本文对地球科学知识图谱的特点和作用、建设的意义和构建方法进行了全面的梳理。知识图谱在数据集成和数据挖掘领域的大规模应用还是一个相对较新的方法和技术，尤其对于地球科学领域知识图谱的构建和应用，还存在着诸多亟待解决的问题。

（1）知识体系的建立缺乏统一性

随着数据、技术、方法的不断更新，科学家对地球各圈层的认识也在不断的更新。在知识点的分类、定义或者技术、方法等各个层面可能还存在争议，难以实现完全的统一。因此，知识体系的建设是一个不断完善和发展的过程，不可能一蹴而就，需要经过不断的更新、维护，才能保证知识体系的相对完整性和科学性。

（2）知识图谱的构建方法还不成熟

知识图谱的构建方法和技术都还不太成熟。本文虽然对知识图谱的构建方法进行了阐述，但是目前还没有一个成熟的方法体系可以借鉴。即便在知识图谱建设比较早的信息科学、生命科学等领域，知识图谱的建设方法和流程也没有形成特别成熟的方案，还需要根据领域特点、构建需求以及关键技术选择合理的构建方法和流程。知识图谱的完整性、可扩展性还有待检验。

（3）基于知识图谱的数据获取、访问和融合机制还不完善

如何根据已经建立的知识图谱在跨学科多源数据之间实现语义映射、翻译和推理的机制还未成熟，这关系到能否有效利用知识图谱将多源异构的大数据集与相关领域的概念联系起来，从而建立基于知识图谱的数据获取、访问和融合机制，实现异构数据之间的互操作、集成和复用，并利用机器学习等人工智能方法实现数据挖掘，开展相关的应用研究。

大数据的建设，实现了孤立、零散数据的数字化和集中化，建立了e-data的概念。而要使得e-data上升为e-science，即真正实现让数据自己说话，就需要让知识图谱充分发挥映射–翻译–推理的桥梁作用，结合数据挖掘、机器学习等人工智能技术，为地球科学问题的求解提供新的思路和认识，使得e-data真正符合“可访问、可获取、可解释、可复用”的FAIR数据原则（Wilkinson et al., 2016），并完成从e-data向e-science的转变，为大数据驱动下的地球科学问题的解答提供语义翻译和数据挖掘的基础，真正开启地球科学研究的新范式。

致谢：感谢中国地质科学院地质研究所王涛研究员、童英研究员、南京大学胡修棉教授、李超博士生在地球科学领域本体建构方面给予的建议，感谢南京大学史宇坤副教授对论文写作给予的帮助，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):

边慧珍，哈斯. 2018. 知识图谱概念获取研究进展[J]. 广西科学院学报，34(1): 46–50.

曹倩，赵一鸣. 2015. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践，38(12): 127–132.

程学旗，靳小龙，王元卓，等. 2014. 大数据系统和分析技术综述[J]. 软件学报，25(9): 1889–1908.

董少春，尹宏伟，许刚. 2010. 地质时间本体在异构数据检索中的应用[J]. 地球信息科学学报，12(2): 2194–2199.

董少春，齐浩，胡欢. 2019. 地球科学大数据的现状与发展[J]. 科学技术与工程，19(20): 1–11.

杜小勇，李曼，王大治. 2004. 语义Web与本体研究综述[J]. 计算机应用，24(10): 14–16.

郭华东，王力哲，陈方，等. 2014. 科学大数据与数字地球[J]. 科学通报，59(12): 1047–1054.

侯志伟，诸云强，高星，等. 2015. 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，17(4): 379–390.

侯志伟，诸云强，高楹，等. 2018. 地质年代本体及其在语义检索中的应用[J]. 地球信息科学学报，20(1): 17–27.

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.

李曼，王琰，赵益宇，等. 2005. 基于关系数据库的大规模本体的存储模式研究[J]. 华中科技大学学报（自然科学版），33(s1): 217–220.

李学龙，龚海刚. 2015. 大数据系统综述[J]. 中国科学：信息科学，45(1): 1–44.

李涛，王次臣，李华康. 2017. 知识图谱的发展与构建[J]. 南京理工大学学报，(1): 26–38.

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.

潘懋，闫东，张文静，等. 2014. 基于本体的地质领域知识服务系统研究[C]// 第十三届全国数学地质与地学信息学术研讨会论文集，110–115.

漆桂林，高桓，吴天星. 2017. 知识图谱研究进展[J]. 情报工程，3(1): 4–25.

汪方胜，侯立文，蒋馥. 2005. 领域本体建立的方法研究[J]. 情报科学，23(2): 241–244.

王艳妮，刘刚. 2011. 地质灾害领域本体的研究与应用[J]. 地理与地理信息科学，27(6): 36–40.

王淑斌. 2014. 中西医治疗2型糖尿病的知识图谱分析[D]. 北京：北京中医药大学.

王颖，钱力，谢靖，等. 2019. 科技大数据知识图谱构建模型与方法研究[J]. 数据分析与知识发现，3(1): 15–26.

熊晶，郭磊，徐建良. 2012. 领域本体在海洋生态知识管理中的应用[J]. 现代图书情报技术，(3): 15–22.

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.

杨俊柯，杨贯中，杨建学. 2005. 基于领域本体的学习资源管理系统框架研究[J]. 科学技术与工程，5(11): 708–711.

杨海慈，王军. 2019. 宋代学术师承知识图谱的构建与可视化[J]. 数据分析与知识发现，3(6): 109–116.

姚健鹏，郭艳军，潘懋，等. 2017. 铜矿床领域本体的构建方法研究[J]. 中国矿业，26(8): 140–145+153.

朱月琴，谭永杰，吴永亮，等. 2017. 面向地质大数据的语义检索模型研究[J]. 中国矿业，26(12): 143–149.

Amit S. 2012. Introducing the knowledge graph [R]. America: Official Blog of Google.

Baru C, Chandra S, Lin K, et al. 2009. The GEON service-oriented architecture for Earth science applications [J]. International Journal of Digital Earth, 2(S1): 62–78.

Bonato A, D'Angelo D R, Elenberg E R, et al. 2016. Mining and modeling character networks [J]. Springer International Publishing: 100–114.

Cox S J D and Richard S M. 2005. A formal model for the geologic time scale and global stratotype section and point, compatible with geospatial information transfer standards [J]. Geosphere, 1(3): 119–137.

Graybeal J, Bermudez L, Bogden P, et al. 2005. Marine metadata interoperability project: leading to collaboration [C]// IEEE International Symposium on Mass Storage Systems and Technology: 14–18.

Graybeal J, Isenor A W and Rueda C. 2012. Semantic mediation of vocabularies for ocean observing systems [J]. Computers & Geosciences, 40: 120–131.

Gruber T R. 1993. A translation approach to portable ontology specifications [J]. Knowledge Acquisition, 5(2): 199–220.

Guarino N. 1998. Formal ontology in information systems [C] // Proceedings of the 1st International Conference June 6–8, Trento, Italy: 3–15.

Lin K and Ludascher B. 2003. A system for semantic integration of geologic maps via ontologies [OL]. http://ceur-ws.org/Vol-83/sia_2.pdf.

Ludascher B, Lin K, Brodaric B, et al. 2003. GEON: toward a cyberinfrastructure for the geosciences-a prototype for geological map interoperability via domain ontologies [OL]. https://pubs.usgs.gov/of/2003/of03-471/ludascher/index.html.

Ma X G, Carranza E J M, Wu C, et al. 2012. Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services [J]. Computers & Geosciences, 40: 107–119.

Rajbhandari S and Keizer J. 2012. The AGROVOC concept scheme-a walkthrough [J]. Journal of Integrative Agriculture, 11(5): 694–699.

Raskin R G and Pan M J. 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET) [J]. Computers & Geosciences, 31(9): 1119–1125.

Richard S M, Matti J and Soller D R. 2003. Geoscience terminology development for the national geologic map database [OL]. https://pubs.usgs.gov/of/2003/of03-471/richard1/index.html.

Rueda C, Bermudez L and Fredericks J. 2009. The MMI ontology registry and repository: a portal for Marine Metadata Interoperability [C]// OCEANS 2009, 1–6.

Seber D, Keller R, Sinha K, et al. 2003. GEON: cyberinfrastructure for the Geosciences [J]. EOS Transactions, 84(S): F8.

Sinha A, Lin K, Raskin R, et al. 2006a. Cyberinfrastructure for the geosciences-ontology-based discovery and integration [OL]. https://www.nsf.gov/geo/geo-ci/index.jsp.

Sinha A, Rezgui A, Malik Z, et al. 2006b. Discovery, Integration, and Analysis (DIA) engine for ontologically registered Earth science data [R]. http://geon.geol.vt.edu/pubreps/Virginia Tech Annual Report 2006.doc.

Soller D R and Berg T M. 2005. The national geologic map database project: overview & progress [OL]. https://pubs.usgs.gov/of/2003/of03-471/soller1/index.html.

Soller D R, Berg T M and Stamm N R. 2005. Standards development for the U. S. national geologic map database [C]// Agu Fall Meeting Abstracts, IN33C-1194.

Studer R. 2008. Knowledge engineering: Principles and methods [J]. Data & Knowledge Engineering, 25(1–2): 161–197.

Wilkinson M D, Dumontier M, Aalbersberg I J, et al. 2016. The FAIR guiding principles for scientific data management and stewardship [J]. Scientific Data, 3: 160018.