Geoscience Frontiers xxx (xxxx) xxx

Contents lists available at ScienceDirect

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

A knowledge graph and service for regional geologic time standards

Chao Ma [a,b], Amruta Suresh Kale [a], Jiyin Zhang [a], Xiaogang Ma [a,*]

[a] Department of Computer Science, University of Idaho, 875 Perimeter Drive, MS 1010, Moscow, ID 83844-1010, USA
[b] State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, Chengdu 610059, China

ARTICLE INFO

Article history:
Received 14 December 2021
Revised 11 July 2022
Accepted 6 August 2022
Available online xxxx

Keywords:
Knowledge graph
Regional geologic time standard
Ontology
Semantic technology

ABSTRACT

Geologic time is an important dimension in geological research. Geologic time data are commonly collected from multiple sources in data-intensive studies of Earth's history and raise an issue of data cleansing and integration. A knowledge graph of the international geological time scale has been established to harmonize heterogeneous data to facilitate effective and efficient data-driven discovery. Although many regional geologic time standards are also used in various databases and literature, there is limited discussion or development of knowledge graph for them. In this research, we construct a knowledge graph for the geologic time standards in 17 regions at the Epoch and Age levels. This regional geologic time knowledge graph is integrated with the international geologic time knowledge graph as a comprehensive deep-time knowledge base. A SPARQL endpoint has been established to provide open and free online service to the knowledge base. Several use cases are presented here to demonstrate the functionality of the knowledge graph we built as well as its application in open data exploration. Our work addresses the shortage of machine-readable knowledge graphs for regional geologic time standards and will help accelerate geologic data integration from multiple sources in data-intensive studies. All data and code in this paper are made open source and are accessible on GitHub and Zenodo.

1. Introduction

The co-evolution of the geosphere and the biosphere has been recognized as one of the grand research questions for the 21st century Earth science (NRC, 2008; NASEM, 2020). Exploring the positive and negative feedbacks between the living and non-living components in the Earth's long history will lead to many innovative scientific discoveries. The exploration needs data from multiple disciplines such as mineralogy, paleobiology, petrology, geochronology, geochemistry, proteomics, and more. In the past decades, many open data facilities have been built. However, researchers of co-evolution are facing the shortage of automated and efficient methods to synthesize the datasets from multiple facilities. Geologic time is an essential topic in the co-evolution study and can be used as a central axis to connect various data silos. Yet in many of the existing data facilities, heterogeneous terminologies of geologic time are used, which are not only hard for scientists to understand but also a hurdle for data integration.

Ontologies and vocabularies have been proven as an effective way to address conceptual heterogeneity and facilitate data interoperability (Gil et al., 2019). Key concepts and relationships in geologic time have been represented in machine-readable ontologies (Cox and Richard, 2005, 2015; Cox, 2011; Perrin et al., 2011; Ma and Fox, 2013), and vocabularies for the international geologic time scale have been built and shared as services on the Web (Ma et al., 2014; Cox et al., 2016; Wang et al., 2018; Ma et al., 2020b). In recent years, the term "knowledge graph" has been increasingly used by scientists to represent their work of conceptual models, ontologies, and vocabularies (Sheth et al., 2019; Hogan et al., 2022). Hogan et al. (2022) defined a knowledge graph as a graph of data that uses nodes to represent entities of interest in the real world and edges to represent relationships between entities. In this paper, we use the term "deep time knowledge base" as a general concept to represent the work of knowledge graphs, ontologies, vocabularies, and other models for the geologic time scale. Where necessary, we will also use specific terms such as class, property, ontology, and vocabulary to specify the technological approach in detail.

Our review of the above-mentioned literature shows that most existing work of knowledge graphs and models of deep time focuses on the international geologic time standards, while limited studies explored the regional standards. In fact, in real-world geological studies, those regional standards can be seen in many databases and publications, such as the regional geologic time scales for North America, Russia, West Europe, Britain, China,

* Corresponding author.
E-mail address: max@uidaho.edu (X. Ma).

https://doi.org/10.1016/j.gsf.2022.101453
1674-9871/

Japan, Australia, New Zealand, and more (Haq and van Eysinga, 2007). In multi-disciplinary and fine-scale studies, such as topics in the co-evolution, geoscience datasets of various regions will be collected. To establish efficient workflows of data integration among those datasets, it is necessary to construct a knowledge graph for those regional geologic time standards and build service of it for a machine to access.

In this study, we construct a knowledge graph of regional geologic time standards to (1) enhance the interoperability of data that contain those regional standards and (2) facilitate efficient data synthesis and data-driven discovery in the co-evolving geosphere and biosphere. The research will benefit automating geoscience data access and integration in the open data environment and will support executable workflows for the data-intensive co-evolution research. The remainder of the paper is organized as follows. Section 2 presents the structure of the developed knowledge graph. Section 3 describes the construction of the knowledge graph and its online services. Section 4 discusses the contributions of this research, and Section 5 gives a brief conclusion.

2. Design a knowledge graph structure for regional geologic time standards

The international geological time scale has been established and published by the International Commission on Stratigraphy (ICS; stratigraphy.org). A representative standard is the International (Chrono) Stratigraphic Chart (ISC), which presents a hierarchal and ordinal structure (Cox and Richard, 2005, 2015; Michalak, 2005). There are two top classes of geologic time concepts within the structure, i.e., interval and instant. For example, Jurassic is an instance of interval and the base boundary of Jurassic is an instance of instant. Geologic time intervals are divided into different levels, which from highest to lowest order are Supereon, Eon, Era, Period, Epoch, and Age (Gradstein et al., 2020). The regional geologic time standards are only limited to the level of Epoch and Age. There are four main differences between international and regional standards: (1) Different interval names; (2) The same name but different definition of boundaries; (3) The international standard covers whole geological history while a regional standard may only cover a part of it; and (4) Due to differences in boundaries, a regional Epoch or Age interval may stand across two higher-level intervals in the international geologic time scale.

There are many detailed regional geologic time scales around the world. The varied disciplinary focuses and different levels of details in those time scales generate extremely heterogeneous terminology. To have a focus, we began the work by collecting regional geologic time intervals from the Time Scale Creator (Ogg and Lugowski, 2020), and we selected 17 regional geologic time standards in our first round of vocabulary construction. In our investigation, we found a few regions with geologic time intervals at the Epoch level: Iberian-Morocco, East Avalonian, Russia Platform, West Europe, Baltoscania, South China, and New Zealand. Also, the following regions have geologic time intervals at the Age level: North America, Boreal, California, Iberian-Morocco, Russia Platform, N–E Siberia, Kazakhstan, Tethyan, West Europe, British, Baltoscania, South China, North China, Japan, Australia, and New Zealand. The time range of the investigated regional standards only focuses on Phanerozoic.

2.1. Existing ontologies and vocabularies

We adopted well-established semantic web standards including RDF (Resource Description Framework) and RDFS (RDF Schema) that provide fundamental building blocks to construct the knowledge graph of regional geologic time standards. Other ontologies and vocabularies utilized in our study for deep time knowledge are dc, dcterms, void, gts, isc, skos, time, and ts. Table 1 gives a list of the core ontologies, schemas, and vocabularies in the current knowledge graph and briefly describes their roles. The source code and details of them are hosted on a GitHub Repo maintained by Dr. Simon Cox (Cox, 2020). The knowledge graph of international geologic time standards (Cox and Richard, 2005, 2015; Cox, 2011) is also used in our knowledge graph.

2.2. Design of the knowledge graph for regional geologic time standards

Each identified region is assigned a unique namespace prefix in our knowledge graph (Table 2). The namespace prefixes are also used in the identifiers of corresponding knowledge graphs. The structure is shown in Figs. 1 and 2. For example, the namespace prefix of regional geologic time standard in North America is "tsna". It is a concept collection with the label of "Geologic Timescale Elements in North America" in English. We then define "tsna2019" as a vocabulary scheme for that standard. The suffix "2019" is used here for version control, in case the vocabulary scheme will be updated in the future.

Under the vocabulary scheme of each region, there is a list of geologic time intervals and instants. Each interval is defined with (1) type of the interval (rdf:type); (2) label of the interval (rdfs:label); (3) a similar interval online (rdfs:seeAlso); (4) the broader interval of the current interval (skos:broader, e.g. an interval that is one level higher than the current interval); (5) skos:broaderTransitive (all intervals that are more than one level higher the current interval); (6) concept scheme (skos:inScheme) that denotes the interval's region and version; (7) the bottom boundary of the interval (time:hasBeginning); and (8) the top boundary of the interval (time:hasEnd) (Figs. 3 and 4). Following ontology patterns in the knowledge graph of the international geologic time scale (Cox, 2020), for the last two properties (7 and 8) we define the instances of instant and the associated properties. In the example shown in Figs. 3 and 4, the top and bottom boundaries of the Ochoan Epoch are each an instant, and they both have properties to show the numerical values and units in the time reference system. In the

Table 1
Core ontologies, schemas and vocabularies in the existing deep-time knowledge graph (Cox and Richard, 2015; Cox et al., 2016).

Prefix	Namespace	Role in the deep time knowledge graph
dc	<https://purl.org/dc/elements/1.1/>	Specify metadata of vocabulary schemes and time intervals and instants
dcterms	<https://purl.org/dc/terms/>	Specify metadata of vocabulary schemes and time intervals and instants
gts	<https://resource.geosciml.org/ontology/timescale/gts#>	Based on THORS and ISO 19156; Specify the structure of core classes and relationships in the geologic time scale
isc	<https://resource.geosciml.org/classifier/ics/ischart/>	Specify the deep time intervals and instants in the ISC charts
skos	<https://www.w3.org/2004/02/skos/core#>	Specify hierarchical structure and multilingual labels of deep time intervals and instants
time	<https://www.w3.org/2006/time#>	Specify the reference system and topological relationships of deep time intervals and instants
ts	<https://resource.geosciml.org/vocabulary/timescale/>	Specify the different versions of vocabulary schemes for the ISC charts

Table 2
Namespace prefix for regional geologic time standards.

Region	Namespace prefix	Region	Namespace prefix
North America	tsna	West Europe	tswe
East Avalonian	tsav	British	tsbr
California	tsca	Baltoscania	tsba
Iberian-Morocco	tsibmo	South China	tssc
Boreal	tsbo	North China	tsnc
Russia Platform	tsru	Japan	tsjp
N–E Siberia	tssi	Australia	tsau
Kazakhstan	tska	New Zealand	tsnz
Tethyan	tste		

ordinal and hierarchical structure of a geologic time scale, the intervals share boundaries: for two adjacent intervals (e.g. Ochoan Epoch and Guadalupian Epoch), the bottom boundary of the later interval in time (Ochoan Epoch) is always the top boundary of the earlier interval (Guadalupian Epoch). This pattern enables us only to define the shared boundary once to reduce redundancy in the resulting knowledge graph. For example, the time:hasEnd of Guadalupian Epoch is not tsna:TopGuadalupian but tsna:BaseOchoan, because they share this boundary, and there is no need to create a redundant boundary record. In regional geologic time standards, sometimes there are gaps between intervals. In this case, we define each boundary separately. For example, the top of the Ochoan Epoch is not shared with any other interval so we define it separately (Figs. 3 and 4).

The property skos:broader is used to define the lower-to-higher link between two hierarchical intervals. This is used in our knowledge graph to define the relationship between a geologic interval and its parent interval that is one level higher. For example: "tsna:Ochoan skos:broader isc:Permian" (Figs. 3 and 4). The lower-level interval must be at the left of "skos:broader". For more than one level of higher hierarchical intervals, we use skos:broaderTransitive. For example: "tsna:Ochoan skos:broaderTransitive isc:Paleozoic" and "skos:broaderTransitive isc:Phanerozoic" (Figs. 3 and 4). Regional geologic time intervals are only at the Epoch and Age levels. Thus, the broader interval of a regional Epoch-level interval is a Period interval in the international geologic time scale, and the broader transitive intervals are at the levels of Eon and Era. For a regional Age-level interval, there are two possibilities: (1) If there is a corresponding Epoch interval in the same region, then use skos:broader as the relationship; (2) If not, then find a corresponding Epoch interval in the international geologic time scale and use skos:broader to connect them.

The above-mentioned properties skos:broader and skos:broaderTransitive are both from the SKOS (Simple Knowledge Organization System) schema (Miles and Bechhofer, 2009). As reflected in its name, the properties in the schema are for a simple and lightweight description of the properties of concepts and the inter-relationships between those concepts. In our work, they can help present a quick overview of the hierarchical structure among the geologic time intervals, such as visualizing a hierarchical structure among time intervals. Nevertheless, the meaning of "broader" and "broader transitive" are too vague to support precise reasoning and inference, and they are not recommended for use in scientific explorations. Instead, in the developed knowledge graph there are classes and properties reused from other ontologies and schemas to provide more precise and meaningful capability of reasoning. Those ontologies and schemas set up the framework of strict and logic constraints on reasoning and inference among the geologic time intervals and instants. For example, Figs. 3 and 4 show the triples "tsna:Ochoan rdf:type gts:Epoch" and "tsna:Ochoan rdf:type time:ProperInterval". In the gts ontology and the Time Ontology (Figs. 5 and 6), there are detailed definitions of those two classes in the frameworks of geologic time scale and temporal topology, respectively. These can lead to interesting use cases. For example, the knowledge graph can immediately tell a user that "isc:Ochoan" is not within "isc:Changhsingian" because the former is an instance of "gts:Epoch" and the latter an instance of "gts:Age", and in the definition of "gts:Epoch", a "gts:Epoch" interval can be only within exactly-one "gts:Period" or "gts:SubPeriod" interval. The numerical information in the knowledge graph is also able to enable precise comparison and reasoning. For example, "isc:Ochoan" has base and top ages 259.81 Ma and 259.41 Ma, respectively. In comparison, the base and top ages of "isc:Changhsingian" are 254.14 Ma and 251.902 Ma, respectively. Using such numerical information together with the classes (i.e., "isc:Epoch" and "isc:Age") of those two intervals, a reasoning engine can definitely tell that "isc:Ochoan" is not within "isc:Changhsingian". Moreover, with the collected numerical information, we will be able to incorporate the existing ontologies and schemas to give more precise description of the geologic time intervals, instants, and their topological relationships (see next section).

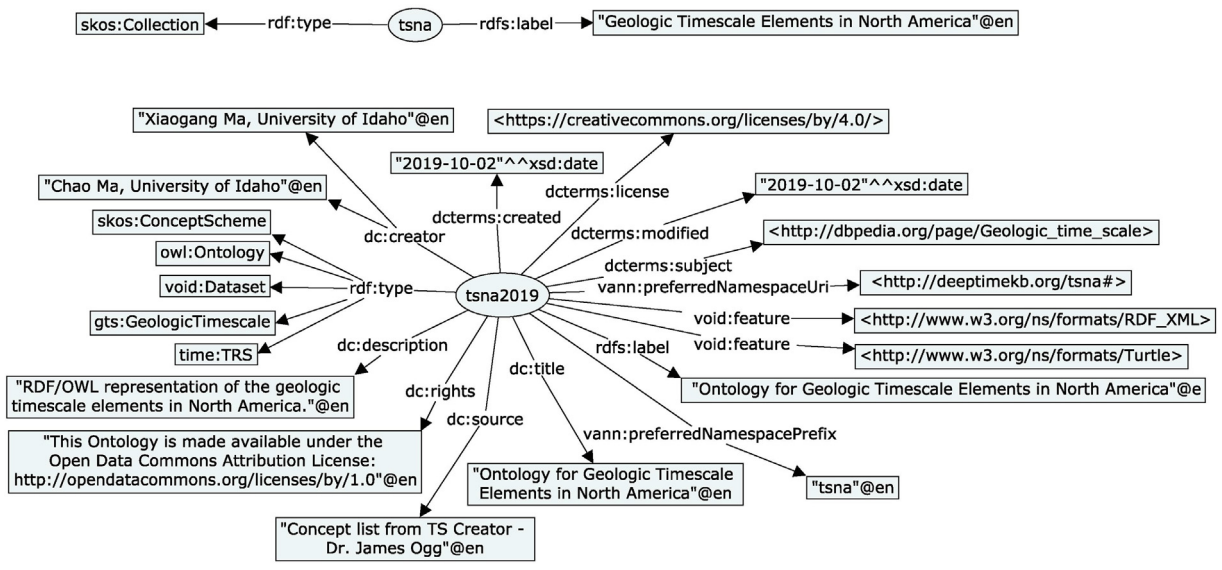


Fig. 1. Metadata and structure of the region's namespace prefix (e.g., tsna) and its corresponding vocabulary scheme (e.g., tsna2019).

```
tsna:
  rdf:type skos:Collection ;
  rdfs:label "Geologic Timescale Elements in North America"@en
.

ts:tsna2019
  rdf:type void:Dataset ;
  rdf:type gts:GeologicTimescale ;
  rdf:type owl:Ontology ;
  rdf:type skos:ConceptScheme ;
  rdf:type time:TRS ;
  dc:creator "Xiaogang Ma, University of Idaho"@en ;
  dc:creator "Chao Ma, University of Idaho"@en ;
  dc:description "RDF/OWL representation of the geologic timescale elements in North America."@en ;
  dc:rights "This Ontology is made available under the Open Data Commons Attribution License: http://opendatacommons.org/licenses/by/1.0"@en ;
  dc:source "Concept list from TS Creator – Dr. James Ogg"@en ;
  dc:title "Ontology for Geologic Timescale Elements in North America"@en ;
  dcterms:created "2019-10-02"^^xsd:date ;
  dcterms:license <https://creativecommons.org/licenses/by/4.0/> ;
  dcterms:modified "2019-10-02"^^xsd:date ;
  dcterms:subject <http://dbpedia.org/page/Geologic_time_scale> ;
  vann:preferredNamespacePrefix "tsna"@en ;
  vann:preferredNamespaceUri <http://deeptimekb.org/tsna#> ;
  void:feature <http://www.w3.org/ns/formats/RDF_XML> ;
  void:feature <http://www.w3.org/ns/formats/Turtle> ;
  rdfs:label "Ontology for Geologic Timescale Elements in North America"@en
.
```

Fig. 2. RDF code defining the concept collection of North America and the vocabulary scheme.

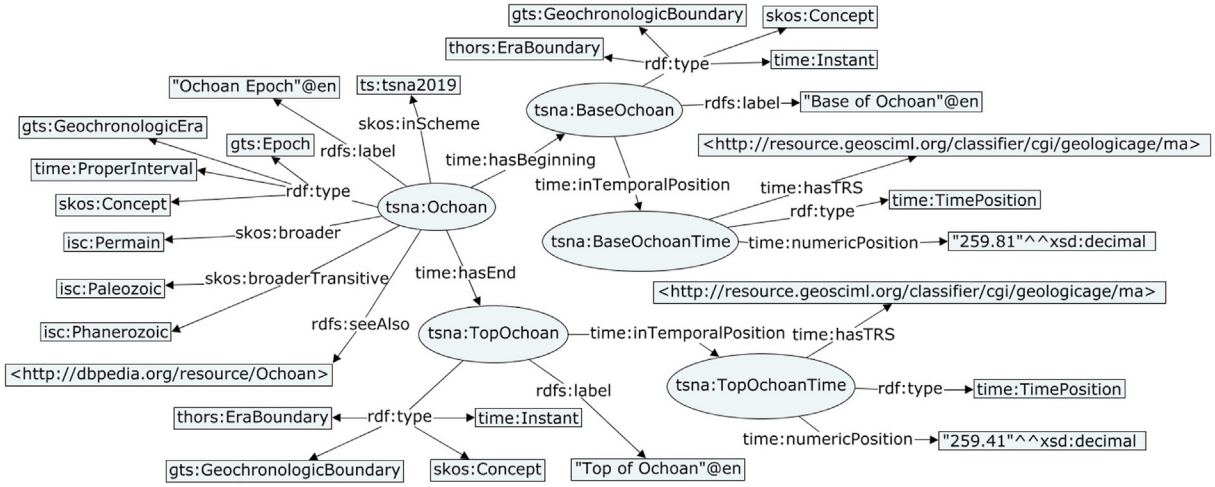


Fig. 3. Definition of Ochoan Epoch (tsna:Ochoan) in the vocabulary scheme for North America.

2.3. Topology of geologic time concepts

The Time Ontology (Cox and Little, 2020) defines the relations among intervals and instants of time. Its critical logic is from the analysis of temporal topology (Allen, 1984). Most geologic time concepts (e.g., Epoch) are time intervals. In the recent knowledge graph of the international geologic time scale (Cox, 2020), the temporal topology has been incorporated. In our knowledge graph for the regional geologic time standards, we have incorporated the temporal topology in two ways: (1) directing adding topological relationships between intervals in the knowledge graph; (2) developing functions to infer those topological relationships by using the existing properties (e.g., top and bottom boundaries of each interval) in the knowledge graph. In an R package that is currently under development (Ma et al., 2020a), we have a function to return the topological relationship between any two geologic time concepts (intervals and/or instants).

3. Deep time knowledge base construction and online service

3.1. An automated workflow for transforming spreadsheet into a knowledge graph

The raw data we collected for the regional geologic time standards are in a spreadsheet that has the regional Epoch, Age, and their boundaries' ages. We automated the transformation of the raw data to a knowledge graph in RDF code by using a template and an R workflow. The materials and code include several files: template.txt, data.csv, code.R, result.html, and pbdb.R, and all of them are shared on GitHub (see link in Code Availability section at the end of this paper). Additionally, a version of the GitHub Repo was archived in early September 2020 and is accessible on Zenodo (Ma, 2020). Every regional interval at the Epoch or Age level should have one (Fig. 7) or two (Fig. 4) boundaries defined, according to which we created an RDF code template (template.txt) that only certain places ("REPLACE#" in RDF code template) need to be replaced with regional geologic time intervals from the raw data. We formatted the raw data to a structure (data.csv) that can be manipulated by the R code. It has 11 columns: "Region", "NamespacePrefix", "ConceptName", "ConceptLevel", "Top (Ma)", "Bottom (Ma)", "broader is local?", "broader", "broaderTransitive", "broaderTransitive", and "broaderTransitive". The procedures of replacing "REPLACE#" with the raw data (Fig. 8) are coded in the R script and explained in code comments (code.R). This automated process makes the data transformation very quick and can be used for future extension of other regional geologic time standards.

The structure of the outputting knowledge graph follows the ontology patterns designed in Cox and Richard (2015) and Cox (2020). With minor modification to the established workflow in R, we are also able to add more information to the knowledge graph for other topics of interest. For example, in the knowledge

```
tsna:Ochoan
  rdf:type gts:Epoch ;
  rdf:type gts:GeochronologicEra ;
  rdf:type skos:Concept ;
  rdf:type time:ProperInterval ;
  rdfs:label "Ochoan Epoch"@en ;
  rdfs:seeAlso <http://dbpedia.org/resource/Ochoan> ;
  skos:broader isc:Permian;
  skos:broaderTransitive isc:Paleozoic ;
  skos:broaderTransitive isc:Phanerozoic ;
  skos:inScheme ts:tsna2019 ;
  time:hasBeginning  tsna:BaseOchoan ;
  time:hasEnd tsna:TopOchoan ;
.

tsna:BaseOchoan
  rdf:type gts:GeochronologicBoundary ;
  rdf:type thors:EraBoundary ;
  rdf:type skos:Concept ;
  rdf:type time:Instant ;
  rdfs:label "Base of Ochoan"@en ;
  skos:prefLabel "Base of Ochoan"@en ;
  time:inTemporalPosition tsna:BaseOchoanTime ;
  rdfs:seeAlso isc:BaseLopingian ;
.

tsna:BaseOchoanTime
  rdf:type time:TimePosition ;
  time:hasTRS <http://resource.geosciml.org/classifier/cgi/geologicage/ma> ;
  time:numericPosition "259.81"^^xsd:decimal
.

tsna:TopOchoan
  rdf:type gts:GeochronologicBoundary ;
  rdf:type thors:EraBoundary ;
  rdf:type skos:Concept ;
  rdf:type time:Instant ;
  rdfs:label "Top of Ochoan"@en ;
  skos:prefLabel "Top of Ochoan"@en ;
  time:inTemporalPosition tsna:TopOchoanTime ;
.

tsna:TopOchoanTime
  rdf:type time:TimePosition ;
  time:hasTRS <http://resource.geosciml.org/classifier/cgi/geologicage/ma> ;
  time:numericPosition "259.41"^^xsd:decimal
.
```

Fig. 4. RDF code of for Ochoan Epoch (tsna:Ochoan) in the vocabulary scheme for North America.

Fig. 5. Definition of "gts:Epoch" in the gts ontology (see Web link in Table 1).

Class:	time:ProperInterval
Definition:	A temporal entity with non-zero extent or duration, i.e. for which the value of the beginning and end are different
Subclass of:	time:Interval
Disjoint with:	time:Instant

Fifteen properties :intervalBefore, :intervalAfter, :intervalMeets, :intervalMetBy, :intervalOverlaps, :intervalOverlappedBy, :intervalStarts, :intervalStartedBy, :intervalDuring, :intervalContains, :intervalFinishes, :intervalFinishedBy, :intervalEquals :intervalDisjoint :intervalIn support the set of interval relations defined by Allen [al-84] and Allen and Ferguson [af-97].

Fig. 6. Definition of "time:ProperInterval" in the time ontology (see Web link in Table 1).

```
tsna:Guadalupian
  rdf:type gts:Epoch ;
  rdf:type gts:GeochronologicEra ;
  rdf:type skos:Concept ;
  rdf:type time:ProperInterval ;
  rdfs:label "Guadalupian Epoch"@en ;
  rdfs:seeAlso <http://dbpedia.org/page/Category:Guadalupian> ;
  rdfs:seeAlso isc:Guadalupian ;
  skos:broader isc:Permian ;
  skos:broaderTransitive isc:Paleozoic ;
  skos:broaderTransitive isc:Phanerozoic ;
  skos:inScheme ts:tsna2019 ;
  time:hasBeginning tsna:BaseGuadalupian ;
  time:hasEnd tsna:BaseOchoan ;
.

tsna:BaseGuadalupian
  rdf:type gts:GeochronologicBoundary ;
  rdf:type thors:EraBoundary ;
  rdf:type skos:Concept ;
  rdf:type time:Instant ;
  rdfs:label "Base of Guadalupian"@en ;
  skos:prefLabel "Base of Guadalupian"@en ;
  time:inTemporalPosition tsna:BaseGuadalupianTime ;
.

tsna:BaseGuadalupianTime
  rdf:type time:TimePosition ;
  time:hasTRS <http://resource.geosciml.org/classifier/cgi/geologicage/ma> ;
  time:numericPosition "272.3"^^xsd:decimal
.
```

Fig. 7. The Guadalupian Epoch (tsna:Guadalupian) that is below and adjacent to the Ochoan Epoch (tsna:Ochoan) in Fig. 4.

graph for the international geologic time scale (Cox, 2020), there are records about the correlation events associated with each ratified Global Boundary Stratotype Sections and Points (GSSP, or 'Golden Spike'). The same structure can also be reused for recording the correlation events and publications related to stratigraphic points and boundaries (i.e., instants) in the regional geologic time scales. In our work, the dataset retrieved from the Time Scale Creator (Ogg and Lugowski, 2020) actually has a column for such correlation events. However, many of them are just a short note written in free text. The structure is not consistent and the reference is often missing. In the current knowledge graph, we decided not include those correlation event records. We will contact researchers in the community of stratigraphy to verify and update those records in order to include them in a future version of our knowledge graph.

3.2. A SPARQL endpoint for the deep time knowledge base

Example 1: In this example we queried the beginning time and end time of "Wordian Age", assuming that we don't know whether it appears in regional or international geologic time standards. The query returns the result both from regional (tsna: North America) and different versions of the international geologic time scale. The time of each boundary in the version history of geologic time standards is within "dc:description []" (Ma et al., 2020b) while the knowledge graph for regional geologic time standards does not have such structure (Figs. 3 and 4), thus the queries were written in two different patterns and combined by "UNION". We were also able to retrieve the time range of any regional or international geologic time intervals by replacing the "Wordian" in this query code. In this example, we assumed that a user does not know whether the time interval is regional or international. On the other hand, if the user knows the region or version, they can modify the "skos:inScheme" clause in the query code for a specific query.

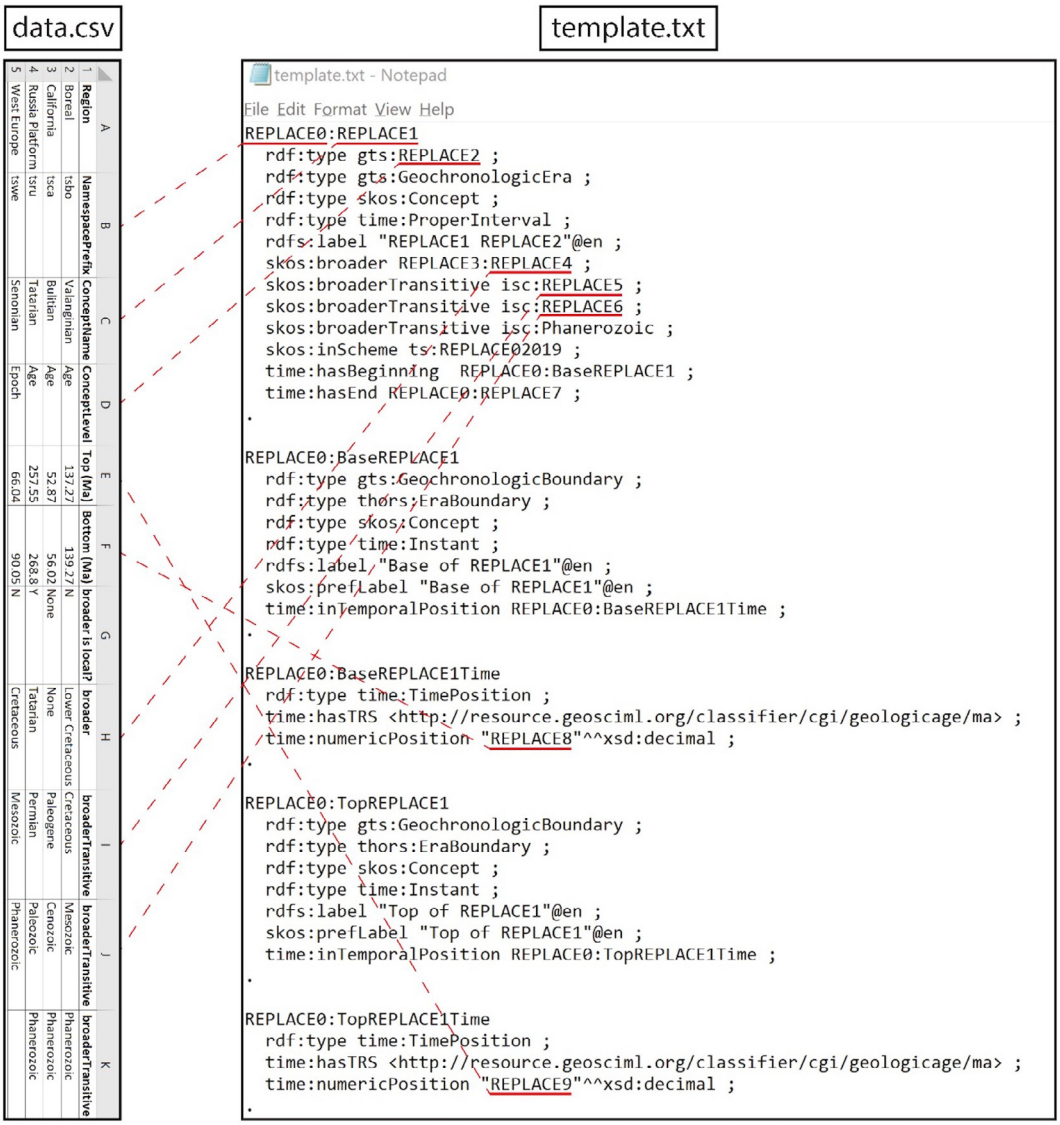
Example 2: Query all geologic time intervals within the interval from 13 to 20 Ma. Because of the same reason as in Example 1, "UNION" is applied to query international and regional geologic time standards. "skos:inScheme" can be modified to query different regional standards. "skos:inScheme" can be modified to query different versions of international standards. Number "13" and "20" can be edited to set a different interval. The returned result has two columns: scheme ID and geologic time concepts. The result includes intervals from several regional standards and one same interval from different versions of the international standard.

Example 3: Query geologic time intervals that are crossed with the 10 Ma boundary. The code structure is the same as that in Examples 1 and 2 and can be modified to change between regions and between versions of the international standard. Number "10" representing "10 Ma" can also be changed to represent a different boundary in geologic time.

These queries in the above three examples can be easily modified to query different time intervals in different regions and in different versions of the international geologic time scale. We are currently developing an R package to implement these functions and make them more common to use in workflow environments (Ma et al., 2020a).

3.3. Use the knowledge graph and service in open data exploration

The developed knowledge graph of regional geologic time standards can be used for data exploration with a lot of open data repositories in the geoscience community, as long as geologic time is a dimension of the data. One recent use case was tested with the Paleobiology Database (PBDB) (Peters and McClennen, 2016). PBDB is a public database that provides global occurrence and taxonomic data for organisms throughout Earth history. It is one of the most scientifically productive databases and platforms that enable data-driven discovery in Earth science. An R package, paleobioDB (Varela et al., 2015), was developed for assisting data exploration through the API of PBDB. In our interactions with PBDB, we found that, if the intervals from the international geologic time scale are used as inputs, then PBDB can recognize the time span of the interval and then return corresponding results. For example, the Canidae fossil occurrences in Oligocene can be obtained by running *pbdb_occurrences (limit="all", base_name="Canidae", interval="Oligocene", show = c("coords", "phylo", "ident"))* in R with the paleobioDB package. The parameter "interval" is for inputting a geologic time



	A	B	C	D	E	F	G	H	I	J	K
1	Region	NamespacePrefix	ConceptName	ConceptLevel	Top (Ma)	Bottom (Ma)	broader is local?	broader	broaderTransitive	broaderTransitive	broaderTransitive
2	Boreal	tsbo	Valanginian	Age	137.27	139.27 N		Lower Cretaceous	Cretaceous	Mesozoic	Phanerozoic
3	California	tsca	Bulitian	Age	52.87	56.02 None		None	Paleogene	Cenozoic	Phanerozoic
4	Russia Platform	tsru	Tatarian	Age	257.55	268.8 Y		Tatarian	Permian	Paleozoic	Phanerozoic
5	West Europe	tswe	Senonian	Epoch	66.04	90.05 N		Cretaceous	Mesozoic	Phanerozoic	

Fig. 8. Interaction of the raw data and the templated RDF code, which is processed by an R workflow (see Code Availability section at the end of this paper).

interval. If it is replaced with a regional geologic time interval (e.g. Otaian Age in New Zealand), the time span of the interval will not be recognized, and no results of fossil occurrences will be returned. This is where the knowledge graph of regional geologic time standards is needed. The top and bottom ages of a regional geologic time interval can be obtained through our knowledge graph, and then be used in querying PBDB. The details of this use case are shown in the file pbdb.R in the folder "Paper_Materials" on our GitHub Repo (See link in the Code Availability section at the end of this paper).

4. Discussion

This research addresses the shortage of machine-readable knowledge graphs for regional geologic time standards. The usage of regional geologic time standards is relatively low in comparison with the intentional geologic time scale, but the concepts (both intervals and instants) in regional time scales are able to give finer details about time and they appear in many publications and databases. Also, regional geologic time standards and terminologies are large in numbers, as described above. In research projects that assemble datasets for various regions, the heterogeneity in the regional geologic time terminologies will be a big hurdle for data integration. The developed knowledge graph and service are open and queryable on the Web. With this service, the intervals and instants in the regional geologic time scales are no longer just labels but are detailed with machine-readable information about their type, attributes, and inter-relationships. Moreover, in the service structure, the developed knowledge graph is integrated with the knowledge graph for the international geologic time scale in a single SPARQL endpoint, which makes it a comprehensive resource for geologic time concepts.

We reused the ontologies developed by Cox and Richard to make the result consistent with community standards. The framework for the vocabulary scheme of each region is also derived from the vocabularies for the international geologic time scale. Because of the consistency with existing ontologies and vocabularies for geologic time, we were able to load them in the same SPARQL endpoint and treat them together as a single knowledge source (see examples in Section 3.2). There are two small differences between the knowledge graphs for regional and international standards. The

first is that along with the knowledge graph for regional standards we developed functions to infer topology from existing records. The other difference is that we have not introduced version control in the knowledge graph for regional standards. This could be extended in future work.

The service-oriented architecture will make the developed knowledge graph easy to use in workflow platforms. We are developing an R package to realize several common functions for querying the knowledge graph (including both regional and international standards), such as (1) querying and mapping the Global Boundary Stratotype Section and Point (GSSP) data, (2) topological relationship between two geologic time concepts (intervals and/or instants), (3) querying start time, end time and duration of a specific geologic time interval, (4) getting the unit level of a geologic time interval, and (5) getting the broader and narrower intervals of a specified geologic time interval. These functions are designed to access the knowledge graph and are easy to implement in workflow platforms for data-intensive work. An end-user, especially a geoscientist, will have no need to know the ontologies and conceptual framework inside the knowledge graph. The user only needs to load the R package in a workflow platform and then run functions to retrieve information of interest from the knowledge graph.

The knowledge graph of regional geologic time standards will promote data interoperability across various data sources. For any given geologic time intervals or instants, their machine-readable meanings can be retrieved from the knowledge graph and they are comparable to each other in a common temporal framework. This functionality can be used in open data exploration of many topics. Besides the use cases in Section 3.3, it can also be applied at a larger scale with distributed databases focusing on the geology of different regions. Algorithms can be developed to automatically translate heterogeneous regional geologic time concepts into quantitative records about their start and end ages and time coverage. Such functionality will reduce the burden for human researchers and smooth the workflow of data cleansing and integration. Though those use cases we will collect feedback from geoscientists and plan the other regional geologic time standards to be collected and transformed into the knowledge graph. Another topic of interest is to use the knowledge graph in text mining with massive literature data. We are planning to add multilingual labels to the knowledge graph for regional geologic time standards and extend the number of regions covered in the knowledge graph. Those will make the resulting knowledge graph more functional to support texting mining.

Our knowledge base can be used in a spectrum of applications ranging from database engineering to data analysis in the aspect of semantic support. Databases can embed the machine-readable concepts into their data. For example, a database can replace the label "Ochoan Epoch" with the concept tsna:Ochoan in our knowledge base. These can improve the interoperability of data, a key part of the FAIR principles. In another scenario, our knowledge base can provide semantic support to the work of processing data from different databases or texts, as it can automatically transfer textural labels into machine-readable concepts with geologic meanings. In the data-driven discovery of Earth evolution, there will be massive geologic time concepts buried in the big data. The form and meaning of these concepts are heterogenous, which has always been a challenge for efficient data integration and accurate analysis. For example, the paleogeographic reconstruction of Jurassic needs data that belongs to the corresponding time span, however, part of the Jurassic data from 2004 (following GTS 2004) may not fall into today's time span of Jurassic (following GTS 2020). The knowledge base is able to recognize these differences and quantify the uncertainty.

The ultimate goal of building the knowledge base is to help automate the knowledge and data mining in geoscience. Although the current result has already shown usefulness in several examples, it still has limitations and challenges. Although we have designed a structure to record version changes (Ma et al., 2020b), there may still be some changes in the meanings of boundaries that are not stored in our knowledge base. For instance, the base of Pleistocene Epoch (and also the base of Quaternary Period) was proposed to be changed from the base of Calabrian Age to that of the Gelasian Age in 2008, and it was approved in 2009. Our knowledge base is able to record that version change, however, the detailed background information for that change has not been captured in the knowledge base yet. Moreover, the 17 regional geologic time scales in our knowledge base currently have no information for historical versions, uncertainties, and multilingual labels. Nevertheless, the knowledge base is based on community standards and has a good extendibility. We plan to develop an input module for new updates to be quickly incorporated into the knowledge base.

5. Conclusion

We present the design and construction of a knowledge graph for regional geologic time standards. By reusing community-level standards, the resulting knowledge graph is consistent with the knowledge graph of international geologic time standards. We integrate the knowledge graphs of regional and international geologic time standards in a single SPARQL endpoint and make it an open and comprehensive resource for geologic time concepts (both intervals and instants). Experimental use cases in this study proved the usability of the endpoint as a machine-readable reference for various regional and international geologic time concepts. As such, the knowledge graph and service will mitigate the barriers of geoscience data interoperability across multiple sources. An R package is under development for accessing the endpoint from workflow platforms such as Jupyter Notebook and R Markdown. In our future work, we will incorporate geologic time standards of more regions into the knowledge graph and collect multilingual labels for the regional geologic time standards.

6. Code availability

The source code of the knowledge graphs developed in this work is open and accessible on GitHub at https://github.com/xgmachina/DeepTimeKB. The materials and code mentioned in Section 3 for knowledge graph construction (template.txt, data.csv, code.R, result.html, pbdb.R, and examples.docx) are in the folder "Paper_Materials" of this GitHub Repo. Additionally, the GitHub Repo is archived on Zenodo with a permanent DOI at https://doi.org/10.5281/zenodo.4025479 (Ma, 2020), where all the new releases can also be found. A SPARQL endpoint has been set up for the comprehensive knowledge graph of international and regional geological time standards, which is accessible at http://virtuoso.nkn.uidaho.edu:8890/sparql/ using the graph name http://deeptimekb.org/iscallnew.

Author contribution

X. Ma designed the work. C. Ma and X. Ma led the coding with assistance from A.S. Kale and J. Zhang. X. Ma and C. Ma drafted and edited the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work presented in the paper was supported by the National Science Foundation, United States (No. 1835717). Additional support was provided by the IUGS Deep-time Digital Earth (DDE) Big Science Program, the Deep Carbon Observatory, the Alfred P. Sloan Foundation, and the Carnegie Institution for Science for communicating the research progress at several workshops and meetings. We thank three anonymous reviewers for their constructive comments on an earlier version of the manuscript.

References

Allen, J.F., 1984. Towards a general theory of action and time. Artif. Int. 23 (2), 123–154.
Cox, S.J.D., Little, C., 2020. Time Ontology in OWL. https://www.w3.org/TR/owl-time/ (accessed 30 July 2020).
Cox, S.J.D., 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 1 (3), 119–137.
Cox, S.J.D., Richard, S.M., 2015. A geologic timescale ontology and service. Earth Sci. Inform. 8 (1), 5–19.
Cox, S.J.D., Yu, J., Rankine, T., 2016. SISSVoc: A Linked Data API for access to SKOS vocabularies. Semant. Web 7 (1), 9–24.
Cox, S.J.D., 2011. OWL representation of the geologic timescale implementing stratigraphic best practice, in: Proceedings of 2011 AGU Fall Meeting, San Francisco, 5-9 December 2011, Abstract IN31B-1440.
Cox, S.J.D., 2020. Geologic Timescale, GitHub. https://github.com/CGI-IUGS/timescale-data (accessed 09 September 2020).
Gil, Y., Pierce, S.A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., Horel, J., 2019. Intelligent systems for geosciences: an essential research agenda. Commun. ACM 62 (1), 76–84.
Gradstein, F.M., Ogg, J.G., Schmitz, M.B., Ogg, G.M., 2020. Geologic Time Scale 2020. Elsevier, Amsterdam, Netherlands, p. 1342.
Haq, B.U., van Eysinga, F.W.B., 2007. The Geological Time Table. Elsevier, Amsterdam, Netherlands.
Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.-C.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A., 2022. Knowledge graphs. ACM Comput. Surv. 54 (4), 71. https://doi.org/10.1145/3447772.
Ma, X., 2020. Deep Time Knowledge Base. Zenodo. https://doi.org/10.5281/zenodo.4025479.
Ma, C., Ma, X., Crump, R., Kale, A.S., 2020a. Knowledge graphs for global and regional geologic time scales and an associated R package, in: 2020 AGU Fall Meeting, Abstract.
Ma, X., Fox, P., 2013. Recent progress on geologic time ontologies and considerations for future works. Earth Sci. Inform. 6, 31–46.
Ma, X., Fox, P., Rozell, E., West, P., Zednik, S., 2014. Ontology dynamics in a data life cycle: challenges and recommendations from a Geoscience Perspective. J. Earth Sci. 25 (2), 407–412.
Ma, X., Ma, C., Wang, C., 2020. A new structure for representing and tracking version information in a deep time knowledge graph. Comput. Geosci. 145, 104620.
Michalak, J., 2005. Topological conceptual model of geological relative time scale for geoinformation systems. Comput. Geosci. 31, 865–876.
Miles, A., Bechhofer, S., 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendataion. https://www.w3.org/TR/skos-reference (accessed 08 February 2021).
NASEM (National Academies of Sciences, Engineering, and Medicine), 2020. A Vision for NSF Earth Sciences 2020-2030: Earth in Time. The National Academies Press, Washington, D.C., 172 pp., https://doi.org/10.17226/25761.
NRC (National Research Council), 2008. Origin and Evolution of Earth: Research Questions for A Changing Planet. National Academies Press, Washington, D.C., p. 150.
Ogg, J.G., Lugowski, A., 2020. TimeScale Creator - a visualization system - Tour and Exercises. https://timescalecreator.org/download/TS_tour_Exercises.pdf (accessed 03 September 2020).
Perrin, M., Mastella, L.S., Morel, O., Lorenzatti, A., 2011. Geological time formalization: an improved formal model for describing time successions and their correlation. Earth Sci. Inform. 4 (2), 81–96.
Peters, S.E., McClennen, M., 2016. The Paleobiology Database application programming interface. Paleobiology 42 (1), 1–7.
Sheth, A., Padhee, S., Gyrard, A., 2019. Knowledge graphs and knowledge networks: The story in brief. IEEE Internet Comput. 23 (4), 67–75.
Varela, S., González-Hernández, J., Sgarbi, L.F., Marshall, C., Uhen, M.D., Peters, S., McClennen, M., 2015. paleobioDB: an R package for downloading, visualizing and processing data from the Paleobiology Database. Ecography 38 (4), 419–425.
Wang, C., Ma, X., Chen, J., 2018. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. Comput. Geosci. 115, 12–19.