Geoscience Frontiers xxx (xxxx) xxx

Contents lists available at ScienceDirect

Geoscience Frontiers

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

A comprehensive construction of the domain ontology for stratigraphy

Huiqing Xu [a], Yingying Zhao [a], Hao Huang [b,*], Shaochun Dong [a], Yukun Shi [a], Chunju Huang [c], Huaichun Wu [d,e], Zhiqi Qian [c], Qiang Fang [d,e], Huaguo Wen [f], Zhongtang Su [f], Shuang Dai [g,h], Ronghua Wang [g,h], Chao Li [i], Chao Sun [j], Junxuan Fan [a,k]

[a] School of Earth Sciences and Engineering and Frontiers Science Center for Critical Earth Material Cycling, Nanjing University, Nanjing 210023, China
[b] Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China
[c] State Key Laboratory of Biogeology and Environmental Geology, School of Earth Sciences, China University of Geosciences, Wuhan 430074, China
[d] State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Beijing 100083, China
[e] School of Ocean Sciences, China University of Geosciences, Beijing 100083, China
[f] State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation & Institute of Sedimentary Geology, Chengdu University of Technology, Chengdu 610059, China
[g] School of Earth Sciences and Key Laboratory of Mineral Resources in Western China (Gansu Province), Lanzhou University, Lanzhou 730000, China
[h] Key Laboratory of Strategic Mineral Resources of the Upper Yellow River, Ministry of Natural Resources, Lanzhou 730000, China
[i] State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology and Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Nanjing 210008, China
[j] Institute of Geology, Chinese Academy of Geological Sciences, Beijing 100037, China
[k] State Key Laboratory for Mineral Deposits Research, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 23 January 2022
Revised 30 June 2022
Accepted 22 August 2022
Available online xxxx

Keywords:
Domain ontology
Stratigraphy
Biostratigraphic unit
Biostratigraphic horizon
Fossil

ABSTRACT

Stratigraphic knowledge, the cornerstone of geoscience, needs to be represented by the Knowledge Graph based upon ontology, in order to apply the state-of-the-art big-data techniques. This study aims to comprehensively construct the ontologies for the stratigraphic domain. This has been achieved by a federated, crowd intelligence-based collaboration among domain experts of major stratigraphic subdisciplines. The initial step is to enumerate key terms from authoritative references and incorporate them into the Geoscience Professional Knowledge Graphs (GPKGs) of Deep-time Digital Earth Project. During this process, semantic heterogeneities were meticulously addressed by professional judgement aided by an automatic detection of Homonyms at the GPKGs platform. Afterwards, these terms were further differentiated as either classes or properties and arranged in a hierarchical framework in a top-down process. Consequently, seven ontologies are constructed for major stratigraphic branches, i.e., Lithostratigraphy, Biostratigraphy, Chronostratigraphy, Chemostratigraphy, Magnetostratigraphy, Cyclostratigraphy and Sequence Stratigraphy. The ontology of Biostratigraphy, among them, is elaborated here, as no biostratigraphic ontology has been attempted before to our knowledge. The constructed biostratigraphic ontology comprises following major root classes: Fossil, Biostratigraphic unit, Biostratigraphic horizon. Altogether, they contribute to the eventual dating and correlating of strata in another root class: Biostratigraphic correlation. In summary, the achievements of this study are probably heretofore the most comprehensive ontologies for the stratigraphic domain. Moreover, a proto model of semantic search engine was conceived to discuss potential application of our work for better querying stratigraphic references, utilizing the semantic liaison of the classes in the constructed ontologies.

1. Introduction

Knowledge Graph (KG) based on domain ontologies has been increasingly gaining traction in academic communities, as it enables integration, reuse, semantic interoperability and automated reasoning of heterogeneous data in a machine-understandable way (Uschold and Gruninger, 1996; Sinha, 2006; Ehrlinger and Wöβ, 2016). As a typical data-intensive domain, the geoscience community is currently facing an evolutionary transition from traditional encyclopedic discipline knowledge system to the machine-understandable KG (Zhou et al., 2021). Among diverse geological fields, the Stratigraphy KG is indispensable for such transition, as the stratigraphic sequence forms the

* Corresponding authors at: Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China (H. Huang).
E-mail addresses: hhuang@mail.iggcas.ac.cn (H. Huang), jxfan@nju.edu.cn (J. Fan).

https://doi.org/10.1016/j.gsf.2022.101461
1674-9871/

Please cite this article as: H. Xu, Y. Zhao, H. Huang et al., A comprehensive construction of the domain ontology for stratigraphy, Geoscience Frontiers, https://doi.org/10.1016/j.gsf.2022.101461

fundamental spatio-temporal framework for exploring the geological history and thus almost all other geo-disciplines are more or less dependent on stratigraphic data. In particular, the International Commission on Stratigraphy (ICS) has recently emphasized one of its major tasks to provide data services concerning the Global Boundary Stratotype Sections and Points (GSSPs) and the International Chronostratigraphic Chart (Cohen et al., 2013). The Stratigraphy KG forms one essential infrastructure to fulfill such goals of the ICS.

As the core base of KG, the ontology provides formal, explicit specification of a shared conceptualization by high semantic expressiveness (Gruber, 1995; Studer et al., 1998; Feilmayr and Wöβ, 2016). The work on the Ontology of Geosciences was pioneered by the GEOscience Network (Keller, 2003) and the semantic web for Earth and environmental terminology (SWEET) (Raskin and Pan, 2005), subsequently followed by several case studies. Among them, the construction and application of the ontologies of the Geologic Time Scale (GTS) and Chronostratigraphy have received considerable attention (see review of Ma and Fox, 2013; Ma et al., 2020). For instance, the ontology models of GTS and GSSPs have been constructed in various versions, e.g., models of UML (Unified Modeling Language), SKOS (Simple Knowledge Organization System) or OWL (Web Ontology Language) (Cox and Richard, 2005, 2015; Raskin and Pan, 2005; Dong et al., 2010; Ma et al., 2011, 2012, 2020; Hou et al., 2015, 2018; Cox, 2016; Wang et al., 2018; Cox and Little, 2020). An interactive International Chronostratigraphic Chart is available by utilizing Time Ontology in OWL (https://stratigraphy.org/timescale/). Other authors also tried to apply strata-related ontologies to integrate borehole data, petroleum data or mineral data for automatic lithostratigraphic correlation or visual interpretation of sedimentary processes (Wu et al., 2008; Perrin, et al., 2011; Carbonera et al., 2015; Wang et al., 2018; Yuan et al., 2021). Nevertheless, several fundamental components of the Stratigraphy disciplines, e.g., Biostratigraphy, Sequence Stratigraphy, Cyclostratigraphy, Magnetostratigraphy etc., have thus far hardly been touched in terms of ontology construction. It is thus indeed necessary, although challenging, to comprehensively construct the ontology for the stratigraphic domain under the framework of earth system science.

This study aims to construct the Stratigraphy Ontology, comprehensively, for seven major subdisciplines by federated, crowd intelligence-based collaboration. Our work is part of the Deep-time Digital Earth Project (DDE), one of whose motivations is to establish an all-domain geoscience KG with the collaboration of professionals in diverse geo-domains across the world (Zhou et al., 2021). Moreover, a proto model was conceived to demonstrate potential application of our work. Results of this study lay the foundation for the interoperability of stratigraphical data in a machine-understandable way and would hopefully benefit data-driven theoretical breakthroughs in future geological studies.

The rest of the paper is organized as follows: Section 2 describes the methods and procedures to construct the ontology models; Section 3 presents the major outcomes of our work, namely ontology models for seven stratigraphic subdisciplines, with emphasis on the biostratigraphy due to limited space; we further demonstrate, in Section 4, the potential application of our work in semantic search for improving data queries in the stratigraphic domain and present the conclusion in Section 5.

2. Construction methods

The seven-step method of Noy and McGuinness (2001) was critically consulted during the ontology construction of this study. After carefully reviewing the existing strata-related ontologies mentioned above, expert groups focusing on different stratigraphic subdisciplines manually extracted terms from authoritative monographs, and further defined classes (concepts), properties and relationships. The construction was progressively updated in an iterative process, with repetitive revisions according to reviews of domain experts. The ontology was eventually coded by the open-source Protégé software (Musen, 2015). The details of major procedures of our work are presented as follows (Fig. 1).

2.1. Determine domain and enumerate terms

The stratigraphic study describes rocks at a global scale by diverse tangible properties, e.g., lithology, fossil, geochemistry, age, etc. Accordingly, several categories of stratigraphic units are classified to express these different aspects. Therefore, we chose seven major subdisciplines in the stratigraphic domain: Lithostratigraphy, Biostratigraphy, Chronostratigraphy, Cyclostratigraphy, Magnetostratigraphy, Chemostratigraphy and Sequence Stratigraphy (Fig. 1). The reason to choose these seven subdomains is twofold: (1) they constitute the bulk of the stratigraphic knowledge; (2) the academic and industrial activities pertaining to them have been flourishing.

Vast number of terms bearing scientific significance were extracted and screened by experts from authoritative monographs for each of these seven domains. The qualified terms are essential conceptions, usually rigorously defined and explained in case studies in each chapter of monographs and oftentimes indispensable for the completeness of the knowledge framework. Glossaries appended in monographs are often best candidates for such terms. Previous compilation such as geoscience vocabularies of Commission for the Management and Application of Geoscience Information (CGI, https://cgi-iugs.org/project/geosciceterminology/) are also valuable sources.

Terms of Lithostratigraphy, Biostratigraphy and Chronostratigraphy were mainly selected from *the International Stratigraphic Guide* (Salvador, 1994), supplemented by *North American Stratigraphic Code* (North American Commission on Stratigraphic Nomenclature, 2005), *Stratigraphic Guide of China* (China Commission of Stratigraphy, 2017) and *The Geologic Time Scale 2012* (Gradstein et al., 2012). The core vocabularies of Chemostratigraphy, Magnetostratigraphy, Cyclostratigraphy and Sequence Stratigraphy were garnered from, respectively, *Magnetic Stratigraphy* (Opdyke and Channell, 1996), *Cyclostratigraphy–concepts, Definitions, and Applications* (Strasser et al., 2006), *Time-series Analysis and Cyclostratigraphy: Examining Stratigraphic Records of Environmental Cycles* (Weedon, 2003), *Sequence Stratigraphy* (Ji, 2005) and *High-resolution Sequence Stratigraphy* (Zheng et al., 2010).

2.2. Define classes and develop a hierarchy

A top-down development process (Uschold and Gruninger, 1996) was adopted for our work because this logic is consistent with the inherent knowledge structure in the aforementioned monographs. Besides, it would be impossible to list all stratigraphic terms to apply the bottom-up process, given the complex and long evolution of stratigraphic ideas among scholars from disparate countries. Furthermore, we selected the independent terms of first-order significance to be classes, and the others with marginal significance or describe attributes of the objects to be properties. The hierarchy of ontology was further developed by stepwise decentralization of general classes to more specialized ones (subclasses) and establishing relations between different classes (e.g., part-of, attribute-of).

All enumerated terms were uploaded to the Geoscience Professional Knowledge Graphs of DDE (hereafter as GPKGs and can be accessed in the Knowledge hub of the DDE website: https://edi-

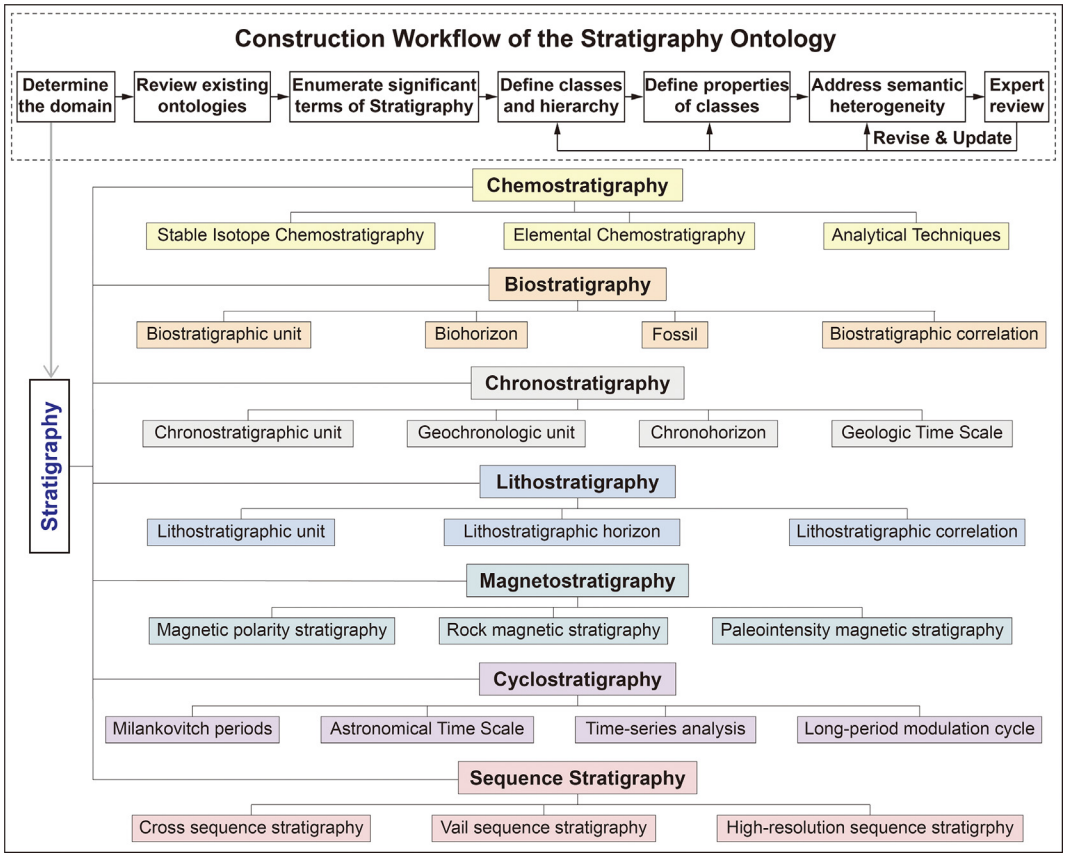

Fig. 1. Construction workflow for the Stratigraphy Ontology and accomplished seven branches for major stratigraphic subdisciplines.

tor.deep-time.org/KgEditorWeb/#/graph) as knowledge nodes. For convenience, three types of nodes were further classified: Defined node, Cited node and Auxiliary node. In one particular ontology, defined nodes were vital concepts accompanied with explicit definition, while cited nodes are those used in the present ontology but defined by another. For instance, the Cyclostratigraphy includes one technique "Filtering" in the "Time Series Analysis", whose definition was originally given in the Mathematical Geosciences. Consequently, "Filtering" is a defined node in the ontology of Mathematical Geosciences, and a cited one in the stratigraphic ontology. Owing to coordination of DDE, some concepts in the Stratigraphy Ontology could be cited from ontologies of other geological domains, e.g., Paleontology, Sedimentology, Geochronology, etc. Moreover, auxiliary nodes are those without rigorous scientific definitions, but necessary for connecting nodes in the hierarchy. Oftentimes, their names are self-explanatory. For instance, in the "Elemental Chemostratigraphy" branch in the Chemostratigraphy Ontology, an auxiliary node "Basic concepts of Element Chemostratigraphy" was created as the father node of "Geochemical fingerprinting", "Geochemical signature", "Excursion", etc. to better organize the hierarchical structure.

2.3. Define properties of classes

Three types of properties were further discriminated: Object Property indicates relation between instances of classes; Datatype Property describes the detailed content of the instances; Annotation Property annotates metadata. Defined nodes share some common properties, e.g., "label, Definition, Equivalent to, References, Subject" and may be distinguished by more specific ones explaining unique features, e.g., index fossils, measured values, stratigraphic sections, boundaries, geological ages etc. In contrast, cited nodes often only contain limited properties denoting the source of their definition. For instance, the "Lithology" node in the ontology of Lithostratigraphy was labeled as "[Sedimentology]. Lithology – 55", which suggests a link to the Defined node "Lithology" in the ontology of Sedimentology (Fig. 2). Properties may include subproperties as shown by examples in the ontology of Biostratigraphy below (Fig. 3). Moreover, some well-constructed databases were cross-referenced to help defining the properties, as they already include voluminous factual instances described by diverse properties, e.g., OneStratigraphy (https://onestratigraphy.ddeworld.org/), MacroStrat (https://macrostrat.org/) and GEOLEX (Stamm et al., 2000).

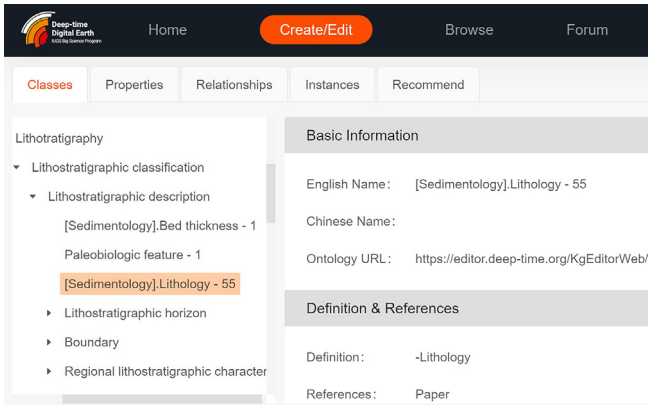

Fig. 2. An example of the cited node "Lithology" in the ontology of Lithostratigraphy.

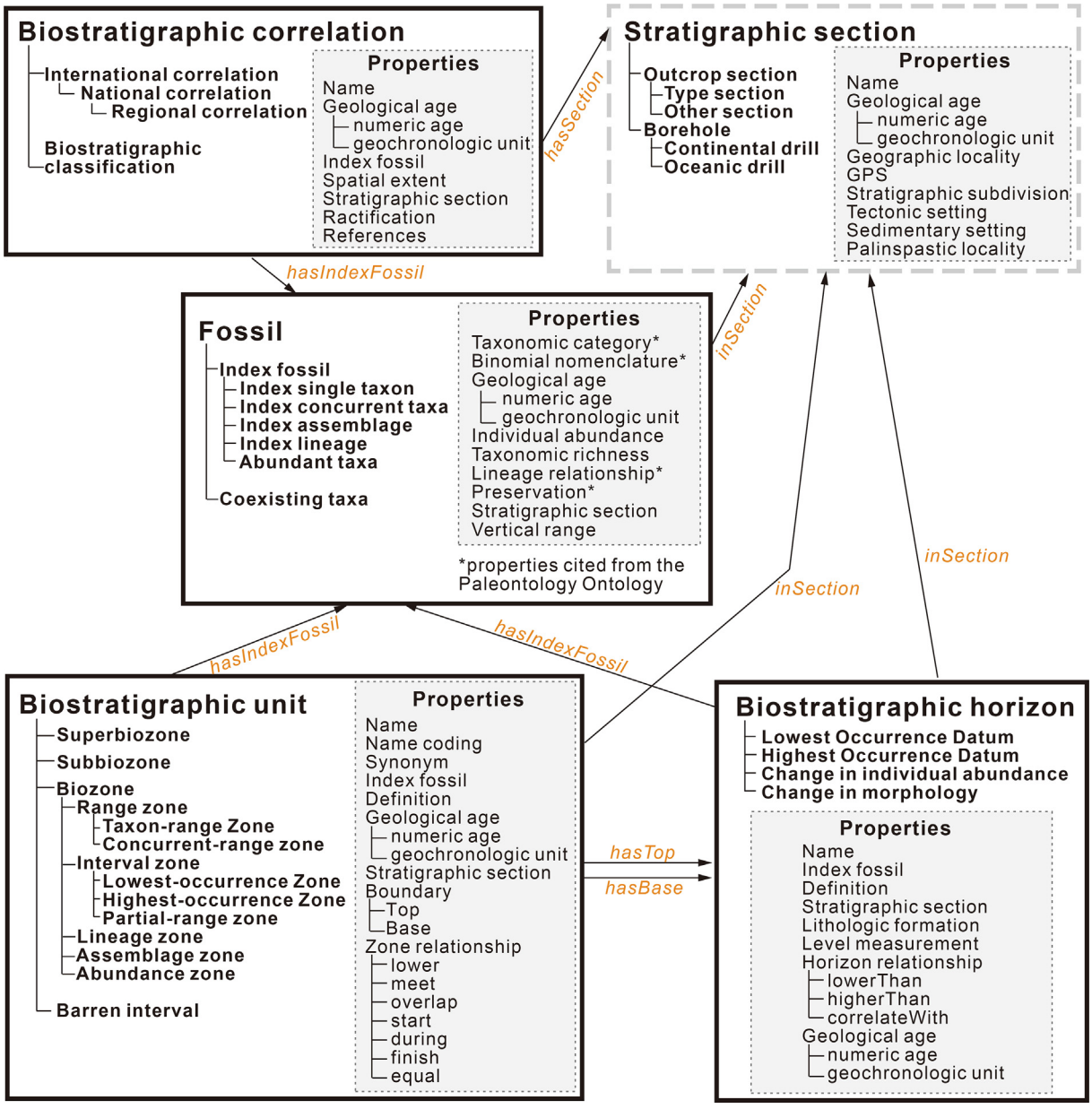

Fig. 3. The framework of the Biostratigraphy Ontology and the branch of Stratigraphy Section, showing key classes and properties.

2.4. Address semantic heterogeneity

Semantic heterogeneity is one major issue to solve for ontology construction. This issue exists not only between the seven stratigraphic ontologies constructed here, but also among all ontologies of 18 geological domains in GPKGs. This is because the diverse geological domains are especially characterized by a large corpus of sophisticated concepts and relationships on the regional or global scales during the evolved long history of geological ideas.

Semantic heterogeneity was addressed by professional opinions with the aid of GPKGs system. To avoid synonyms as much as possible, the latest definitions were always given priority during term extraction. If necessary, a synonym list could be added to each node for manual cross-checking in the GPKGs. In addition, the GPKGs could organize all uploaded nodes by the RDF format (Resource Description Framework, Brickley and Guha, 1999) and automatically detect semantic conflicts, so that only unique name for each node is allowed. Homonyms, if exist in different ontologies, have to be differentiated by adding postfix in the format as "name (branch subject)". For example, "Period" denotes "the interval of geologic time during which the rocks of the corresponding system were formed" in the ontology of Chronostratigraphy, but is "the time taken for one complete cycle of oscillation of a time series" in the ontology of Mathematical Geosciences. Thus, their IDs have to be modified as "Period (Stratigraphy)" and "Period (Mathematical Geosciences)" respectively.

3. Ontology models of Stratigraphy

The Stratigraphy Ontology is composed of eight branches. In addition to the seven stratigraphic disciplines, Stratigraphic Section was another root class as it physically preserves all stratigraphic data. In total, over 1200 nodes were constructed in the ontology of Stratigraphy. The nodes in the ontologies of Chronostratigraphy, Biostratigraphy and Lithostratigraphy are more repre-

sented by geological objects, such as fossils, stratigraphic units and boundaries (e.g., Lithological Formation as lithostratigraphic units and conformity or unconformity as boundaries in the Lithostratigraphy branch). Many of these nodes are also necessary for other four stratigraphic branches as well as ontologies of other geological domains. This allows the other four branches to focus more on concepts of their specific analytic methods, materials and results (see examples of Chemostratigraphy and Cyclostratigraphy in Fig. 1).

The seven subdiscipline ontologies are more or less dependent upon and reciprocal to each other, despite their emphasis on different aspects of stratigraphic data. For instance, geological age (Chronostratigraphy) is a universal property for all stratigraphic records, and determined by the combination of biozones (Biostratigraphy) and constraints from other subdisciplines (e.g. sequence boundary in Sequence Stratigraphy). Lithostratigraphic units (Lithostratigraphy) are fundamental source to extract varying stratigraphic signals, i.e., geochemical or geomagnetic excursion (Chemostratigraphy and Magnetostratigraphy), paleoclimate proxy (Cyclostratigraphy).

For the rest of this paper, we will focus on the elaboration of the ontology of Biostratigraphy due to limited space, while the other ontologies can be accessed via GPKGs. More importantly, the ontology construction for the biostratigraphy has seldom been attempted before as far as we know.

According to Murphy and Salvador (2000), the Biostratigraphy "deals with the distribution of fossils in the stratigraphic record and organization of strata into units based on their contained fossils". The ontology of Biostratigraphy constructed by us is, thus, founded on the following root classes: Biostratigraphic unit, Biostratigraphic horizon, Fossil, and Biostratigraphic correlation (Fig. 3) and coded in RDF format by Protégé (Fig. 4). The biostratigraphic units and horizons are fundamental elements and characterized by distribution of index fossils in specific stratigraphic sections. Besides, biostratigraphic correlation, whether regionally or globally, between sections in disparate localities was established as another class to embody the application of biostratigraphy.

Biozone is identified by the combination of index fossils, top and base boundaries etc. Biozone represents a subclass of Biostratigraphic unit and further comprises five subclasses: Range zone, Assemblage zone, Abundance zone, Lineage zone and Interval zone (Fig. 5). Except Interval zone, each of the other four is diagnosed by the known stratigraphic and geographic range, respectively, of the selected index taxa (one or two), index assemblage (three or more taxa in association), index lineage (a specific segment of an evolutionary lineage) and significantly abundant taxa. All these characteristic fossils constitute the subclass Index fossil in the class Fossil. In contrast, the interval zone does not emphasize the fossil content within it, but purely relied on the recognized biohorizons as its base and top. In addition, Barren Interval is one special subclass. It represents the strata devoid of fossils and sandwiched between two particular biohorizons, thus disjoint with all other subclasses of biostratigraphic units.

The top and base of biozones are identified as the biostratigraphic horizon (biohorizon), another essential concept in Biostratigraphy. Biohorizons are defined by significant changes in biostratigraphic features at a boundary or surface without rock thickness (Murphy and Salvador, 2000). The change oftentimes is the lowest or highest occurrence datum of index fossils in particular stratigraphic section (LOD and HOD), and less commonly the noticeable change of individual abundance, morphological characters of certain taxa, etc. (Table 1).

Recognition of biozones and biohorizons depends on the paleontological features of fossils and their occurrences in the strata. Therefore, these two core classes of Biozone and Biostratigraphic horizon are linked with another root class, i.e., Fossil. The fossil class has two groups of properties pertaining to the biostratigraphy: properties of taxonomy (e.g., taxonomic category, binominal nomenclature, lineage relationship, taxonomic richness) and those of concrete specimens (e.g., individual abundance, preservation state, and vertical range in the stratigraphic section). Fortunately, many of these properties (indicated by asterisk in Fig. 3) could be conveniently cited from the Paleontology Ontology in the GPKGs. Taking preservation for example, four main factual instances have been assigned to the Preservation Class in the Paleontology Ontology: biocoenosis, thanatocoenosis, reworked and infiltrated. Then we introduced the "hasPreservationState" property to describe the particular preservation state of the fossil class. Among them, reworked and infiltrated are not reliable biostratigraphic index, as they represent fossil redeposited into the sediments either younger or older than the life span of the fossils.

One key goal of biostratigraphic work is the subdivision and sequence of stratigraphic units in a single section (Biostratigraphic classification) and further demonstrating correspondence between such units among sections in separate areas (regional, national or international Biostratigraphic correlation). These activities arrange rocks in relative temporal sequence and allow us to explore the deep-time events in consistent geochronological order from a broad perspective. Two properties i.e., "Zone relationship" and "Horizon relationship" are designed in our ontology to determine,

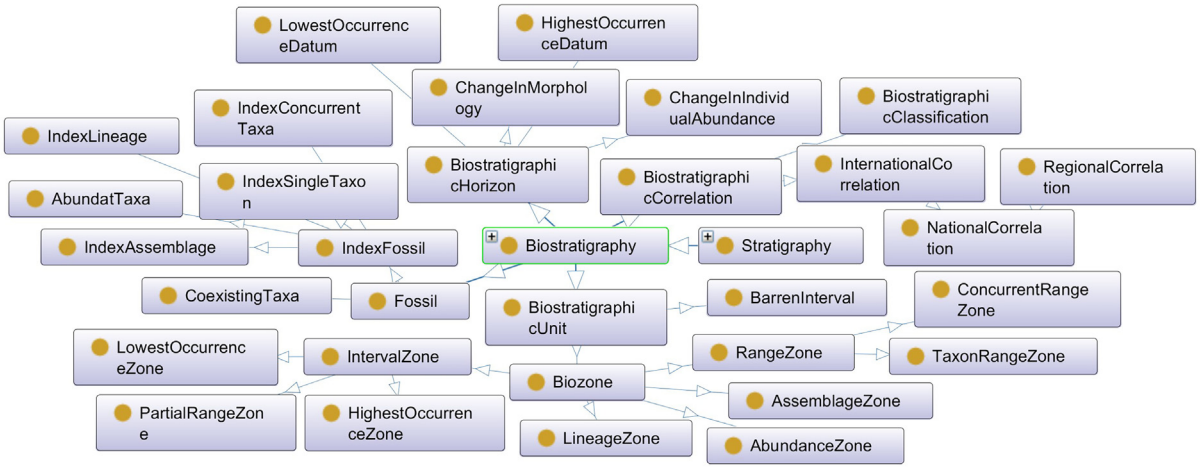

Fig. 4. The results of implementation and visualization of the Biostratigraphy Ontology by the Protégé program.

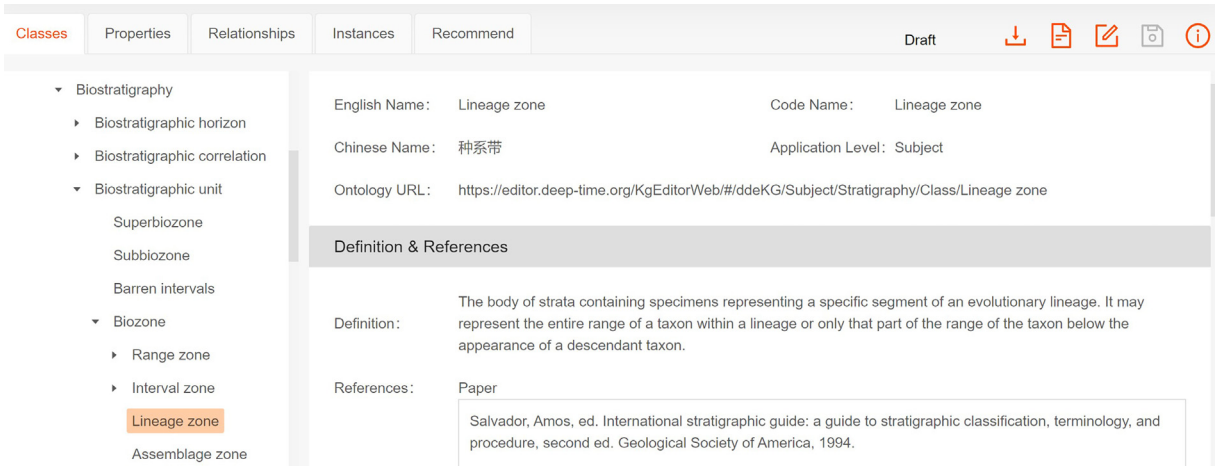

Fig. 5. The knowledge nodes of the Stratigraphy Ontology as uploaded to the Geoscience Professional Knowledge Graphs: the Lineage zone in the Biostratigraphy Ontology as a case.

Table 1
Main Object Properties in the Biostratigraphy Ontology.

Defined Property	Domain	Range
hasIndexFossil	Biozone/Biohorizon	Index single taxon, Index concurrent taxa, Index assemblage, Index lineage, Abundant taxa
hasTop/hasBase	Biozone	FAD, LAD, LOD, HOD, change in individual abundance or morphology, etc.
hasPreservationState	Fossil	biocoenosis, thanatocoenosis, reworked, infiltrated etc.
hasAge	almost all geological classes	Geological age either numeric age or geochronologic unit
inSection	Biozone/Biohorizon/Fossil	Specific stratigraphic section (e.g. the Meishan section in South China)
lower/equal/meet/overlap/start/contain	Biozone	–

respectively, the sequence of biozones and biohorizons. Three types of relation between biohorizons, i.e., higher, lower and correlateWith, are straightforward, while the relationships between biozones are more complex, especially when biozones defined by disparate fossil categories are considered. The 13 relationships between temporal intervals summarized by Allen (1983) were adopted to describe the relationship between biozones (Fig. 6). Besides three relations just mentioned above, one stratigraphic zone may meet, start, finish, contain or overlap another one. The relationship meet indicates that the top of a zone is same as

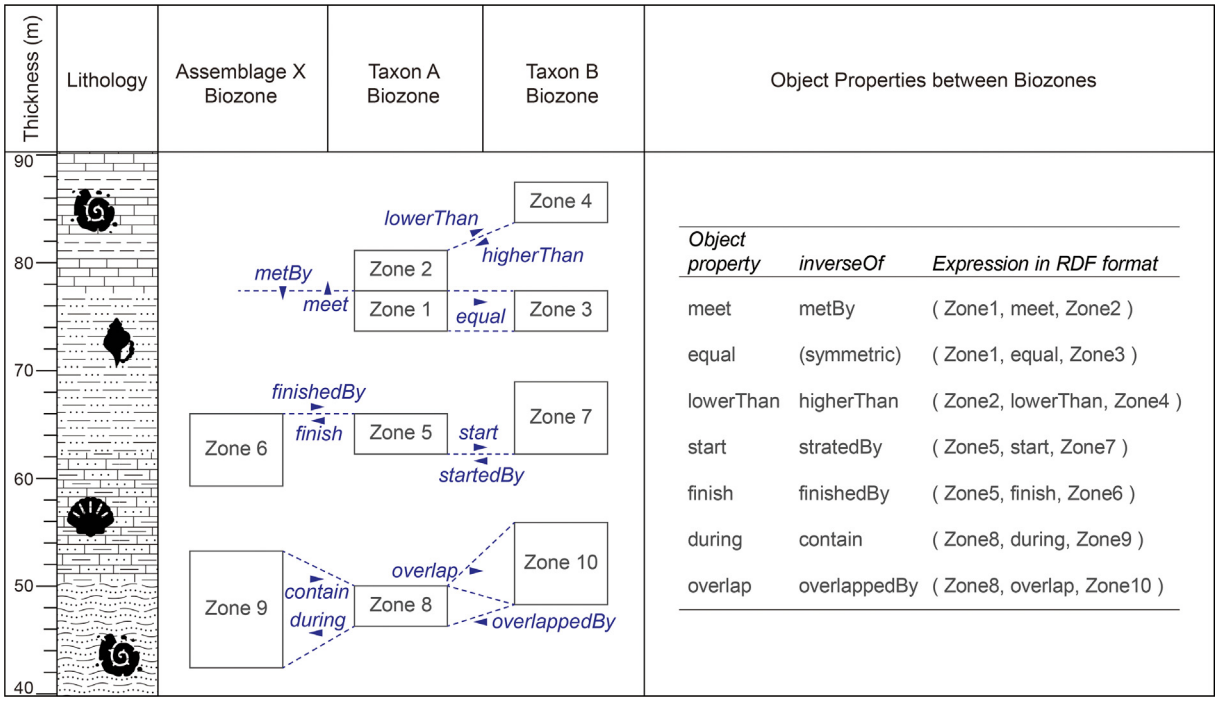

Fig. 6. The stratigraphic relationships between biostratigraphic units in the Biostratigraphy Ontology.

(OWL: sameAs) the base of another zone. This is the most common scenario in successive zonation based on one and the same taxon in the stratigraphic section. The property "start" or "finish" means two zones share the same base or top, respectively. Besides, the stratigraphic range of one zone may contain that of another zone. Finally, overlap indicates that one zone has its top located within another zone for comparison, while its base lower than that of the compared zone (Zone 8 and 10 in Fig. 6). It is comprehensible that all these five properties have their corresponding inverse properties (owl: inverseOf) (e.g., metBy, startedBy, finishedBy, during, overlappedBy).

All root classes in the ontology of Biostratigraphy share a common property "Geological age". This property is described by two subproperties: numeric age and geochronologic unit. These two are significant concepts, respectively, in the Geochronology and Chronostratigraphy, and have been well defined by previous works mentioned in the Introduction Section. In our work, they are expressed by the citation, respectively, of NumericEraBoundary and GeochronologicEra as explained by Cox and Richard (2015) based on the iSO-19108 (Temporal Schema).

4. Application for semantic search

Based on the Stratigraphy Ontology, the Knowledge Graph of Stratigraphy could be progressively fulfilled by integrating instances from versatile sources, more conveniently from well-established stratigraphic databases. Then, the semantic searching or reasoning by machines would become feasible. As a preliminary example, a prototype of semantic search specialized for stratigraphic references is designed (Fig. 7). The constructed ontologies by this study are probably heretofore the most comprehensive for stratigraphic knowledge and all seven ontologies are interconnected by cited nodes. Only by such scope and liaison can a search engine, in ways that mimic human intelligence, infer the association of instances with related semantic meaning from different stratigraphic subdisciplines.

To query a particular keyword in the contrived engine, any of the seven stratigraphic subdisciplines can be further specified to decide the scope for search. The returned results would include not only the references that directly contain the keyword, but also, according to the constructed ontologies, (1) the related subclasses of this keyword; (2) the concepts equivalent, in terms of geological age, to this keyword and those subclasses. For instance, when the geochronological term "Paleogene" was searched among limited instances exported from the OneStratigraphy database, returned results include references exactly containing, in the title or abstract, the keyword and concepts of this time interval (foraminiferal zone, lithologic formation) (the middle column in the Fig. 7). Meanwhile, a knowledge graph is accompanied to visually illustrate the connection of these concepts. According to the Chronostratigraphy Ontology, the Paleocene and Eocene are subclasses of Paleogene. The Eocene, in turn, is characterized by a significant carbon isotope excursion at the P/E boundary. The search results also include a foraminiferal *Morozovella angulatai* zone, the Ensa Shale Formation and Polarity Chron C21r, all dated to be Paleogene.

5. Conclusions

We constructed the Stratigraphy Ontology of seven stratigraphic subdisciplines, i.e., Chronostratigraphy, Biostratigraphy, Lithostratigraphy, Chemostratigraphy, Cyclostratigraphy, Magnetostratigraphy and Sequence Stratigraphy. The results probably represent thus far the most comprehensive ontology of the stratigraphic domain, including over 1200 nodes describing essential

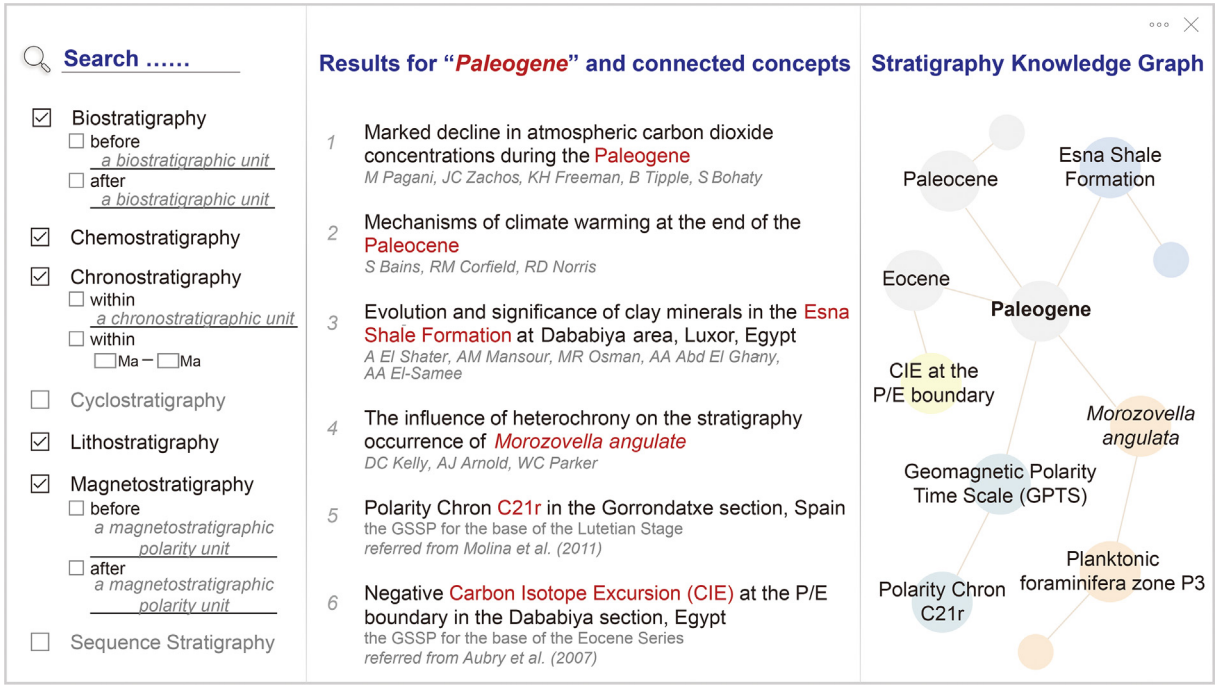

Fig. 7. A prototype of semantic search for querying stratigraphic literatures based on the constructed ontology of Stratigraphy.

concepts and their properties in total. Except for Lithostratigraphy and Chronostratigraphy, the ontologies of other stratigraphic domains have not been established before. Especially, the ontology of Biostratigraphy is contrived based on the root classes of Biostratigraphic unit, Biostratigraphic horizon, Fossil and Biostratigraphic correlation. All these ontologies are connected by cited nodes not only with each other, but also with ontologies of other major geological domains, e.g., Paleontology, Geochronology etc., in the framework of DDE Geoscience Professional Knowledge Graphs. A prototype of semantic search engine focusing on stratigraphic literatures was also designed by utilizing the semantic liaison among different subdisciplines in our constructed ontologies.

The ontology of the stratigraphic domain is not the ultimate goal in itself, but a kicking off for transition of the geological research paradigm towards big-data methodology. Further efforts need to be devoted to incorporate massive instances or factual data and eventually fulfill a comprehensive Stratigraphy Knowledge Graph. Currently, we are still in the initial stage of manually aggregating rather limited instances from well-known databases and published literature. We hope this process would be assisted by applying advanced Nature Language Processing (NLP) technology in the near future. The Stratigraphy Ontology in this study would undoubtedly be dynamically evaluated, debugged and updated according to the feedbacks from stratigraphic experts and other geologists during the application of these ontologies for specific tasks.

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.

Acknowledgements

We appreciate the constructive reviews from two anonymous reviewers and the editor, which significantly improved the initial manuscript. This research is supported by the National Natural Science Foundation of China (Grant No. 41725007), National Key R&D Program of China (Grant No. 2018YFE0204201), Fundamental Research Funds for the Central Universities (0206-14380121), and Frontiers Science Center for Critical Earth Material Cycling Fund (JBGS2101). This paper contributes to the "Deep-time Digital Earth" Big Science Program.

References

Allen, J.F., 1983. Maintaining knowledge about temporal intervals. Commun. ACM 26, 832–843. https://doi.org/10.1145/182.358434.

Brickley, D., Guha, R.V., 1999. Resource description framework (RDF) schema specification. Proposed Recommendation, World Wide Web Consortium. http://www.w3.org/TR/PR-rdf-schema (accessed 7 September 2021).

Carbonera, J.L., Abel, M., Scherer, C.M.S., 2015. Visual interpretation of events in petroleum exploration: An approach supported by well-founded ontologies. Expert Syst. Appl. 42, 2749–2763. https://doi.org/10.1016/j.eswa.2014.11.021.

Cohen, K.M., Finney, S.C., Gibbard, P.L., Fan, J.X., 2013. The ICS international chronostratigraphic chart. Episodes 36, 199–204.

China Commission of Stratigraphy, 2017. Stratigraphic Guide of China, 2016 ed. Geological Publishing House, Beijing, pp. 62 (in Chinese).

Cox, S.J.D., 2016. Time ontology extended for non-Gregorian calendar applications. Semantic Web 7, 201–209. https://doi.org/10.3233/SW-150187.

Cox, S.J.D., Little, C., 2020. Time ontology in OWL. https://www.w3.org/TR/owl-time (accessed 7 September 2021).

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

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

Dong, S.C., Yin, H.W., Xu, G., 2010. Heterogeneous data searching based on Geologic Time Ontology. J. Geo-inform. Sci. 12, 194–199 (in Chinese with English abstract).

Ehrlinger, L., Wöß, W., 2016. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS) 48, 1–4.

Feilmayr, C., Wöß, W., 2016. An analysis of ontologies and their success factors for application to business. Data Knowl. Eng. 101, 1–23. https://doi.org/10.1016/j.datak.2015.11.003.

Gradstein, F.M., Ogg, J.G., Schmitz, M.D., Ogg, G.M., 2012. The Geologic Time Scale 2012. Elsevier, Oxford, pp. 1–1144.

Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum. Comput. Stud. 43, 907–928. https://doi.org/10.1006/ijhc.1995.1081.

Hou, Z.W., Zhu, Y.Q., Gao, X., Luo, K., Wang, D.X., Sun, K., 2015. A Chinese Geological Time Scale Ontology for geodata discovery. In: Proc. 23rd International Conference on Geoinformatics, pp. 1–5. https://doi.org/10.1109/GEOINFORMATICS.2015.7378648.

Hou, Z.W., Zhu, Y.Q., Gao, Y., Song, J., Qin, C.Z., 2018. Geologic Time Scale Ontology and its applications in semantic retrieval. J. Geo-inform. Sci. 20, 17–27. https://doi.org/10.12082/dqxxkx.2018.170328 (in Chinese with English abstract).

Ji, Y.L., 2005. Sequence Stratigraphy. Tongji University Press, Shanghai, pp. 1–203 (in Chinese).

Keller, G.R., 2003. GEON (GEOscience Network): A first step in creating cyberinfrastructure for the Geosciences. Seismol. Res. Lett. 74, 441–444. https://doi.org/10.1785/gssrl.74.4.441.

Ma, X.G., Carranza, E.J.M., Wu, C.L., van der Meer, F.D., Liu, G., 2011. A SKOS-based multilingual thesaurus of geological time scale for interoperability of online geological maps. Comput. Geosci. 37, 1602–1615. https://doi.org/10.1016/j.cageo.2011.02.011.

Ma, X.G., Fox, P., 2013. Recent progress on geologic time ontologies and considerations for future works. Earth Sci. Inform. 6, 31–46. https://doi.org/10.1007/s12145-013-0110-x.

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

Ma, X.G., Ma, C., Wang, C.B., 2020. A new structure for representing and tracking version information in a deep time knowledge graph. Comput. Geosci. 145,. https://doi.org/10.1016/j.cageo.2020.104620 104620.

Murphy, M.A., Salvador, A., 2000. International subcommission on stratigraphic classification of IUGS international commission on stratigraphy: international stratigraphic guide—an abridged version. GeoArabia 5, 231–266. https://doi.org/10.2113/geoarabia0502231.

Musen, M.A., 2015. The Protégé project: a look back and a look forward. AI Matters 1 (4), 4–12. https://doi.org/10.1145/2757001.2757003.

North American Commission on Stratigraphic Nomenclature, 2005. North American Stratigraphic Code. AAPG Bull. 89 (11), 1547–1591. https://doi.org/10.1306/07050504129.

Noy, N.F., McGuinness, D.L., 2001. Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05 and SMI-2001-0880, Stanford Knowledge Systems Laboratory and Stanford Medical Informatics. http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf (accessed 7 September 2021).

Opdyke, M.D., Channell, J.E.T., 1996. Magnetic Stratigraphy. Academic Press, London, pp. pp.1–346.

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, 81–96. https://doi.org/10.1007/s12145-011-0080-9.

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

Salvador, A., 1994. International Stratigraphic Guide: a guide to stratigraphic classification, terminology, and procedure, second ed. International Union of Geological Sciences and Geological Society of America, Colorado, 214 pp.

Sinha, A.K., 2006. Geoinformatics: Data to Knowledge. Geological Society of America, Colorado, pp. pp. 1–282. https://doi.org/10.1130/SPE397.

Stamm, N.R., Wardlaw, B.R., Soller, D.R., 2000. GEOLEX-The National Geologic Map Database's Geologic Names Lexicon. Open-file Report 00-325, US Geological Survey. https://pubs.usgs.gov/of/2000/of00-325/stamm.html (accessed 7 September 2021).

Strasser, A., Hilgen, F.J., Heckel, P.H., 2006. Cyclostratigraphy-concepts, definitions, and applications. Newsl. Strat. 42 (2), 75–114. https://doi.org/10.1127/0078-0421/2006/0042-0075.

Studer, R., Benjamins, V.R., Fensel, D., 1998. Knowledge engineering: Principles and methods. Data Knowl. Eng. 25, 161–197.

Uschold, M., Gruninger, M., 1996. Ontologies: Principles, methods and applications. Knowl. Eng. Rev. 11 (2), 93–136. https://doi.org/10.1017/S0269888900007797.

Wang, C.B., Ma, X.G., Chen, J.G., 2018. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. Comput. Geosci. 115, 12–19. https://doi.org/10.1016/j.cageo.2018.03.004.

Weedon, G.P., 2003. Time-series Analysis and Cyclostratigraphy: Examining Stratigraphic Records of Environmental Cycles. Cambridge University Press, Cambridge, pp. pp. 1–259.

Wu, L.X., Xu, L., Che, D.F., 2008. Stratum-ontology and its application in borehole data integration. Geomat. Inform. Sci. Wuhan Univ. 33, 144–148 (in Chinese with English abstract).

Yuan, M., Li, S.R., Liu, X.Y., 2021. Research on standardized model of geological knowledge graph. J. Jilin Univ. 39, 215–222 (in Chinese with English abstract).

Zheng, R.C., Wen, H.G., Li, F.J., 2010. High-resolution Sequence Stratigraphy. Geological Publishing House, Beijing, 354 pp. (in Chinese).

Zhou, C.H., Wang, H., Wang, C.S., Hou, Z.Q., Zheng, Z.M., Shen, S.Z., Cheng, Q.M., Feng, Z.Q., Wang, X.B., Lv, H.R., Fan, J.X., Hu, X.M., Hou, M.C., Zhu, Y.Q., 2021. Geoscience knowledge graph in the big data era. Sci. China Earth Sci. 64, 1105–1114. https://doi.org/10.1007/s11430-020-9750-4.