The Geological Society of America
Special Paper 558

Text mining and knowledge graph construction from geoscience literature legacy: A review

Chengbin Wang*
Yuanjun Li
Jianguo Chen
State Key Laboratory of Geological Processes and Mineral Resources & School of Earth Resources, China University of Geosciences, Wuhan 430074, China

ABSTRACT

In the recent decade, knowledge graph has been a key technique under quick development in artificial intelligence. Due to its great potential for tackling big data and solving complex scientific questions in the geosciences, it has attracted the attention of both computer scientists and geoscientists. In this paper, we review concepts and technologies relevant to the knowledge graph, the workflow of geoscience knowledge graph construction, and state-of-the-art examples from several geoscience disciplines. There are two general strategies for constructing geoscience knowledge graphs: top-down and bottom-up. The detailed technologies include geoscience domain knowledge modeling, data collection, knowledge extraction, knowledge cleaning and fusion, knowledge storage, and knowledge service and discovery. A few recent studies have shown that knowledge graph is a useful tool for improving our understanding of the evolution of the Earth and can assist in data-intensive geoscience studies. At the end of the paper, we discuss the best practices from the studies reviewed and propose research topics for future work. Both knowledge and rules in existing human-curated databases and text mining from the literature should be leveraged in constructing geoscience knowledge graphs. Moreover, development of a higher level schema for existing ontology models and a comparable training corpus should be considered.

1. INTRODUCTION

Geoscience is a data-intensive field. Since the first geological map was produced by William Smith in 1815, geoscience research has produced massive heterogeneous data sets. The data set associated with the geosciences consists of structured, semi-structured, and unstructured data obtained from different sources using varied methods (Zhu et al., 2017; Wang et al., 2018b, 2021). Structured data are stored in the spreadsheet and relational database in terms of rows and columns (Jatana

*Corresponding author: wangchb@cug.edu.cn

Wang, C., Li, Y., and Chen, J., 2022, Text mining and knowledge graph construction from geoscience literature legacy: A review, *in* Ma, X., Mookerjee, M., Hsu, L., and Hills, D., eds., Recent Advancement in Geoinformatics and Data Science: Geological Society of America Special Paper 558, p. 11–28, https://doi.org/10.1130/2022.2558(02).

et al., 2012; Adam and Schultz, 2015) in databases such as EarthChem[1], Geobiodiversity Database (GBDB)[2], Macrostrat[3], Mindat[4], and the RRUFF project[5]. The unstructured data do not have a predefined schema and usually require data processing to yield semantic information and relational data (Sint et al., 2009; Li et al., 2015), such as satellite imagery, geoscience literature, scanned geological map, and image. Semi-structured data is a special form of structured data that combines the content and data structure and uses tags to label semantic elements. It does not obey the relational data model (Madani et al., 2013; Tekli, 2016), such as the XML-based online geological map in the OneGeology[6] database.

Geoscience also produces massive unstructured data that are mainly stored in the form of literature written in natural languages. Based on the estimation of Google Scholar, there are 389,000,000 records of academic literature, of which 114,000,000 are written in English (Khabsa and Giles, 2014; Gusenbauer, 2019). The GeoRef[7], a comprehensive bibliographic database in geoscience, contains over 4,200,000 records of the geoscience literature, and it continues to increase by more than 100,000 documents annually. National Geological Library of China[8] contains more than 370,000 records of geoscience documents written in Chinese. The massive geoscience literature is a valuable digital legacy from which geoscientists can make further data mining and knowledge discoveries.

In the geosciences, the study of biological evolution in the nineteenth century and the study of geo-plate theory in the twentieth century promoted our understanding of the Earth and the evolution of life. In the 1980s, Earth system science was proposed to study and explore global Earth system behaviors from a unified system viewed at multiple temporal and spatial scales through deep, cross-interdisciplinary research involving geoscience, life science, chemistry, mathematics, information science, and social science (Jacobson et al., 2000). The Earth system science initiative supported by academic organizations and governments has accelerated the accumulation of geoscience data. Therefore, how to deal with the diverse and heterogeneous geoscience data becomes a big challenge for Earth system science.

Because the knowledge graph can integrate multiple data and promote domain knowledge discovery, the knowledge graph has been favored by geoscientists in the recent decade (Wang et al., 2021). The knowledge graph can employ the triple structure to link and represent all of the knowledge from different data sources based on a schema. Beyond processing structured data, it can also extract knowledge from unstructured literature based on its supporting techniques. Knowledge reasoning also provides a powerful tool that geoscientists can use to better understand the Earth's evolution. Knowledge graph and its related technologies provide a powerful tool for processing diverse and heterogeneous geoscience big data and exploring complex questions in the geosciences.

Research related to the geoscience knowledge graph has been carried out in the fields of paleontology, geological survey, petroleum geology, geological disasters, mineral exploration, and more (Fan et al., 2020; Holden et al., 2019; Peters et al., 2017; Li et al., 2018; Zhou et al., 2020; Zhu et al., 2017). The research mainly involves the following steps. (1) The design and construction of a domain ontology model for a certain topic in geosciences to guide the construction of a knowledge graph. (2) Mapping the relational database and data model to a knowledge graph. (3) Knowledge graph construction by mining unstructured text. (4) Semantic query of knowledge graph and knowledge service. (5) Data mining and knowledge discovery.

In the core science system of the U.S. Geological Survey (USGS), multidisciplinary data management and integration were proposed to promote the work of solving complex scientific and social problems (Bristol et al., 2012). The National Science Foundation (NSF) has supported the EarthCube[9] program for a decade to develop sharable tools and cyberinfrastructure to transform geoscience research. The Deep-Time Digital Earth[10] (DDE) program was initiated in 2019 to harmonize global deep-time Earth data, share global geoscience knowledge, and transform Earth science. Gil et al. (2019) proposed the essential research agenda of an intelligent system of geosciences. In these important scientific programs and agendas, the knowledge graph is involved at different levels and is regarded as an important tool for future research in the geosciences. In the knowledge graph, the triple graph structure is employed to represent knowledge. In this way, a knowledge graph can not only organize the relational data set, but also process the unstructured text data for extracting the triple-structure knowledge from the geoscience literature. Knowledge graphs have different research focuses for different domains and scenarios. In the domain of the geosciences, knowledge graph has great potential for knowledge organization, the representation of big data, and knowledge discovery in the geosciences.

In the era of big data and artificial intelligence, using knowledge graphs and related technologies to improve the paradigm of geoscience research for solving complex problems has gradually gained consensus among geoscientists and computer scientists (Gil et al., 2019; Ma, 2021; Wang et al., 2021). In this paper, we review the status of research into text mining and knowledge graph construction from the geoscience literature and propose a few topics for the future work of geoscience knowledge graph construction and applications. The remainder of this paper is organized as follows. Section 2 introduces concepts relevant to

[1]https://www.earthchem.org/
[2]http://www.geobiodiversity.com/home
[3]https://macrostrat.org/
[4]https://www.mindat.org/
[5]https://rruff.info/
[6]http://portal.onegeology.org
[7]https://pubs.geoscienceworld.org/georef
[8]http://www.cgl.org.cn/
[9]https://www.earthcube.org/
[10]http://www.ddeworld.org/

the knowledge graph and technical systems. Section 3 reviews the workflow and key technologies of geoscience knowledge graph construction based on text mining from the geoscience literature and data mapping from relational data, and introduces the construction and application of knowledge graph in the domain of porphyry copper deposits briefly. Section 4 provides a discussion and recommends a few topics for future work. Section 5 provides a conclusion.

2. REFLECTING ON CONCEPTS AND TECHNOLOGIES RELATED TO KNOWLEDGE GRAPH

2.1. History of Knowledge Graph

A knowledge graph is a graph-structured model that employs a node–edge–node structure to organize the knowledge of a domain (Singhal, 2012; Ehrlinger and Wöß, 2016). Every piece of knowledge can be expressed using the triple structure of subject–predicate–object. The nodes represent the entities of subject and object while the edge represents the semantic relation of the predicate. The rapid development of knowledge graph in recent years has benefited from many related research fields such as semantic network, expert system, natural language processing (NLP), semantic web, database, deep learning algorithm, and high-performance computing (Feigenbaum and Buchanan, 1993; Berners-Lee and Hendler, 2001; Annervaz et al., 2018; Chen and Luo, 2019; Wang et al., 2019). Recent knowledge graph developments have absorbed the concepts and frameworks of the semantic web in terms of knowledge organization and representation, making knowledge exchange easier between computers and humans.

Although knowledge graph is now a popular topic in the academic and industrial sectors, it has taken a long time to develop (Fig. 1). Knowledge graph is an intuitive product of knowledge representation in artificial intelligence. In the 1960s, semantic network (or frame network) was proposed as a form of knowledge representation among concepts (Quillan, 1963). A semantic network was composed of interconnected nodes and edges and was mainly used in the field of natural language processing. Nodes represent concepts, and edges represent the relations among them. In this stage, the semantic links contained in the edges were weak, simple, and unable to support complex reasoning in the semantic networks. In the 1970s, the term of knowledge graph was first proposed in the instructional system of courses (Schneider, 1973). In the early stage, knowledge graphs were used to restrict the semantic relation in the semantic networks (Nurdiati and Hoede, 2008; Gebretensae, 2019). With the release of Google knowledge graph, the term "knowledge graph" had a new meaning that is now widely accepted by academia and industry.

To address the drawback of weak semantics, extensive work was carried out to enhance the semantic relations. In the 1970s, artificial intelligence research focused on expert systems composed of knowledge bases and inference engines (Fig. 1; Feigenbaum and Buchanan, 1993). In the context of an expert system, the knowledge base is a sub-system that contains a series of human knowledge used by a computer system to emulate the human decision-making process (Gaschnig, 1982; Duan et al., 2005). The knowledge graph can be regarded as a graph-structured knowledge database (Hogan et al., 2022). The inference engine is another sub-system for an expert system that employs logic rules (e.g., IF-THEN rules) in the knowledge base to deduce new information. A knowledge base with strong semantic relations was proposed as a necessary infrastructure to support the reasoning functions in an expert system (Gaschnig, 1982; Liao, 2005). Description logic model is a knowledge representation language that describes concept classification and their relations (Fig. 1) and can strengthen the semantic links for logical reasoning (Nardi and Brachman, 2010). It provides a logical formalism for ontology and the semantic web. In computer science, each ontology is the formal, explicit, and detailed description of a shared conceptual model (Gruber, 1995). The ontology model was usually used to design computational models for artificial intelligence systems (Jepsen, 2009). In the construction of a domain-specific knowledge graph, an ontology model design is a pioneering work as it is able to define a framework of entity classes, semantic relationships, and instances.

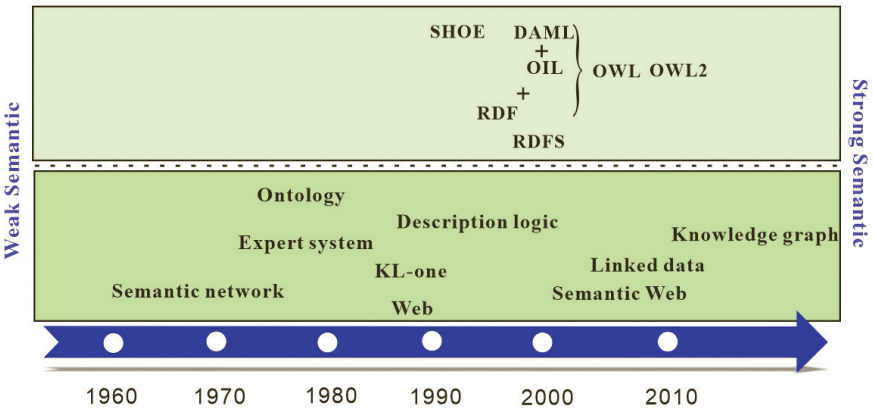

Figure 1. Timeline shows the history from semantic network to knowledge graph. OWL— Web Ontology Language; RDF—Resource Description Framework; OIL—Ontology Interchange Language; DAML—DARPA Agent Markup Language; RDFS—Resource Description Framework Schema; SHOE—Simple HTML Ontology Extensions.

The internet and the web led to explosive growth in data; however, in the early days, they could not meet the requirement for processing the complex tasks of reasoning and decision-making. The semantic network was also extended into the web domain. The semantic web proposed by T. Berners-Lee not only links textual pages through hyperlinks but also defines and links entities together to build a cyberspace knowledge base (Berners-Lee, 1998). To address the increasingly complex semantic web technology stack, linked data—a set of design principles for sharing machine-readable interlinked data on the web—was proposed to build a linked open data ecosystem (Fig. 1; Berners-Lee et al., 2008). In 2012, Google's knowledge graph was released; it is a commercial realization of some ideas from the semantic web (Singhal, 2012). Since then, the knowledge graph has entered a stage of rapid development and has been widely used in a series of research and commercial applications.

The representation of knowledge graph requires corresponding descriptive language. In the early stage, knowledge representation was based on the traditional syntactic ontology languages (e.g., Cycl[11], KL-one[12], F-Logic[13], and DOGMA[14]). After 1995, XML-based web markup language became popular for representing description logic and defining the structure of knowledge. In 1999, the first web markup ontology language was released and was further expanded into DARPA Agent Markup Language (DAML) by the U.S. Defense Advanced Research Projects Agency (DARPA) (Hendler et al., 2000). At the same time, European scientists also developed a similar markup language, the Ontology Interchange Language (OIL) (McGuinness et al., 2002). In 2004, the World Wide Web Consortium (W3C) released a new language, Web Ontology Language (OWL), which is based on the integration of OIL and DAML. In recent years, several new markup languages, serialization formats, schemas, and database structures were developed for knowledge representation, such as JSON-LD[15], Schema.org, RDFa[16], and Graph DB.

2.2. Technological Ecosystem of Knowledge Graph

With the rapid development over the last few decades, the knowledge graph is more than just a tool for improving search engines or creating a knowledge representation. It has formed a technological ecosystem with varied approaches and applications. In addition to the schema and markup language of the knowledge graph, the construction and application of the knowledge graph benefit widely from the associated techniques of data mapping, deep learning, NLP, data fusion, visualization, knowledge reasoning, and database development. In the construction of a knowledge graph, data mapping can transform structured relational data into a triple knowledge graph, while NLP is used to extract entities and semantic links from the unstructured literature. Deep learning algorithms are the supporting infrastructure for NLP and knowledge graph to train models and extract information. The database is not only used to store structured knowledge, it also supports data mapping for quick construction of knowledge graphs by mapping the existing relational databases. Visualization is the representation interface of a knowledge graph for users. Data fusion is used for semantic and entity alignment among multi-source knowledge graphs. For scientific research, knowledge discovery is a fundamental task of knowledge graph, in which semantic reasoning is the corresponding tool in many applications.

2.3. Strategy for Constructing Knowledge Graph

The strategy for knowledge graph construction can be categorized into top-down and bottom-up (Fig. 2). In the top-down strategy, the knowledge model, which includes the domain ontology model, semantic description framework, knowledge exchange syntax, and entity tagging system, is created first as a schema for a domain knowledge graph. The scope of information extraction of entities and semantic links from the unstructured literature is defined in the domain ontology model. In the bottom-up strategy, the scope and schema of information extraction are unclear at the early stage. The information extraction of entities and semantic links is based on some basic syntax rules. The extracted results require post-processing (e.g., data filtering and knowledge fusion) before the cleansed knowledge can be stored in the knowledge graphs.

The top-down strategy first determines the schema and then inputs data according to the schema constraints. The bottom-up strategy first collects the data and then extracts the schema from the data. In a domain, the scheme is relatively stable and can be derived from the domain expert knowledge. The top-down strategy can design the required entities and semantic relations for

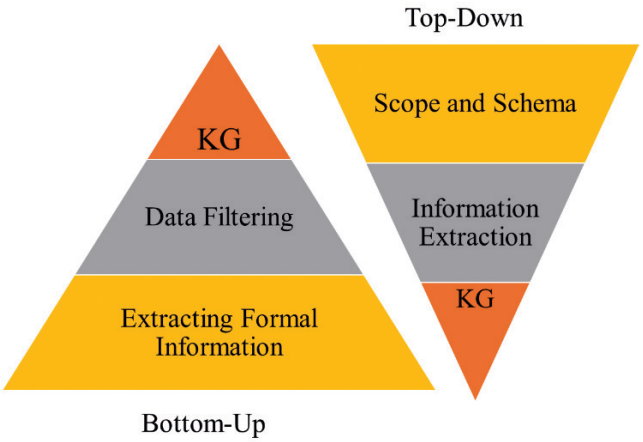

Figure 2. Two strategies for constructing knowledge graph (KG).

[11]https://en.wikipedia.org/wiki/CycL
[12]https://en.wikipedia.org/wiki/KL-ONE
[13]https://en.wikipedia.org/wiki/F-logic
[14]https://en.wikipedia.org/wiki/DOGMA
[15]https://json-ld.org/
[16]https://www.w3.org/TR/rdfa-primer/

application scenarios. Due to the clear domain schema in the geoscience domain, the top-down strategy is the preferred method for extracting information from unstructured geoscience literature and integrating the existing relational data for knowledge graph construction. Increasing data will cause the evolution of schema in the general field; the bottom-up strategy is the appropriate way to construct a general knowledge graph.

3. KNOWLEDGE GRAPH CONSTRUCTION IN THE GEOSCIENCES

Due to the clear and stable domain schema in the geoscience domain, the top-down strategy is preferred for constructing a geoscience knowledge graph. Construction of a geoscience knowledge graph is a systematic work that requires cooperation among geoscientists and computer scientists. The construction process mainly includes six aspects (Fig. 3). In geoscience knowledge modeling, computer scientists assist geoscientists in designing the knowledge model for the selected domain. In the construction process from knowledge extraction to knowledge storage (Fig. 3), the main works are performed by computer scientists, while geoscientists review the early knowledge graph and provide the support of expert knowledge. In the landing scenario of a knowledge graph, profound cooperation is needed among geoscientists and computer scientists to provide knowledge service and enable knowledge discovery in applications.

3.1. The Workflow of Knowledge Graph Construction

3.1.1. Geoscience Domain Knowledge Modeling

Knowledge modeling mainly includes ontology modeling, semantic description framework, knowledge exchange syntax, and entity-relation tagging systems (Fig. 3). The semantic description framework is used to define the basic data model and logical structure in the knowledge graph. The exchange syntax refers to the data exchange formats. Some mature solutions for semantic description and knowledge exchange syntax are provided by W3C, which recommends the Resource Description Framework (RDF) that uses the triple structure of subject–predicate–object to represent knowledge. The knowledge exchange syntax can use JSON-LD, Turtle[17], and other formats recommended by W3C.

The ontology model and entity-relation tagging system are the key tasks in knowledge modeling. Each ontology is the formal specification of the shared conceptualization of a domain of study (Gruber, 1995). The ontology model can define the classes, property, semantic relation, and vocabulary in a geoscience knowledge graph. The ontology model is designed to satisfy the objectives and requirements of a geoscience knowledge graph based on the existing ontology model, data model, terminology, expert knowledge, and relational database model in the geoscience domain. The ontology model can be edited and designed on the platforms of Protégé, Ontolingua, OntoSaurus, WebODE, OntoEdit, OilEd, WebOnto, and TopBraid Composer (Lambrix et al., 2003; Roche, 2003). The entity tagging system employs a series of tokens to mark the semantic information in the geoscience text data and is used to recognize the semantic units that carry the greatest text data information. Therefore, the ontology model can provide terminology and domain knowledge for designing the entity tagging system and the corresponding geoscience tokens.

Constructing an ontology model and entity tagging system in the geosciences is not a simple task. Both must be continuously improved. Therefore, a hybrid method combining collaboration, loop iterative evolution, and expert evaluation is often used in the construction process. That is, in the early stage, a small working group of geoscientists and computer scientists draws up a geoscience domain ontology model and entity tagging system, and then the established model is corrected and updated using iterative evolution. Finally, domain experts are invited to review the ontology model and entity tagging system to ensure that the knowledge model is robust and representative.

Geoscience ontology models had been studied before knowledge graph was widely studied and applied in the academic and industrial fields. As a shared conceptualization of domain knowledge, the geoscience ontology model has been designed for geoscience domain knowledge representation, linked data in the geosciences, knowledge integration, and development of a geoscience information management system (Cox and Richard, 2005; Raskin and Pan, 2005; Fox et al., 2009; Ma et al., 2012, 2014; Wang et al., 2018a; Garcia et al., 2020; Mantovani et al., 2020). A typical geoscience ontology model is in the domain of the geologic time scale, which is a chronological framework for Earth history. Continuous discoveries in stratigraphy and paleontology result in the geologic time scale being updated frequently by the International Commission on Stratigraphy. A series of ontology models has been designed to represent the geological time scale chart from different viewpoints (Cox and Richard, 2005, 2015; Ma et al., 2011, 2012, 2020; Wang et al., 2018a). In addition to geologic time scale ontology, ontology models in other geoscience branch domains have also been designed, such as the Semantic Web for Earth and Environmental Terminology (SWEET) designed by NASA (Raskin and Pan, 2005) and those in mineral exploration (Mentes, 2012), petroleum (Li et al., 2010), structural geology (Babaie et al., 2006; Zhong et al., 2009), geological map (Mantovani et al., 2020), geological hazard (Liu et al., 2010), and marine science (Rueda et al., 2009).

3.1.2. Geoscience Data for Constructing Knowledge Graph

In the geosciences, the data set that is used to construct knowledge graphs mainly includes structured databases and unstructured text data (Fig. 3). The structured data consist of a geoscience relational table and database. The relational geoscience database consists of a series of tables, and tables containing relational records can be regarded as a special type of

[17]https://www.w3.org/TR/turtle/

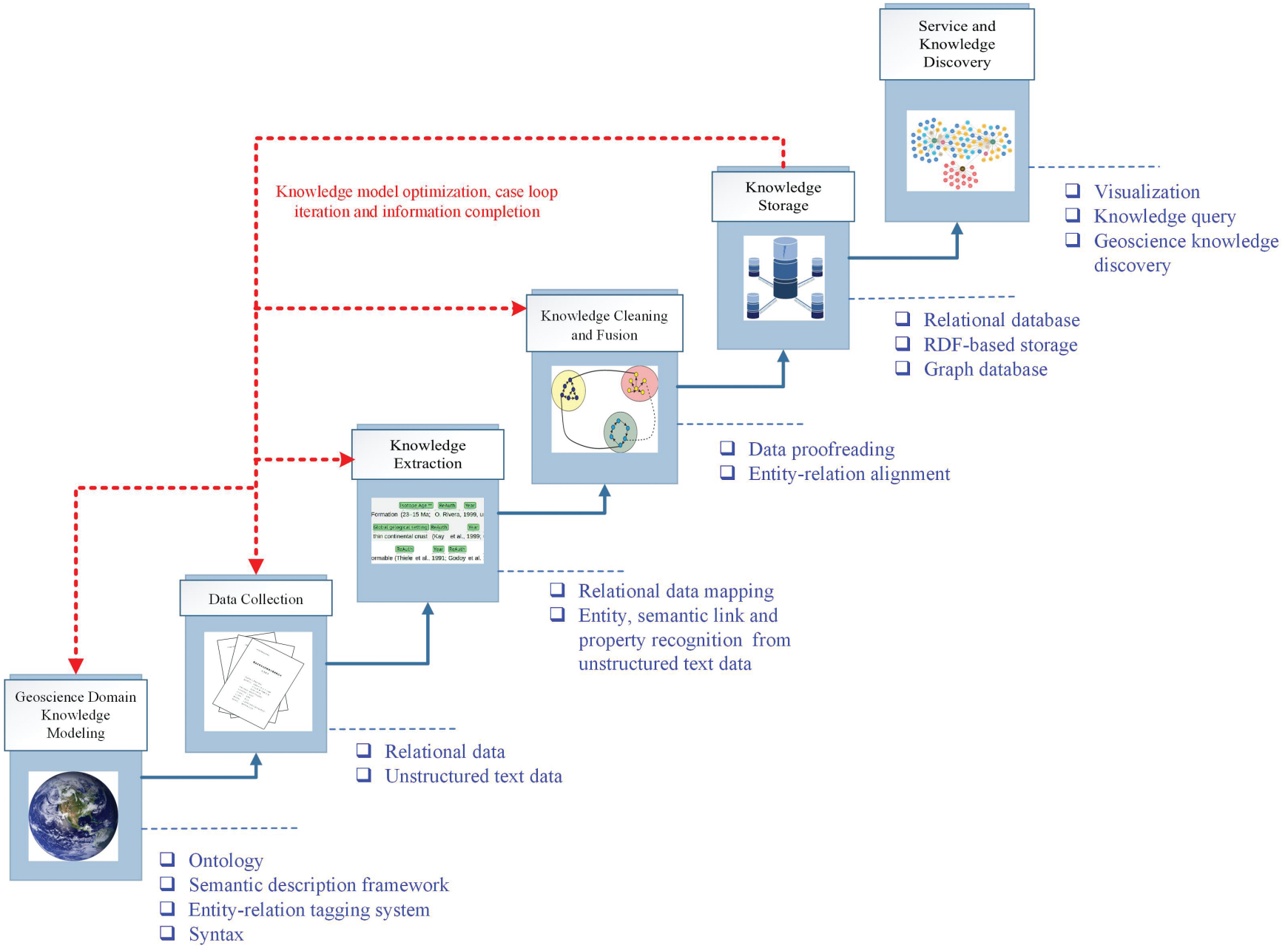

Figure 3. The workflow of knowledge graph construction in the geosciences. RDF—Resource Description Framework.

relational database. The relational geoscience data mainly come from online databases. Table 1 lists some representative online geoscience databases. The relational data not only provide the logical table for triple mapping, they also provide a seed of knowledge for producing training data quickly. "Unstructured geoscience data" mainly refers to text data from academic papers and geological survey reports derived from the geological survey programs and the published literature in an academic literature database. Due to the diversity of data sources, data cleaning and calibration of the raw data is required, such as the crawling and downloading of academic literature, text and table recognition from the literature, data correction, unified coding of documents, and the building of text corpus.

3.1.3. Knowledge Extraction from Geoscience Data

In addition to using the manual method, the computer-aided method is a cost-effective way to construct a geoscience knowledge graph. The relational geoscience tables and databases can be transformed into RDF data using mapping languages (Fig. 3). R2RML and direct mapping (DM) are two mapping languages recommended by W3C. DM is used in the simple transformation from the relational database to RDF, while R2RML is a customized mapping language that provides the functions to view relational data in the RDF model. A logical table from the relational data is mapped into a triple structure of subject–predicate–object through R2RML and DM mapping (Fig. 4). A logical table can be a basic table, view, or SQL query result. Each row in the logical table is mapped into several RDF triples that include subject mapping and multi-predicate object mapping. Finally, the RDF triples are combined to form a unified knowledge graph.

Unstructured literature is another main data source for the construction of geoscience knowledge graphs (Fig. 3). Low-dimensional entity and relational information extraction from the high-dimensional geoscience literature is a complex task and a great challenge that is difficult for traditional methods of processing structured data. NLP is the main way to extract entities and their relations from the geoscience literature for text mining. From the view of NLP tasks in the construction of a geoscience knowledge graph, the studies mainly include word segmentation for language without space between words, geological entity recognition, and extraction of the semantic link between entities (e.g., Luo et al., 2017; Wang et al., 2018c; Qiu et al., 2018; Consoli et al., 2020). In geoscience literature, English is not the only written language. In some languages, like Chinese, words are not naturally separated by spaces as in English. For example, geological texts in Chinese need to be segmented into a series of semantic units of vocabulary before they can be processed further. The tasks of word segmentation and geological entity recognition are similar. The difference between them lies

TABLE 1. A LIST OF REPRESENTATIVE ONLINE GEOSCIENCE DATABASES

Database	Discipline	Content	Reference
National Mineral Deposit Database of China	Geology and mineral resources	Deposit name, location, mineralization type, size, utilization status, genetic type, geological work level	http://ngac.org.cn/kuangchandi/
Mineral Resources Online Spatial Data		Geological map, global mineral deposits, geochemical and geophysical survey data	https://mrdata.usgs.gov
OneGeology		Global geological map	http://portal.onegeology.org/OnegeologyGlobal
PBDB	Paleobiology	Distribution and classification of fossil animals, plants, and microorganisms	https://paleobiodb.org/
GBDB		Section-based stratigraphic and paleontological information	geobiodiversity.com
EarthChem	Petrology, geochemistry, geochronology	One-stop-shop databases including PetDB, SedDB, NAVDAT, MetPetDB, the U.S. Geological Survey National Geochemical Database, GEOROC, and GANSEKI	https://earthchem.org/
GeoKem	Petrology, geochemistry	Composition of all volcanic and igneous centers	http://www.geokem.com/
LEPR	Geochemistry	Elemental partitioning in magmatic systems	http://traceds.ofm-research.org/access_user/login.php
Mindat	Mineralogy, mineral deposit	Minerals and their localities, deposits, and mines	https://www.mindat.org/
RRUFF	Mineralogy	Raman spectra, X-ray diffraction, and chemistry data for minerals	https://rruff.info/
Macrostrat	Sedimentary, petrology, paleobiology	Spatial and temporal distribution of sedimentary, igneous, and metamorphic rocks as well as data extracted from them	https://macrostrat.org/

Note: PBDB—Paleobiology Database; GBDB—Geobiodiversity Database; LEPR—Library of Experimental Phase Relations.

A

ID	Deposit name	Altname	Lat	Lon	Age (Ma)	Ore tonnage	Cu grade	Mo grade	Au grade	Ag grade
1	Agua Rica	Mi Vida	-27.37	-66.28	65	1761	0.42	0.03	0.18	3.2

B

<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_longtitude> "-66.28" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_cugrad> "0.42" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_latitude> "-27.371" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_aggrade> "3.2" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_augrade> "0.18" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_depositage> "65" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_oreton> "1761" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_altName> "Mi Vida" .
<http://mineraldeposit/Agua_Rica> <http://mindep/mineraldeposit_mograd> "0.033" .
<http://mineraldeposit/Agua_Rica> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://mindep/mineraldeposit> .

Figure 4. A demo case of data mapping from relational data to Resource Description Framework (RDF) data. (A) The records in the relational database; (B) the mapping result of RDF.

in the different tokens used. The token in the word segmentation indicates the start and end of a multi-character word, whereas the token in the entity recognition indicates the start and end of an entity word that is customized by domain knowledge. Word segmentation is performed to divide the high-dimensional text data into words and character combinations with semantics separated by spaces or slashes based on the tokens. Entity recognition is performed to extract the named entities defined in the entity tagging system from the unstructured geoscience literature and treat non-entity words as a class of OTHERS.

The methods of entity recognition are categorized into rule-based, machine learning, and deep learning. The rule-based methods use the customized rules and domain dictionary to extract entities based on string matching. Dictionary matching is a typical method for entity recognition, such as the forward maximum matching method (Li et al., 2009; Zhang et al., 2006), backward maximum matching (Bhasuran et al., 2016; Gao et al., 2005a), and the shortest path method (Gao et al., 2005b). The features of syntax and text data parts of speech (e.g., inverse document frequency, typical prefixes, and suffixes) were also used to design rules for entity recognition from the literature (Kim and Woodland, 2000; Chiticariu et al., 2010; Zhang et al., 2013; Eftimov et al., 2017). The rule-based methods usually perform with high precision and low recall in a domain. They are weak in generalization ability and are difficult to apply to a different domain (Karystianis et al., 2017).

Recently, many methods of machine learning and deep learning have been proposed and used in the NLP tasks of entity recognition. As the supervised methods, the machine learning and deep learning methods both train entity recognition models based on the training corpus. The biggest difference between the two methods is that the deep learning method uses neural networks to design an end-to-end pipeline rather than designing a series of features for a machine learning algorithm. The supervised methods in the NLP tasks include three steps of distributed representation, context encoder, and decoder. The text data are regarded as a high-dimensional column vector that is a challenge for memory management and computer processing in the NLP tasks (Dhillon et al., 2002; Hotho et al., 2005). To enable the computer to process text data, the text data need to be transformed into numerical data and can be represented by one-hot encoding. For a corpus of length N, every word can be represented as an N-dimensional vector. In a one-hot vector, the position of the encoded word is set to 1 while the remaining positions are set to 0. However, the one-hot vector is sparse, and it is difficult to describe the contextual semantics (Johnson and Khoshgoftaar, 2020). Similar words have similar semantics, and the semantics of the words are determined by their context. The word in the corpus can be mapped into a *k*-dimensional word vector based on the neural networks, and the vector transformation of text is named as word embedding or distributed representation. The semantic similarity is determined by the word vector distance. The continuous bag of words model (CBOW) and skip-gram are widely used for word embedding in the NLP (Guthrie et al., 2006; Mikolov et al., 2013). The encoder is designed to mine the hidden patterns between contexts in the word embedding sequence. The commonly used encoders include convolutional neural network, recurrent neural network, recursive neural network, deep transformer, and language model.

The decoder takes the features from encoders as input and outputs the entity labels.

The machine learning methods used in entity recognition mainly contain hidden Markov model (Morwal et al., 2012), decision tree (Szarvas et al., 2006), maximum entropy (Chieu and Ng, 2003), support vector machine (Ekbal and Bandyopadhyay, 2008), and conditional random field (CRF) (Sobhana et al., 2010). Recently, there have been more and more case studies of deep learning applications in entity recognition, such as convolutional neural network (Chen et al., 2019), recurrent neural network (Liu et al., 2017), recursive neural network (Li et al., 2017), transfer learning (Lee et al., 2017), active learning (Shen et al., 2017), reinforcement learning (Fang et al., 2019), generative adversarial network (Zhang et al., 2019), and long short-term memory (Hu et al., 2018).

Most studies on the text mining of geoscience literature are focused on geological entity recognition (Table 2). The method of entity recognition includes dictionary matching, machine learning (e.g., CRF), and deep learning methods (e.g., BiLSTM, attention-based BGRU, and deep belief networks). The performance of geological entity recognizers trained in different studies varies greatly. Due to differences in training test corpus, entity types, and algorithms in these studies (Table 2), the performances of these entity recognizers are not comparable. However, these case studies have proven that it is feasible to extract entities from the unstructured geoscience literature based on existing techniques.

In addition to geological entities, the geological literature contains a large amount of property information of a special entity type, which is embedded in the geological text in the form of tables and text descriptions. This property information has obvious similarities to the forms of numerical values plus physical units, such as age, temperature, element concentration, and other properties. Therefore, the regular expression method is preferred for extracting attribute information from the geoscience literature.

Geological entities represent the nodes of the literature, while semantic relationships provide links between entities to form the information network of the geoscience literature. The entity represents the subject and object information in the triple structure. The semantic relationship represented by the predicate is also important semantic information in the geoscience literature. In the process of constructing the knowledge graph, the semantic link between entities needs to be extracted. According to how the training corpus is built, semantic link extraction methods can be divided into supervised learning and weakly supervised learning methods. The supervised learning methods require the labeling of a large amount of corpus while weakly supervised learning requires only a small amount of labeled data for model learning.

In recent research of the weakly supervised methods, multi-task transfer learning, bootstrapping, active learning, and label propagation were used to extract the semantic relations (Chen et al., 2006; Jiang, 2009; Zhou et al., 2010; Angeli et al., 2014; Pawar et al., 2017; Qu et al., 2018). Besides, some unsupervised methods such as clustering and the template-based approach were also used in the relation extraction (Chambers and Jurafsky, 2011; Sun et al., 2011; Adel et al., 2018; Das et al., 2019). A few studies were focused on semantic link extraction from the geoscience literature. The simplest way to extract semantic links is to use the co-occurrence frequency to represent the semantic relationship between adjacent geological entities (Wang et al., 2018c). Moreover, factor graph and attention-based BGRU were also used in the extraction of semantic links from the geoscience literature (Table 2; Peters et al., 2014; Zhang, 2015; Zhou et al., 2020).

PaleoDeepDive and GeoDeepDive are the most influential cases of text mining in the geosciences. PaleoDeepDive, a machine reading and learning system, was developed to extract fossil information from the literature and update the paleobiology database (PBDB) (Peters et al., 2014). GeoDeepDive, the updated version of PaleoDeepDive, is still ongoing and extracts geological unit information for North America from the geological literature and builds the macrostrata (https://macrostrat.org/). Two factors have led to the success of these two cases: (1) The research has received in-depth technical support from geoscientists and computer scientists, especially the support of Stanford CoreNLP and DeepDive. (2) The paleontology database collected in the early stage provides a large number of training data, dictionaries of geological terms, and semantic rules.

3.1.4. Knowledge Fusion

Due to the diverse and multilingual data sources, the knowledge acquired in the knowledge extraction process is vague and heterogeneous and includes such items as different names of the same entity, different references of the same entity, and other heterogeneous problems. Entity disambiguation and entity alignment are required to address these issues (Fig. 3). Entity alignment mainly depends on similarity-based methods, rule-based methods, and division-based methods. The similarity-based method uses the triangular inequality feature of the metric space to filter out the entity pairs that do not satisfy the mapping conditions, realize the alignment of entities, and eliminate the heterogeneity (Euzenat and Valtchev, 2004). The division-based method divides the large-scale knowledge graph into several small knowledge graphs for matching to reduce the total time required for similarity calculation (Rahm et al., 2004). The rule-based method usually requires different rules for different source data. Probability and semi-supervised learning methods are introduced into the rule-based method to build matching rules automatically and reduce subjective interference (Suchanek et al., 2011; Niu et al., 2012). Furthermore, broad geoscience knowledge graph construction also requires knowledge of fusion technologies to integrate the knowledge graphs of different disciplines.

3.1.5. Knowledge Storage

Knowledge graph storage methods include relational databases, RDF-based storage, and graph databases (Fig. 3). The method based on a relational database needs to map the triple relationship in the knowledge graph into a relational database,

TABLE 2. A LIST OF TEXT MINING TECHNOLOGIES AND THEIR APPLICATIONS IN THE GEOSCIENCES

No.	Topic	Research content	Method	Language	P*	R*	F*	Reference
1	Word segmentation	Chinese word segmentation	BiLSTM	Chinese	85.59	85.52	85.55	Qiu et al., 2018
2		Geological word segmentation in Chinese	CRF	Chinese	94.14	91.40	92.75	Wang et al., 2018c
3		Chinese word segmentation	CRF	Chinese	91.50	92.20	91.80	Huang et al., 2014
4	Named entity recognition	Extract entities of geologic time, geological structure, rock, and stratum from geological reports in Chinese	Attention-based BiLSTM model	Chinese	86.74	86.05	86.39	Qiu et al., 2019
5		Eon, era, period, epoch, age, siliciclastic sedimentary rock, carbonate sedimentary rock, chemical sedimentary rock, organic-rich sedimentary rock, Brazilian sedimentary basin, basin geological context, lithostratigraphic unit, and miscellaneous data relevant to the oil and gas industry	BiLSTM-CRF	Portuguese	86.63	82.71	84.63	Consoli et al., 2020
6		Mineral, rock, ore deposit, timescale, strata, and location from geological report in English	Dictionary matching	English	78.44	84.19	80.96	Enkhsaikhan, 2021
			Character-Level LSTM + Word-Level BiLSTM		76.35	79.24	77.59	
			Word-Level BiLSTM		77.54	79.98	78.53	
7		Extract entities related to location, method, and data from geological hazard literature in Chinese	CRF	Chinese	82.10	77.56	79.81	Fan et al., 2020
			BiLSTM-CRF		92.05	94.19	93.10	
			The deep, multi-branch BiGRU-CRF model		94.13	94.25	94.19	
			Att-BiLSTM-CRF	Chinese	89.69	89.52	89.61	
8		Extract entities related to eon, era, period, and sedimentary basin from literature in Portuguese	CRF	Portuguese	76.78	43.27	54.33	Amaral et al., 2017
9		Extract entities related to country, state, waterbodies, mineral, person, organization, city, region, mountain, island, river, village, measures, year, fault, rock, and time	CRF	English	77.05	77.27	75.81	Sobhana et al., 2010
10		Minerals, commodity names, geological eras, rocks, stratigraphic units, mineralization styles, location names, mines, tectonic setting names, and regions	Dictionary matching	English				Enkhsaikhan et al., 2018
11		Time and spatial features, property, entity relationship, and feature relationship	Deep Belief Networks	Chinese	93.16	95.90	94.51	Zhang et al., 2018
12		Incident description, tools and equipment, workover steps from the text in the oil and gas fields	Bi-LSTM+SoftMax	Chinese	83.00	91.00	87.00	Zhong et al., 2020
			Bi-LSTM+CRF		85.00	98.00	92.00	
			Word2Vec-Bi-LSTM+CRF		87.00	100.00	93.00	
13		Stratum, geological history, paleobionts, geological structure, rock, and other	Dic-Att-BiLSTM	Chinese			86.55（RGR）	Qiu et al., 2020
							91.18（GJP）	
14		Geological timescales, mineralogy, host rock types, and alteration types	Dictionary matching, TF-IDF, TextRank, POS	Chinese				Holden et al., 2019
15	Semantic link	Semantic information in geological text in Chinese	Attention-based BGRU and highway network	Chinese	76.80			Luo et al., 2017
16	Entity and semantic link	Extract entity and semantic information related to paleontology	DeepDive	Multilingual				Peters et al., 2014; Zhang, 2015
		Extract entity and semantic link to construct petro knowledge graph	Bert, language technology platform, manual rule	Chinese				Zhou et al., 2020

*P—Precision; R—Recall; F—F-measure.

which requires a serious workload for the query and storage of semantic information. The method based on RDF storage is incompatible with the massive property information. The graph database is a NoSQL database and mainly uses nodes and edges to organize data. Nodes represent entities in the knowledge graph, and edges represent semantic links among entities. Because there is a large amount of valuable property information in the geosciences, the graph database can add labels and key values to the nodes to represent the classification and property information of the entity, which is appropriate for knowledge graphs.

3.1.6. Geoscience Knowledge Service and Knowledge Discovery

The purpose of knowledge graph research is to construct a knowledge graph through various knowledge acquisition methods and provide the service for knowledge semantic query, semantic reasoning, and visualization of temporal and spatial information (Fig. 3). There are some knowledge services associated with knowledge graph available online for geoscientists and the public. For example, PBDB, Macrostrat, PetroKG (Zhou et al., 2020), and GeoDocA (Holden et al., 2019) are well-known data and knowledge service systems.

The PBDB was started in 1998 (Callaway, 2015). In the early stage, it is just a relational database for paleontologists to store fossil information collected from field surveys and published papers. There are 410 contributors from over 130 institutions in 24 countries who have contributed to PBDB[18]. In 2013, S.E. Peters and Chris Ré collaborated to explore the feasibility of using text mining to extract paleontological information from the published literature (Peters et al., 2014; Callaway, 2015; Peters and McClennen, 2016). Then PaleoDeepDive, a customized version of DeepDive for paleobiology, was created to extract fossil information from the literature and update the PBDB database. As of April 2021, The PBDB includes 76,068 references, 434,743 taxa, 819,564 opinions, 219,016 collections, and 1,515,784 occurrences. PBDB is not only a paleontology database; it also provides excellent data query, Web app, API interface, mobile app, R library, and paleontology analysis tools (Peters and McClennen, 2016; Varela et al., 2015).

Macrostrat[19] is the world's largest homogenized geologic map database developed based on the GeoDeepDive[20] and integrates the spatial and temporal distribution information of sedimentary, igneous, and metamorphic rocks. It provides a cyberinfrastructure for geoscientists to study crustal formation and destruction and paleontological evolution (Husson et al., 2016; Peters et al., 2018). The database not only contains PBDB paleontological data, it also contains a large amount of rock and stratigraphic chronology data organized in columns, units, polygons, and packages. As of April 2021, it includes 1534 regional rock columns, 35,478 rock units, 2,540,323 geologic map polygons, and 51,212 stratigraphic names from North America, the Caribbean, New Zealand, and the deep sea. Macrostrat provides access to the web app, mobile app, and API to explore the data for research and education.

In addition to online services, there are also some non-open services and databases associated with knowledge graphs in the geoscience field, such as GeoDocA, PetroKG, and GeoCloud. GeoDocA was developed by the University of Western Australia to assist in the search and analysis of mineral exploration reports (Holden et al., 2019). PetroKG is a knowledge graph in the upstream of PetroChina (Zhou et al., 2020). China University of Geosciences in Wuhan developed a system of automatic indexing and summarization of geological reports in Chinese. Zhu et al. (2017) developed the knowledge graph for mineral deposits. These functions are integrated into the GeoCloud, which was developed by the China Geological Survey.

Knowledge graph is helpful for improving data mining and promoting knowledge discovery in the geosciences (Fig. 3). Knowledge graph can enhance the interpretability of semantic information and data models in the data mining process and improve the reasoning ability of geoscience domain knowledge. In the petroleum industry, well log interpretation is necessary for potential reservoir detection and classification. The expert rule and feature engineering defined in the PetroKG improved the accuracy of well log interpretation by more than 7.69% over the traditional machine learning approaches (Zhou et al., 2020).

PBDB and Macrostrat not only provide the data and knowledge service for research and education, but also promote our understanding of geological and biological evolution. For instance, the stratum statistics of stromatolites and their production based on natural language processing show that the appearance of stromatolites has a strong correlation with the growth of the total amount of dolomite rather than with mass extinction (Peters et al., 2017). The amount of sedimentary rock is related to the change of oxygen and the evolution of life, which indicates that the unstable evolution of sedimentary rocks drives changes in oxygen, which in turn drives the evolution of life (Husson and Peters, 2017, 2018). During the transition period between the Neoproterozoic and Paleozoic, the sediment volume increased by up to five times. A large amount of sediment was eroded before the Cambrian and corresponds to the great unconformity in North America (Husson and Peters, 2018). These new knowledge discoveries are supported by PBDB, Macrostrat, and even GeoDeepDive.

3.2. Porphyry Copper Deposit Knowledge Graph

Mineral deposits are produced by the coupling of multiple geological processes and always have a close relationship with geological events related to the evolution of the Earth. Mineral deposits are regarded as a window for studying geodynamics. Porphyry copper deposits are located in the island arc, continental marginal arc, and collision orogenic setting and are an important raw material for the global economy (Sillitoe, 1972, 2010; Singer

[18]https://paleobiodb.org/#/
[19]https://macrostrat.org/
[20]https://geodeepdive.org/

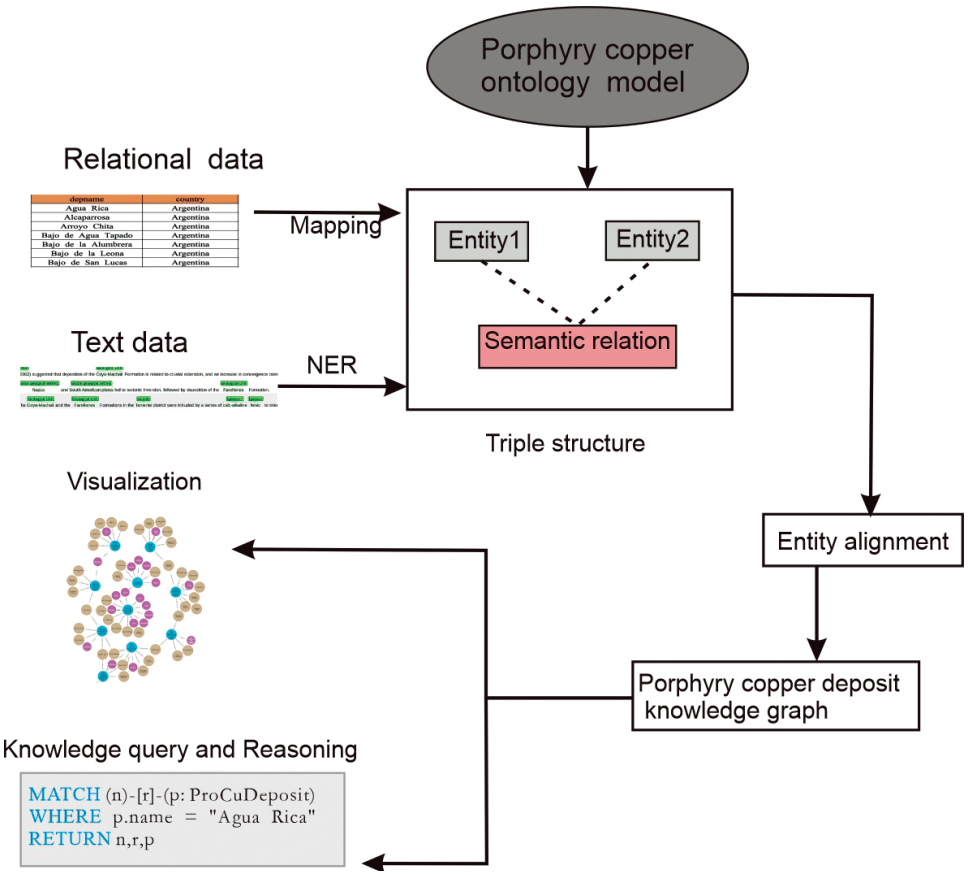

Figure 5. The workflow for constructing a knowledge graph in the domain of porphyry copper deposits. NER—named entity recognition.

et al., 2005). The formation and preservation of porphyry copper deposits are closely related to volcanic rocks, strata, paleoenvironment, geological structure, geodynamic environment, and mineralogy. Understanding porphyry copper deposits depends on geochemistry, geophysical, mineral exploration, and other research techniques. Therefore, the porphyry copper deposit was selected as the domain for constructing a knowledge graph.

The data used in constructing the porphyry copper deposit knowledge graph include the USGS global porphyry copper deposit database[21] and 37 academic papers with entity annotation. The workflow of knowledge graph construction in the porphyry copper domain is shown in Figure 5. First, an ontology model was designed based on the geological information related to the porphyry copper deposit model. There are 47 entity types, and five relation types were created based on the classes in the porphyry copper deposit ontology model (Table 3). Second, the relational data were mapped into the triple structure of the knowledge graph according to the customized rule based on the Neo4j platform[22]. The entities and relations from the geoscience literature were aligned with the mapping knowledge graph based on the similarity coefficient. Finally, a knowledge graph of porphyry copper deposit can be built and represented by the network graph (Figs. 5–6). Based on the knowledge graph, knowledge queries and reasoning can also be carried out for knowledge discoveries (Fig. 6).

4. DISCUSSION AND RECOMMENDATIONS FOR FUTURE WORK

Geoscience knowledge graph provides a solution for dealing with complex geoscience questions in the era of big data and is gradually being accepted by geoscientists and computer scientists. Although geoscience knowledge graph research began in recent decades, results show that it is valuable for understanding the Earth. Based on reviewing the state-of-the-art work in the geoscience knowledge graph, the following issues merit discussion and study in the future.

4.1. Knowledge Representation Is the Key to Knowledge Graph

The state-of-the-art progress in the geoscience knowledge graph gives us some inspiration for further application and development in numerous geoscience disciplines. There are many studies on text mining for geoscience knowledge graph construction. Text data mining is one of the supporting technologies that can

[21]https://mrdata.usgs.gov/porcu/
[22]https://neo4j.com/

TABLE 3. A LIST OF ENTITY AND RELATION TYPES USED IN THE CONSTRUCTION OF KNOWLEDGE GRAPH IN THE DOMAIN OF PORPHYRY COPPER DEPOSITS

Entity type	Relation type
PorphyryCopperDeposit, Geographic, GeologicalSetting, Petrology, Mineralization, Mineral, Alteration, Geochemistry, Geochronology, Fluid, Document, Continent, Country, Province, City, Locality, GlobalGeologicalSetting, RegionalGeologicalSetting, IgneousRock, MetamorphicRock, SedimentaryRock, IgneousForm, MagmaType, MineralizationType, MineralResourceSpecies, MineralizationStage, OreBody, OreTextureStructure, MetalMineral, PrimaryMineral, AlterationMineral, AlterationStageZonation, AlterationType, MajorElement, TraceElement, EconomicElement, IsotopeElement, AnalysisMethod, AnalysisInstrument, GeologicTimescale, IsotopeAge, FluidInclusion, FluidSource, GeologicalStructure, GeologicalUnit, Author, Journal	subclass_of, is_a, has_object_property, formed_by, related_to

be used to extract information from the unstructured geoscience literature to construct geoscience knowledge graphs. However, it cannot fully represent the precise framework of a geoscience knowledge graph. The existing human-curated geoscience database is also a valuable legacy. These databases are organized in the form of relational data. The knowledge contained in the relational database is an important structured knowledge resource that can be used in many ways.

The existing geoscience databases can quickly produce the prototype of geoscience knowledge graphs through mapping from the relational data set to a knowledge graph, and provide basic knowledge and rules for further text mining of the unstructured geoscience literature. For example, because of the attraction of the early human-curated database of PBDB, Ré and Peters collaborated on text mining of the paleontology literature and upgraded DeepDive to PaleoDeepDive to extract fossil information from the geoscience literature. Due to the great success of DeepDive and its derivative versions, DeepDive has been updated to xDD, which is a digital library and cyberinfrastructure that facilitates the discovery and utilization of data and knowledge in published documents. Although xDD has a powerful ability to mine geoscience documents, success also depends on domain expert knowledge and the existing relational data. The geoscience dictionary and existing fossil information in the early PBDB database provide the supporting data and knowledge required for model training and testing in PaleoDeepdive. The innovation and knowledge discoveries based on xDD are inseparable from domain expert knowledge and scientific issues in the domain. Therefore, these human-curated databases should be considered when constructing geoscience knowledge graphs.

Although there are already open and online geoscience databases, the semantic gap between different domains and

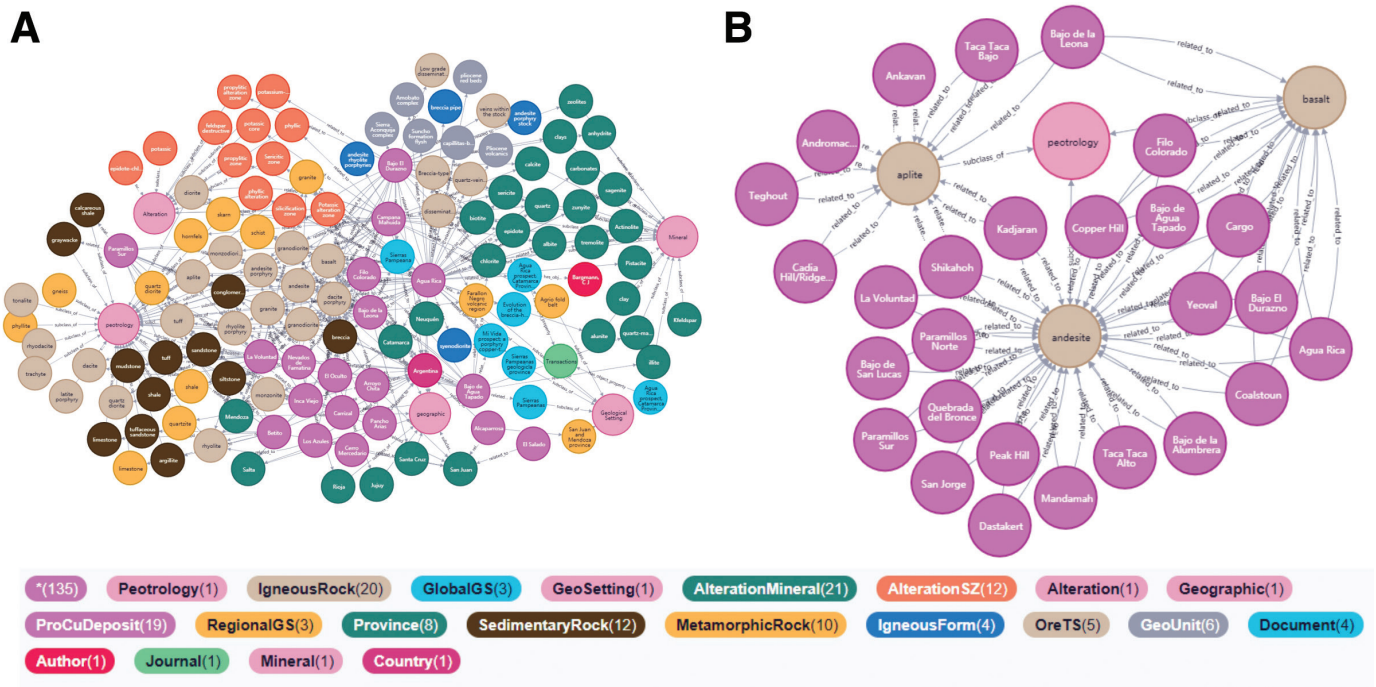

Figure 6. Diagrams show (A) the visualization of knowledge graph in the domain of porphyry copper deposits and (B) a query result of the relationship between porphyry copper deposits, petrology, and intrusive rock.

subdisciplines is not conducive to the deep integration of multisource and heterogeneous data. The lack of semantic information in the relational data and unstructured data cannot support the knowledge reasoning. Therefore, a broad geoscience knowledge graph is necessary for future geoscience cross-disciplinary research. However, it is an impossible task to construct a whole geoscience knowledge graph at one time. The feasible solution is to construct every subdiscipline knowledge graph and then integrate multidisciplinary knowledge graphs to build a broad geoscience graph based on wide-ranging geoscience ontology and knowledge fusion techniques. DDE, a big science program initiated by the International Union of Geological Sciences that is associated with deep-time Earth data, designated 19 subdiscipline working groups to develop its DDE Knowledge graph.

4.2. Domain-Specific Knowledge Models in the Geosciences

Although there are a lot of geoscience ontology models designed for different purposes, they are not suitable for constructing domain-specific knowledge graphs in geoscience directly. The scope of a geoscience knowledge graph is defined by a knowledge model. The knowledge graphs in the geoscience sub-disciplines have their topics and core schema, which results in differences in domain knowledge and terminology in different sub-disciplines. The ontology model in the knowledge model not only determines the geoscience entity tag in the text mining of geoscience literature, it also determines the semantic relations between entities in the construction of knowledge graph. In the process of constructing knowledge graph, some synonym relation mentions can be extracted by NLP and then need to be aligned to the predefined semantic relations in the ontology model. Therefore, it is necessary to build a broad and professional ontology model to standardize entity annotation and semantic relations and guide the construction of many knowledge graphs in the geosciences. The existing ontology models describe the knowledge system in certain geoscience fields, and a higher-level schema needs to be designed by integrating the existing ontology models to guide the construction of geoscience knowledge graphs.

4.3. Comparable Training Corpus

The published case studies (Table 2) show that they have excellent performance in entity recognition and relation extraction. However, the different tokens and training corpus results in these models are not comparable. The studies with fewer entity types performed better than those with more entity types. The studies using deep learning did not perform better than those using the machine learning model. In the field of NLP, there are some open corpora that include tokens and criteria for algorithm comparison. To promote the development of text data mining in the geosciences and enable text data mining to promote the development of geoscience knowledge graphs, it is necessary to build a series of corpora and criteria for different tasks of geoscience text mining.

4.4. Deep Knowledge Discovery

Multidisciplinary integration can promote knowledge discoveries in the geosciences. The Earth has undergone a lengthy evolution of more than 4.5 billion years, which results in unique deep-time properties in geoscience data. Some discoveries of Earth and life evolution are hidden in the heterogeneous and deep-time geoscience dark data set that covers many disciplines and is difficult to process in traditional ways (Soreghan, 2004; National Research Council, 2008; Wang et al., 2021). More and more studies have proven that multidisciplinary collaboration is best for exploring the evolution of the Earth and life. Moore et al. (2018) revealed the biodiversity in the Archean ocean based on the cobalt-bearing mineral ecosystem and chemical characteristics in vitamin B12. A high-resolution biodiversity evolutionary history from the Cambrian to Early Triassic was created by analyzing the Paleozoic big marine data set (Fan et al., 2020). Peters et al. (2017) revealed that the emergence of stromatolite had a close relation with dolomite rather than with mass extinction based on paleontology data extracted from the literature.

Although the above studies have proven that multidisciplinary integration and computer technology are beneficial for knowledge discovery in the geosciences, the heterogeneous and unstructured geoscience data sets with semantic gaps pose a great challenge for cross-disciplinary data mining for knowledge discoveries (Wang et al., 2018a). The geoscience knowledge graph is not only used to harmonize geoscience data and knowledge and visualize the geoscience knowledge. If we have the cyberinfrastructure of a broad geoscience knowledge graph, it can reduce the difficulty of conducting research similar to the examples given above. In this way, it can promote knowledge discoveries in the geosciences and deepen our understanding of the Earth's evolution.

5. CONCLUSIONS

For data-intensive geoscience fields, research in recent decades has proven knowledge graph to be a functional tool for processing massive and heterogeneous geoscience data. In this paper, we reviewed concepts and technologies relevant to knowledge graph and the state-of-the-art progress of geoscience knowledge graphs and then summarized the workflow to build a knowledge graph in the geosciences. Construction of a geoscience knowledge graph requires deep cooperation among geoscientists and computer scientists based on the technologies of NLP, machine learning, visualization, knowledge inference, relational data mapping, data fusion, and database. Although text mining based on NLP and deep learning methods is an important approach for building a knowledge graph, the human-curated database is also an important legacy that can quickly produce the prototype of knowledge geoscience graphs and provide basic knowledge and rules for further text mining of the unstructured geoscience literature. In the future, domain-specific knowledge models and comparable training corpus should be considered

to improve the efficiency and quality of geoscience knowledge graph construction.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (41902305), National Key R&D Program of China (2017YFC0601500 and 2017YFC0601504), Natural Science Foundation of Hubei Province (2019CFB231), and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUG190617).

REFERENCES CITED

Adam, S., and Schultz, U.P., 2015, Towards tool support for spreadsheet-based domain-specific languages, *in* Proceedings of the 2015 ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences: New York, Association for Computing Machinery, p. 95–98.

Adel, H., Oberländer, L.A.M., Papay, S., Padó, S., and Klinger, R., 2018, DERE: A task and domain-independent slot filling framework for declarative relation extraction, *in* Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, p. 42–47.

Amaral, D., Collovini, S., Figueira, A., Vieira, R., and Gonzalez, M., 2017, Building an annotated corpus with geological entities for NER, *in* Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology: Uberlândia, MG, Brazil, Sociedade Brasileira de Computação, p. 63–72 [in Portuguese].

Angeli, G., Tibshirani, J., Wu, J., and Manning, C.D., 2014, Combining distant and partial supervision for relation extraction, *in* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, p. 1556–1567.

Annervaz, K.M., Chowdhury, S.B.R., and Dukkipati, A., 2018, Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing: arXiv preprint arXiv:1802.05930; https://doi.org/10.48550/arXiv.1802.05930.

Babaie, H.A., Oldow, J.S., Babaei, A., Avé Lallemant, H.G., and Watkinson, A.J., 2006, Designing a modular architecture for the structural geology ontology, *in* Sinha, A.K., ed., Geoinformatics: Data to Knowledge: Geological Society of America Special Paper 397, p. 269–282, https://doi.org/10.1130/2006.2397(21).

Berners-Lee, T., 1998, Semantic web road map: https://www.w3.org/DesignIssues/Semantic.html (accessed 21 April 2022).

Berners-Lee, T., and Hendler, J., 2001, Publishing on the semantic web: Nature, v. 410, p. 1023–1024, https://doi.org/10.1038/35074206.

Berners-Lee, T., Hollenbach, J., Lu, K., Presbrey, J., Prud'hommeaux, E., and Schraefel, M.C., 2008, Tabulator redux: Browsing and writing linked data, *in* Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T., eds., Proceedings of the Linked Data on the Web Workshop, Beijing, China, v. 369, paper 12.

Bhasuran, B., Murugesan, G., Abdulkadhar, S., and Natarajan, J., 2016, Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases: Journal of Biomedical Informatics, v. 64, p. 1–9, https://doi.org/10.1016/j.jbi.2016.09.009.

Bristol, R.S., Euliss, N.H., Jr., Booth, N.L., Burkardt, N., Diffendorfer, J.E., Gesch, D.B., McCallum, B.E., Miller, D.M., Morman, S.A., Poore, B.S., Signell, R.P., and Viger, R.J., 2012, Science Strategy for Core Science Systems in the U.S. Geological Survey, 2013–2023: Public Review Release: U.S. Geological Survey Open-File Report 2012-1093, 29 p.

Callaway, E., 2015, TOOLBOX: Computers read the fossil record: Nature, v. 523, p. 115–116, https://doi.org/10.1038/523115a.

Chambers, N., and Jurafsky, D., 2011, Template-based information extraction without the templates, *in* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Portland, Oregon, p. 976–986.

Chen, H., and Luo, X., 2019, An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing: Advanced Engineering Informatics, v. 42, https://doi.org/10.1016/j.aei.2019.100959.

Chen, H., Lin, Z., Ding, G., Lou, J., Zhang, Y., and Karlsson, B., 2019, GRN: Gated relation network to enhance convolutional neural network for named entity recognition: Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, v. 33, p. 6236–6243, https://doi.org/10.1609/aaai.v33i01.33016236.

Chen, J., Ji, D., Tan, C.L., and Niu, Z.Y., 2006, Relation extraction using label propagation based semi-supervised learning, *in* Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, p. 129–136.

Chieu, H.L., and Ng, H.T., 2003, Named entity recognition with a maximum entropy approach: Proceedings of the Seventh Conference on Natural Language Learning at Human Language Technologies–Association for Computational Linguistics 2003, vol. 4, Edmonton, Canada, p. 160–163, https://doi.org/10.3115/1119176.1119199.

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., and Vaithyanathan, S., 2010, Domain adaptation of rule-based annotators for named-entity recognition tasks, *in* Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing: Cambridge, Massachusetts, Association for Computational Linguistics, p. 1002–1012.

Consoli, B., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., and Moreira, V., 2020, Embeddings for named entity recognition in geoscience Portuguese literature, *in* Proceedings of the 12th Language Resources and Evaluation Conference, Marseille: Paris, The European Language Resources Association, p. 4625–4630.

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: Geosphere, v. 1, p. 119–137, https://doi.org/10.1130/GES00022.1.

Cox, S.J.D., and Richard, S.M., 2015, A geologic timescale ontology and service: Earth Science Informatics, v. 8, p. 5–19, https://doi.org/10.1007/s12145-014-0170-6.

Das, P., Das, A.K., Nayak, J., Pelusi, D., and Ding, W., 2019, A graph based clustering approach for relation extraction from crime data: IEEE Access, v. 7, p. 101,269–101,282, https://doi.org/10.1109/ACCESS.2019.2929597.

Dhillon, I.S., Guan, Y., and Kogan, J., 2002, Iterative clustering of high dimensional text data augmented by local search, *in* Proceedings, 2002 Institute of Electrical and Electronics Engineers International Conference on Data Mining: Piscataway, New Jersey, IEEE, p. 131–138.

Duan, Y., Edwards, J.S., and Xu, M.X., 2005, Web-based expert systems: Benefits and challenges: Information & Management, v. 42, p. 799–811, https://doi.org/10.1016/j.im.2004.08.005.

Ehrlinger, L., and Wöß, W., 2016, Towards a definition of knowledge graphs, *in* Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems–SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16), co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany: CEUR Workshop Proceedings 1695.

Ekbal, A., and Bandyopadhyay, S., 2008, Bengali named entity recognition using support vector machine, *in* Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages: Hyderabad, India, Asian Federation of Natural Language Processing, p. 51–58.

Eftimov, T., Seljak, B.K., and Korošec, P., 2017, A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations: PloS one, v.12, e0179488, https://doi.org/10.1371/journal.pone.0179488.

Enkhsaikhan, M., 2021, Geological knowledge graph construction from mineral exploration text [Ph.D. thesis]: Perth, Australia, The University of Western Australia, 153 p.

Enkhsaikhan, M., Liu, W., Holden, E.J., and Duuring, P., 2021, Auto-labelling entities in low-resource text: A geological case study: Knowledge and Information Systems, v. 63, p. 695–715, https://doi.org/e10.1007/s10115-020-01532-6.

Euzenat, J., and Valtchev, P., 2004, Similarity-based ontology alignment in OWL-lite, *in* Proceedings, 16th European Conference on Artificial Intelligence (ECAI): Amsterdam, IOS Press, p. 333–337.

Fan, J.X., Shen, S.-z., Erwin, D.H., Sadler, P.M., MacLeod, N., Cheng, Q.-m., Hou, X.-d., Yang, J., Wang, X.-d., Wang, Y., Zhang, H., Chen, X., Li, G.-x., Zhang, Yi-c., Shi, Yu-k., Yuan, D.-x., Chen, Q., Zhang, L.-n., Li, C., and Zhao, Y.-y., 2020, A high-resolution summary of Cambrian to Early

Triassic marine invertebrate biodiversity: Science, v. 367, p. 272–277, https://doi.org/10.1126/science.aax4953.

Fang, Z., Cao, Y., Li, Q., Zhang, D., Zhang, Z., and Liu, Y., 2019, Joint entity linking with deep reinforcement learning, *in* Proceedings, The World Wide Web Conference, San Francisco: New York, Association for Computing Machinery, p. 438–447.

Feigenbaum, E., and Buchanan, B., 1993, DENDRAL and META-DENDRAL: Roots of knowledge systems and expert system applications: Artificial Intelligence, v. 59, p. 233–240, https://doi.org/10.1016/0004-3702(93)90191-D.

Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J.L., and Middleton, D., 2009, Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience: Computers & Geosciences, v. 35, p. 724–738, https://doi.org/10.1016/j.cageo.2007.12.019.

Gao, H., Huang, D., and Yang, Y., 2005a, Word-level Chinese named entity recognition based on segmentation digraph, *in* Proceedings, 2005 International Conference on Natural Language Processing and Knowledge Engineering: Piscataway, New Jersey, Institute of Electrical and Electronics Engineers, p. 380–383.

Gao, J., Li, M., Huang, C.N., and Wu, A., 2005b, Chinese word segmentation and named entity recognition: A pragmatic approach: Computational Linguistics, v. 31, p. 531–574, https://doi.org/10.1162/089120105775299177.

Garcia, L.F., Abel, M., Perrin, M., and dos Santos Alvarenga, R., 2020, The GeoCore ontology: A core ontology for general use in Geology: Computers & Geosciences, v. 135, https://doi.org/10.1016/j.cageo.2019.104387.

Gaschnig, J., 1982, Prospector: An expert system for mineral exploration, *in* Michie, D., ed., Introductory Readings in Expert Systems: New York, Gordon and Breach, p. 47–64.

Gebretensae, N., 2019, Wikidata: A Free Collaborative Knowledge Graph: http://knowledge-graphs-seminar.s3.amazonaws.com/wikidata.pdf (accessed 31 August 2022).

Gil, Y., Pierce, S.A., Babaie, H., and 30 others, 2019, Intelligent systems for geosciences: An essential research agenda: Communications of the ACM, v. 62, p. 76–84, https://doi.org/10.1145/3192335.

Gruber, T.R., 1995, Toward principles for the design of ontologies used for knowledge sharing?: International Journal of Human-Computer Studies, v. 43, p. 907–928, https://doi.org/10.1006/ijhc.1995.1081.

Gusenbauer, M., 2019, Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases: Cham, Switzerland, Springer International Publishing, v. 118, p. 177–214, https://doi.org/10.1007/s11192-018-2958-5.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y., 2006, A closer look at skip-gram modelling, *in* Proceedings of the International Conference on Language Resources and Evaluation, v. 6, p. 1222–1225, http://www.lrec-conf.org/proceedings/lrec2006/pdf/357_pdf.pdf.

Hendler, J., McGuinness, D.L., et al., 2000, The DARPA agent markup language: Institute of Electrical and Electronics Engineers Intelligent Systems, v. 15, p. 67–73.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., and 12 others, 2022, Knowledge graphs [CSUR]: Association for Computing Machinery Computing Surveys, v. 54, p. 1–37, https://doi.org/10.1145/3447772.

Holden, E.J., Liu, W., Horrocks, T., Wang, R., Wedge, D., Duuring, P., and Beardsmore, T., 2019, GeoDocA—Fast analysis of geological content in mineral exploration reports: A text mining approach: Ore Geology Reviews, v. 111, 102919, https://doi.org/10.1016/j.oregeorev.2019.05.005.

Hotho, A., Nürnberger, A., and Paaß, G., 2005, A brief survey of text mining: LDV Forum, v. 20, no. 1, p. 19–62.

Hu, Y., Huber, A., Anumula, J., and Liu, S.C., 2018, Overcoming the vanishing gradient problem in plain recurrent networks: arXiv:1801.06105, p. 1–20.

Huang, L., Du, Y., and Chen, G., 2015, GeoSegmenter: A statistically learned Chinese word segmenter for the geoscience domain: Computers & Geosciences, v. 76, p. 11–17, https://doi.org/10.1016/j.cageo.2014.11.005.

Husson, J.M., and Peters, S.E., 2017, Atmospheric oxygenation driven by unsteady growth of the continental sedimentary reservoir: Earth and Planetary Science Letters, v. 460, p. 68–75, https://doi.org/10.1016/j.epsl.2016.12.012.

Husson, J.M., and Peters, S.E., 2018, Nature of the sedimentary rock record and its implications for Earth system evolution, *in* Lyons, T.W., Droser, M.L., Lau, K.V., and Porter, S.M., eds., Early Earth and the Rise of Complex Life: Emerging Topics in Life Sciences, v. 2, p. 125–136, https://doi.org/10.1042/ETLS20170152.

Husson, J.M., Peters, S.E., Ross, I., and Czaplewski, J.J., 2016, Macrostrat and GeoDeepDive: A platform for geological data integration and deep-time research: Abstract IN23F-04 presented at 2016 Fall Meeting, American Geophysical Union, San Francisco, California, 12–16 December.

Jacobson, M., Charlson, R.J., Rodhe, H., and Orians, G.H., 2000, Earth System Science: From Biogeochemical Cycles to Global Changes: London, Elsevier Academic Press, 550 p.

Jatana, N., Puri, S., Ahuja, M., Kathuria, I., and Gosain, D., 2012, A survey and comparison of relational and non-relational database: International Journal of Engineering Research & Technology, v. 1, p. 1–5.

Jepsen, T.C., 2009, Just what is an ontology, anyway?: IT Professional Magazine, v. 11, p. 22–27.

Jiang, J., 2009, Multi-task transfer learning for weakly-supervised relation extraction, *in* Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09), Singapore, p. 2–7.

Johnson, J.M., and Khoshgoftaar, T.M., 2020, Semantic embeddings for medical providers and fraud detection, *in* Proceedings, Institute of Electrical and Electronics Engineers 2020 21st International Conference on Information Reuse and Integration for Data Science (IRI): Piscataway, New Jersey, Institute of Electrical and Electronics Engineers, p. 224–230.

Karystianis, G., Thayer, K., Wolfe, M., and Tsafnat, G., 2017, Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews: Journal of Biomedical Informatics, v. 70, p. 27–34, https://doi.org/10.1016/j.jbi.2017.04.004.

Khabsa, M., and Giles, C.L., 2014, The number of scholarly documents on the public web: PLoS One, v. 9, https://doi.org/10.1371/journal.pone.0093949.

Kim, J.H., and Woodland, P.C., 2000, A rule-based named entity recognition system for speech input: Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000), v. 1, p. 528–531

Lambrix, P., Habbouche, M., and Perez, M., 2003, Evaluation of ontology development tools for bioinformatics: Bioinformatics, v. 19, p. 1564–1571, https://doi.org/10.1093/bioinformatics/btg194.

Lee, J.Y., Dernoncourt, F., and Szolovits, P., 2017, Transfer learning for named-entity recognition with neural networks: arXiv preprint arXiv:1705.06273, https://doi.org/10.48550/arXiv.1705.06273.

Li, G., Yang, X., Ye, T., Sun, H., Tang, X., and Han, B., 2010, Design and implementation of ontology-based knowledge base system for marine hydrocarbon geology: Journal of Computer Applications, v. 30, p. 532–536, https://doi.org/10.3724/SP.J.1087.2010.00532.

Li, L., Zhou, R., and Huang, D., 2009, Two-phase biomedical named entity recognition using CRFs: Computational Biology and Chemistry, v. 33, p. 334–338, https://doi.org/10.1016/j.compbiolchem.2009.07.004.

Li, P.-H., Dong, R.-P., Wang, Y.-S., Chou, J.-C., and Ma, W.-Y., 2017, Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks, *in* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: Copenhagen, Denmark, Association for Computational Linguistics, p. 2664–2669.

Li, S., Chen, J., and Xiang, J., 2018, Prospecting information extraction by text mining based on convolutional neural networks—A case study of the Lala copper deposit, China: IEEE access, v.6, p. 52,286–52,297.

Li, Z., Yang, C., Jin, B., Yu, M., Liu, K., Sun, M., and Zhan, M., 2015, Enabling big geoscience data analytics with a cloud-based, map-reduce-enabled and service-oriented workflow framework: PLoS One, v. 10, no. 3, https://doi.org/10.1371/journal.pone.0116781.

Liao, S.-H., 2005, Expert system methodologies and applications—A decade review from 1995 to 2004: Expert Systems with Applications, v. 28, p. 93–103, https://doi.org/10.1016/j.eswa.2004.08.003.

Liu, G., Wang, Y., and Wu, C., 2010, Research and application of geological hazard domain ontology, *in* Proceedings, 2010 18th International Conference on Geoinformatics, Beijing: Piscataway, New Jersey, Institute of Electrical and Electronics Engineers, p. 1–6.

Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., and Xu, H., 2017, Entity recognition from clinical texts via recurrent neural network: BMC Medical Informatics and Decision Making, v. 17, 67, https://doi.org/10.1186/s12911-017-0468-7.

Luo, X., Zhou, W., Wang, W., Zhu, Y., and Deng, J., 2017, Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data: IEEE Access, v. 6, p. 5705–5715, https://doi.org/10.1109/ACCESS.2017.2785229.

Ma, X., 2021, Knowledge graph construction and application in geosciences: A review: Computers & Geosciences, 105082, https://doi.org/10.1016/j.cageo.2022.105082.

Ma, X., Carranza, E.J.M., Wu, C., Van Der Meer, F.D., and Liu, G., 2011, A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps: Computers & Geosciences, v. 37, p. 1602–1615, https://doi.org/10.1016/j.cageo.2011.02.011.

Ma, X., Carranza, E.J.M., Wu, C., and Van der Meer, F.D., 2012, Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services: Computers & Geosciences, v. 40, p. 107–119, https://doi.org/10.1016/j.cageo.2011.07.018.

Ma, X., Zheng, J.G., Goldstein, J.C., Zednik, S., Fu, L., Duggan, B., Aulenbach, S.M., West, P., Tilmes, C., and Fox, P., 2014, Ontology engineering in provenance enablement for the National Climate Assessment: Environmental Modelling & Software, v. 61, p. 191–205, https://doi.org/10.1016/j.envsoft.2014.08.002.

Ma, X., Ma, C., and Wang, C., 2020, A new structure for representing and tracking version information in a deep time knowledge graph: Computers & Geosciences, v. 145, https://doi.org/10.1016/j.cageo.2020.104620.

Madani, A., Boussaid, O., and Zegour, D.E., 2013, Semi-structured documents mining: A review and comparison: Procedia Computer Science, v. 22, p. 330–339, https://doi.org/10.1016/j.procs.2013.09.110.

Mantovani, A., Piana, F., and Lombardo, V., 2020, Ontology-driven representation of knowledge for geological maps: Computers & Geosciences, v. 139, https://doi.org/10.1016/j.cageo.2020.104446.

McGuinness, D.L., Fikes, R., Hendler, J., and Stein, L.A., 2002, DAML+ OIL: An ontology language for the Semantic Web: IEEE Intelligent Systems, v. 17, p. 72–80, https://doi.org/10.1109/MIS.2002.1039835.

Mentes, H.S., 2012, Design and development of a mineral exploration ontology [M.S. thesis]: Atlanta, Georgia State University, 159 p.

Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013, Efficient estimation of word representations in vector space: arXiv:1301.3781, http://arxiv.org/abs/1301.3781.

Moore, E.K., Hao, J., Prabhu, A., Zhong, H., Jelen, B.I., Meyer, M., Hazen, R.M., and Falkowski, P.G., 2018, Geological and chemical factors that impacted the biological utilization of cobalt in the Archean eon: Journal of Geophysical Research: Biogeosciences, v. 123, p. 743–759, https://doi.org/10.1002/2017JG004067.

Morwal, S., Jahan, N., and Chopra, D., 2012, Named entity recognition using hidden Markov model (HMM) [IJNLC]: International Journal on Natural Language Computing, v. 1, p. 15–23, https://doi.org/10.5121/ijnlc.2012.1402.

Nardi, D., and Brachman, R.J., 2010, An introduction to description logics, *in* Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., eds., The Description Logic Handbook: Theory, Implementation and Applications: Cambridge, UK, Cambridge University Press, p. 1–44, https://doi.org/10.1017/CBO9780511711787.003.

National Research Council, 2008, Origin and Evolution of Earth: Research Questions for a Changing Planet: Washington, D.C., The National Academies Press, 150 p., https://doi.org/10.17226/12161.

Niu, X., Rong, S., Wang, H., and Yu, Y., 2012, An effective rule miner for instance matching in a web of data, *in* Proceedings, 21st Association for Computing Machinery International Conference on Information and Knowledge Management: New York, ACM, p. 1085–1094.

Nurdiati, S., and Hoede, C., 2008, 25 years development of knowledge graph theory: The results and the challenge (Memorandum; No. 2/1876): Enschede, University of Twente, 10 p.

Pawar, S., Bhattacharyya, P., and Palshikar, G., 2017, End-to-end relation extraction using neural networks and Markov logic networks, *in* Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, p. 818–827.

Peters, S.E., and McClennen, M., 2016, The Paleobiology Database application programming interface: Paleobiology, v. 42, p. 1–7, https://doi.org/10.1017/pab.2015.39.

Peters, S.E., Zhang, C., Livny, M., and Ré, C., 2014, A machine reading system for assembling synthetic paleontological databases: PLoS One, v. 9, https://doi.org/10.1371/journal.pone.0113523.

Peters, S.E., Husson, J.M., and Wilcots, J., 2017, The rise and fall of stromatolites in shallow marine environments: Geology, v. 45, p. 487–490, https://doi.org/10.1130/G38931.1.

Peters, S.E., Husson, J.M., and Czaplewski, J., 2018, Macrostrat: A platform for geological data integration and deep-time earth crust research: Geochemistry, Geophysics, Geosystems, v. 19, p. 1393–1409, https://doi.org/10.1029/2018GC007467.

Qiu, Q., Xie, Z., Wu, L., and Li, W., 2018, DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain: Computers & Geosciences, v. 121, p. 1–11, https://doi.org/10.1016/j.cageo.2018.08.006.

Qiu, Q., Xie, Z., Wu, L., and Tao, L., 2019, GNER: A generative model for geological named entity recognition without labeled data using deep learning: Earth and Space Science, v. 6, p. 931–946, https://doi.org/10.1029/2019EA000610.

Qiu, Q., Xie, Z., Wu, L., and Tao, L., 2020, Dictionary-based automated information extraction from geological documents using a deep learning algorithm: Earth and Space Science, v. 7, https://doi.org/10.1029/2019EA000993.

Qu, J., Ouyang, D., Hua, W., Ye, Y., and Li, X., 2018, Distant supervision for neural relation extraction integrated with word attention and property features: Neural Networks, v.100, p. 59–69, https://doi.org/10.1016/j.neunet.2018.01.006.

Quillan, R., 1963, A Notation for Representing Conceptual Information: An Application to Semantics and Mechanical English Paraphrasing: Santa Monica, California, Systems Development Corporation, 59 p.

Rahm, E., Do, H.H., and Massmann, S., 2004, Matching large XML schemas: SIGMOD Record, v. 33, p. 26–31, https://doi.org/10.1145/1041410.1041415.

Raskin, R.G., and Pan, M.J., 2005, Knowledge representation in the semantic web for Earth and environmental terminology (SWEET): Computers & Geosciences, v. 31, p. 1119–1125, https://doi.org/10.1016/j.cageo.2004.12.004.

Roche, C., 2003, Ontology: A survey: International Federation of Automatic Control Proceedings Volumes, v. 36, p. 187–192.

Rueda, C., Bermudez, L., and Fredericks, J., 2009, The MMI Ontology Registry and Repository: A portal for Marine Metadata Interoperability: Oceans, v. 2009, p. 1–6, https://doi.org/10.23919/OCEANS.2009.5422206.

Schneider, E.W., 1973, Course Modularization Applied: The Interface System and Its Implications for Sequence Control and Data Analysis: Alexandria, Virginia, Human Resources Research Association, 22 p.

Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., and Anandkumar, A., 2017, Deep active learning for named entity recognition: arXiv preprint arXiv:1707.05928, https://doi.org/10.48550/arXiv.1707.05928.

Sillitoe, R.H., 1972, A plate tectonic model for the origin of porphyry copper deposits: Economic Geology, v. 67, p. 184–197, https://doi.org/10.2113/gsecongeo.67.2.184.

Sillitoe, R.H., 2010, Porphyry copper systems: Economic Geology, v. 105, p. 3–41, https://doi.org/10.2113/gsecongeo.105.1.3.

Singer, D.A., Berger, V.I., Menzie, W.D., and Berger, B.R., 2005, Porphyry copper deposit density: Economic Geology, v. 100, p. 491–514, https://doi.org/10.2113/gsecongeo.100.3.491.

Singhal, A., 2012, Introducing the knowledge graph: Things, not strings: Official Google Blog, v. 5, p. 16, https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed 21 April 2022).

Sint, R., Schaffert, S., Stroka, S., and Ferstl, R., 2009, Combining unstructured, fully structured and semi-structured information in semantic wikis: CEUR Workshop Proceedings, v. 464, p. 73–87.

Sobhana, N., Mitra, P., and Ghosh, S.K., 2010, Conditional random field based named entity recognition in geological text: International Journal of Computers and Applications, v. 1, p. 143–147, https://doi.org/10.5120/72-166.

Soreghan, G.S., 2004, GeoSystems: Probing climate and linked systems of Earth's deep-time dark ages: Abstract H54A-05, presented at 2004 Fall Meeting, American Geophysical Union, San Francisco, California 13–17 December 2004.

Suchanek, F.M., Abiteboul, S., and Senellart, P., 2011, PARIS: Probabilistic alignment of relations, instances, and schema: arXiv preprint arXiv:1111.7164, https://doi.org/10.48550/arXiv.1111.7164.

Sun, A., Grishman, R., and Sekine, S., 2011, Semi-supervised relation extraction with large-scale word clustering, *in* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, p. 521–529.

Szarvas, G., Farkas, R., and Kocsor, A., 2006, A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms, *in* Todorovski, L., Lavrač, N., and Jantke, K.P., eds., Discovery Science. DS 2006. Lecture Notes in Computer Science, Volume 4265: Berlin, Springer, https://doi.org/10.1007/11893318_27.

Tekli, J., 2016, An overview on XML semantic disambiguation from unstructured text to semi-structured data: Background, applications, and ongoing challenges: IEEE Transactions on Knowledge and Data Engineering, v. 28, p. 1383–1407, https://doi.org/10.1109/TKDE.2016.2525768.

Varela, S., González-Hernández, J., Sgarbi, L.F., Marshall, C., Uhen, M.D., Peters, S., and Mcclennen, M., 2015, paleobioDB: An R package for downloading, visualizing and processing data from the Paleobiology Database: Ecography, v. 38, p. 419–425, https://doi.org/10.1111/ecog.01154.

Wang, C., Ma, X., and Chen, J., 2018a, Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information: Computers & Geosciences, v. 115, p. 12–19, https://doi.org/10.1016/j.cageo.2018.03.004.

Wang, C., Ma, X., and Chen, J., 2018b, The application of data pre-processing technology in the geoscience big data: Yanshi Xuebao, v. 34, p. 303–313.

Wang, C., Ma, X., Chen, J., and Chen, J., 2018c, Information extraction and knowledge graph construction from geoscience literature: Computers & Geosciences, v. 112, p. 112–120, https://doi.org/10.1016/j.cageo.2017.12.007.

Wang, C., Hazen, R.M., Cheng, Q., Stephenson, M.H., Zhou, C., Fox, P., Shen, S., Oberhänsli, R., Hou, Z., Ma, X., Feng, Z., Fan, J., Ma, C., Hu, X., Luo, B., and Wang, J., 2021, The Deep-Time Digital Earth program: Data-driven discovery in geosciences: National Science Review, v. 8, https://doi.org/10.1093/nsr/nwab027.

Wang, H., Zhao, M., Xie, X., Li, W., and Guo, M., 2019, Knowledge graph convolutional networks for recommender systems, *in* Proceedings of the 2019 World Wide Web Conference (WWW '19), San Francisco, California, 13–17 May: New York, Association for Computing Machines, p. 3307–3313.

Zhang, C., 2015, DeepDive: A data management system for automatic knowledge base construction [Ph.D. thesis]: Madison, Wisconsin, University of Wisconsin–Madison, 193 p.

Zhang, H., Guo, Y., and Li, T., 2019, Multifeature named entity recognition in information security based on adversarial learning: Security and Communication Networks, v. 2019, https://doi.org/10.1155/2019/6417407.

Zhang, S., Qin, Y., Hou, W.J., and Wang, X., 2006, Word segmentation and named entity recognition for SIGHAN Bakeoff3, *in* Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney: Stroudsburg, Pennsylvania, Association for Computational Linguistics, p. 158–161.

Zhang, S., Elha, N., and Da, D., 2013, Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts: Journal of Biomedical Informatics, v. 46, no. 6, p. 1088–1098, https://doi.org/10.1016/j.jbi.2013.08.004.

Zhang, X.Y., Ye, P., Wang, S., and Du, M., 2018, Geological entity recognition method based on Deep Belief Networks: Acta Petrologica Sinica, v. 34, p. 343–351 [in Chinese with English abstract].

Zhong, J., Aydina, A., and McGuinness, D.L., 2009, Ontology of fractures: Journal of Structural Geology, v. 31, p. 251–259, https://doi.org/10.1016/j.jsg.2009.01.008.

Zhong, Y., Liu, X., Wang, J., Chen, Y., and Zhang, T., 2020, Research of extraction on petroleum unstructured information based on named entity recognition: Journal of Southwest Petroleum University (Science & Technology Edition), v. 42, p. 165–173 [in Chinese with English abstract].

Zhou, G., Qian, L., and Fan, J., 2010, Tree kernel-based semantic relation extraction with rich syntactic and semantic information: Information Sciences, v. 180, p. 1313–1325, https://dl.acm.org/doi/abs/10.1016/j.ins.2009.12.006.

Zhou, X.G., Gong, R.-B., Shi, F.G., and Wang, Z.F., 2020, PetroKG: Construction and application of knowledge graph in upstream area of PetroChina: Journal of Computer Science and Technology, v. 35, p. 368–378, https://doi.org/10.1007/s11390-020-9966-7.

Zhu, Y., Zhou, W., Xu, Y., Liu, J., and Tan, Y., 2017, Intelligent learning for knowledge graph towards geological data: Scientific Programming, v. 2017, https://doi.org/10.1155/2017/5072427.

Manuscript Accepted by the Society 17 March 2022
Manuscript Published Online 30 November 2022

Printed in the USA