Journal of Global Change Data & Discovery. 2024, 8(2): 111-124
www.geodoi.ac.cn

Data Thesaurus of Earth Surface System Sciences

Chen, J.[1] Wang, S.[2, 3*] Zhu, Y. Q.[2, 3] Duan, F. Z.[1] Wang, B.[4*]

1. College of Resources Environment and Tourism, Capital Normal University, Beijing 100048, China;
2. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;
3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China;
4. Command Center for Comprehensive Survey of Natural Resources, China Geological Survey, Beijing 100055, China

Abstract: The earth surface system scientific data thesaurus is a vocabulary that in a standard format describes the terminology of various spheres within the earth surface system and basic semantic relationships among them. As a fundamental data resource, a high-quality thesaurus facilitates concept differentiation and comparison, information organization and retrieval, and data standardization and sharing, thereby advancing interdisciplinary research on the earth surface system. Based on domain-specific thesauri (e.g., the Global Change Master Directory), authoritative domain-specific literature (e.g., geographical dictionaries), domain ontologies (e.g., the Sematic Web for Earth and Environment Terminology), and online resources (e.g., Wikipedia), covering subject headings of the earth surface system in global change, geographic environment, resource systems and other areas, this study clarifies the definition and scope of the earth surface system. We manually constructed a high-quality earth surface system scientific data thesaurus that includes seven layers of concepts: basic space, spheres, systems, subsystems, objects, elements, and attributes, encompassing a total of 3,463 subject headings. Additionally, it describes equivalent, hierarchical, and related relationships among the terms, totaling 4,454 relationships. Researches indicate that thesaurus performs well in terms of scale and functionality, promising to provide data support for network construction, information alignment, information retrieval, knowledge services, and knowledge discovery in the field of earth surface system science. The dataset is archived in. xlsx format and consists of three data files, with a data size of 1.84 MB (compressed into one file, 1.78 MB)

Keywords: earth surface system; scientific data; subject headings; thesaurus; ontology model; knowledge services

DOI: https://doi.org/10.3974/geodp.2024.02.01

CSTR: https://cstr.escience.org.cn/CSTR:20146.14.2024.02.01

Received: 08-04-2024; **Accepted:** 10-06-2024; **Published:** 25-06-2024

Foundations: Ministry of Science and Technology of P. R. China (2022YFF0711601, 2022YFB3904201); National Natural Science Foundation of China (42101467); LREIS (KPI009)

***Corresponding Author:** Wang, S., Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, wangshu@igsnrr.ac.cn; Wang, B., Command Center for Comprehensive Survey of Natural Resources, China Geological Survey, wangbincgs@mail.cgs.gov.cn

Data Citation: [1] Chen, J., Wang, S., Zhu, Y. Q., *et al*. Data thesaurus of earth surface system sciences [J]. *Journal of Global Change Data & Discovery*, 2024, 8(2): 111–124. https://doi.org/10.3974/geodp.2024.02.01. https://cstr.escience.org.cn/CSTR:20146.14.2024.02.01.

[2] Chen, J., Wang, S., Zhu, Y. Q., *et al*. Thesaurus of scientific data for the earth surface system [J/DB/OL]. *Digital Journal of Global Change Data Repository*, 2024. https://doi.org/10.3974/geodb.2024.07.10.V1. https://cstr.escience.org.cn/CSTR:20146.11.2024.07.10.V1.

Dataset Availability Statement:

The dataset supporting this paper was published and is accessible through the *Digital Journal of Global Change Data Repository* at: https://doi.org/10.3974/geodb.2024.07.10.V1 or https://cstr.escience.org.cn/CSTR:20146.11.2024.07.10.V1.

1 Introduction

The earth surface not only focuses on the spatial geographical features of the Earth's surface but also on the interactions between living beings and natural environments, making it a core subject of geographical research[1, 2]. With the development of geography, scholars have gradually recognized that the Earth's surface is an open and complex mega-system with material and energy exchanges[3]. The earth surface system reveals the interactions and impacts among fundamental elements such as climate, biology, water, geology, and soil, as well as their evolution and development over time and space. Meanwhile, the diverse, heterogeneous, voluminous, and complex spatiotemporal knowledge generated with the evolution and development of the various elements within the mega-system drives the earth surface system towards a data-intensive science[4, 5]. Managing and utilizing scientific data from the earth surface system, such as climate change metrics, extreme disaster event forecasts, and ecological environment monitoring, is crucial for better resource management, environmental sustainability maintenance, and natural disaster prediction[6, 7].

A thesaurus, a vocabulary of descriptors used for describing and classifying specific concepts or terms within a field, is an important tool for information organization and retrieval in information management[8]. In research related to the Earth's surface, the construction of thesauri has accumulated a certain foundation in both single-disciplinary and interdisciplinary studies. For example, the Geographical Science Thesaurus[9] covers technical terms in areas like natural sciences, humanities, and regional geography; the Chinese Thesaurus of Geology[10, 11] focuses on descriptors about rocks and minerals and geological structures; the Environmental Science Thesaurus[12] includes specific terms for retrieval in the field of environmental science. There are also comprehensive thesauri that cover geoscience-related terms, such as the Chinese Classified Thesaurus[13] which involves disciplinary and thematic concepts across natural sciences, and the NASA Thesaurus[14] which focuses on natural space sciences while also covers Earth sciences. These thesauri encompass basic geography, resource environment, geological geomorphology and other fields that are related to geoscience. However, none of these thesauri fully cover the core topics of the earth surface system studies on the own. Variations in how different thesauri interpret the same concept make it challenging to share data across thesauri, highlighting a lack of a unified, standardized knowledge system for the earth surface system. In summary, existing thesauri, whether single-disciplinary or interdisciplinary, face issues such as inconsistent concept definitions and an inability to fully cover the core concepts of the earth surface system science. Currently, there's no comprehensive, complete, and accurate thesaurus for earth surface system scientific data.

Therefore, constructing the earth surface system scientific data thesaurus helps better organize key objects, concepts, and their interrelations covered in the earth surface system field, providing a convenient way for organizing, storing, and utilizing earth surface system scientific data. To address the abovementioned issues, this paper manually constructs a high-quality earth surface system scientific data thesaurus, aiming to provide data support for network construction, information association and alignment, information retrieval, knowledge services, and knowledge discovery in the field of the earth surface system science.

2 Metadata of the Dataset

The metadata information of earth surface system scientific data thesaurus[15] is summarized in Table 1. It includes the dataset full name, short name, authors, data format, data size, data files, data publisher, and data sharing policy, etc.

Table 1 Metadata summary of the thesaurus of scientific data for the earth surface system

Items	Description
Dataset full name	Thesaurus of scientific data for the earth surface system
Dataset short name	ESSSD_Thesaurus
Authors	Chen, J., Capital Normal University, cj15160172956@163.com; Wang, S., Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, wangshu@igsnrr.ac.cn Zhu, Y. Q., Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, zhuyq@igsnrr.ac.cn Duan, F. Z., Capital Normal University, duanfuzhou@263.net Wang, B., Command Center for Comprehensive Survey of Natural Resources, China Geological Survey, wangbincgs@mail.cgs.gov.cn
Data format	.xlsx
Data size	1.84 MB, compressed to 1.78 MB
Data files	Thesaurus terms in Chinese and English, descriptions in Chinese and English, relationships between terms, term classification, data sources
Foundations	Ministry of Science and Technology of P. R. China (2022YFF0711601, 2022YFB3904201); National Natural Science Foundation of China (42101467); LREIS (KPI009)
Data publisher	Global Change Research Data Publishing & Repository, http://www.geodoi.ac.cn
Address	No. 11A, Datun Road, Chaoyang District, Beijing 100101, China
Data sharing policy	(1) *Data* are openly available and can be free downloaded via the Internet; (2) End users are encouraged to use *Data* subject to citation; (3) Users, who are by definition also value-added service providers, are welcome to redistribute *Data* subject to written permission from the GCdataPR Editorial Office and the issuance of a *Data* redistribution license; and (4) If *Data* are used to compile new datasets, the 'ten per cent principal' should be followed such that *Data* records utilized should not surpass 10% of the new dataset contents, while sources should be clearly noted in suitable places in the new dataset[16]
Communication and searchable system	DOI, CSTR, Crossref, DCI, CSCD, CNKI, SciEngine, WDS, GEOSS, PubScholar, CKRSC

3 Methods

The establishment of earth surface scientific data thesaurus adopts a strategy that combines "top-down" and "bottom-up" approaches, ensuring the comprehensiveness and professionalism of the thesaurus by integrating authoritative domain-specific dictionaries, monographs, ontologies, and online resources from multiple sources. At the same time, by designing a hierarchical structural framework and semantic relationships, it achieves effective organization and standardization of earth surface system scientific data to support data analysis, application, and sharing. This section elaborates on the methods used to construct the earth surface system scientific data thesaurus.

3.1 Construction of the Data Thesaurus of Earth Surface System Sciences

The data thesaurus construction of the earth surface system sciences involves a combined "top-down" and "bottom-up" approach as illustrated in Figure 1. First, with the consideration of the basic and general application requirements of the thesaurus, such as earth surface system scientific data sharing services and emergency disaster knowledge services, the scope and content of earth surface system scientific data are defined. Second, terms and concepts of earth surface system scientific data are collected and analyzed through manual collection and web crawlers from various data sources, including authoritative domain-specific dictionaries and comprehensive online resources to establish a corpus of related subject headings and lists of potential terms for the thesaurus. Then a, overall structural framework of the thesaurus is established by the "top down" approach from a macro perspective. This involves determining the overall compilation structure, classification system, and hierarchical structure of the thesaurus. Following the structural framework, basic semantic relationships between candidate terms are designed, encompassing equivalence, hierarchical, and associative relationships. Finally, the subject headings are identified among the candidate terms through a "bottom-up" approach.

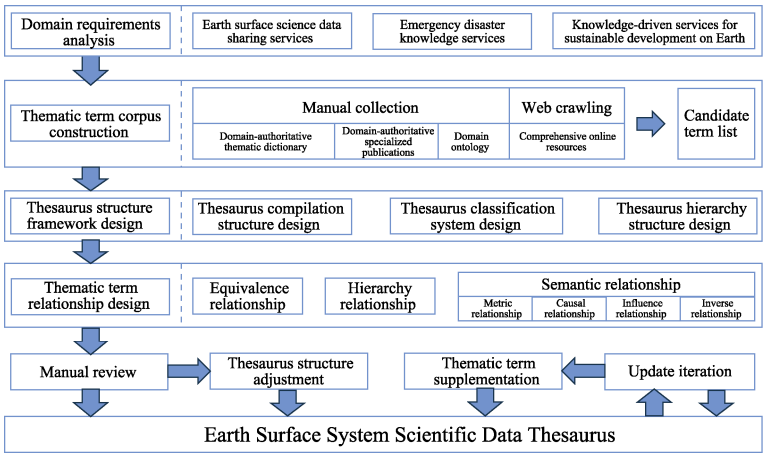

Figure 1 Technical roadmap for the data thesaurus construction of the earth surface system sciences

Through manual review, the structure of the thesaurus is adjusted, refining the categories and hierarchy of subject headings. This iterative process ensures the continuous updating and supplementation of the thesaurus.

3.2 Definition and Scope of the Earth Surface System

The establishment of a comprehensive data thesaurus of earth surface system sciences serves as the foundation for the analysis, application, sharing, and knowledge services of earth surface system data. The accurate definition of the concepts, connotations, and scope of the earth surface system is crucial for the classification and content construction of the thesaurus.

The diverse perspectives of modern scientists on the concept and research scope of the earth surface system hold profound significance in defining the coverage scope of the data thesaurus of earth surface system science. German geographer Richthofen proposed the concept of the "earth surface" in 1883, and Soviet geographer Пётр Иванович Броунов further defined it as concentric layers in 1910[17, 18]. With the emergence of significant theories about continental drift[19], seafloor spreading[20], plate tectonics[21], and the Gaia hypothesis[22], researchers gained a deeper understanding of the connotations and scope of the earth surface system, as shown in Table 2. From a geographical perspective, the earth surface system is considered as a coupled whole encompassing the Earth's various layers around human activities[18, 23]. From a macro perspective in earth science, the earth surface system is viewed as a complex system for the exchange of energy and matter both internally and externally[24, 25]. From an ecological perspective, the earth surface system is defined as a geographical spatial carrier supplying human and ecosystem needs[26]. From the perspective of natural resources, the earth surface system is seen as the core space providing fundamental living conditions for human production and life[25].

The divergence in the academic community regarding the earth surface system primarily concerns the delineation of boundaries, specifically the accurate definition of the lower boundary within the lithosphere) and the upper boundary within the atmosphere). Despite various emphases on how to understand the earth surface system from different perspectives, there is a general consensus on several points. First, in the fundamental understanding of the earth surface system, it is acknowledged as a complex organic system composed of interconnected and notably dynamic layers characterized by frequent material-energy information circulation. Second, it is agreed that the core layers of the earth surface system, from bottom to top, encompass the lithosphere (partial), pedosphere, biosphere, anthroposphere, hydrosphere, and atmosphere (partial). Last, it is recognized that the earth surface system is the most active realm of layer interactions and human activities, providing continuous support for humanity.

Table 2 Different descriptions of the scope of the earth surface system by different researchers

No.	Spheres included	Description of earth surface system scope	Reference
1	Atmosphere, hydrosphere, biosphere, lithosphere	From the top of the troposphere to the surface of the geosphere and the depth of the oceans	[17, 18, 23, 27]
2	Atmosphere, hydrosphere, pedosphere, lithosphere	The near-surface realm from the solid phase of the subsurface to the mobile phase within the operational orbit height of artificial Earth satellites	[28]
3	Atmosphere, hydrosphere, biosphere, anthroposphere, noosphere, pedosphere, lithosphere	Mutually permeable layered concentric spheres consisting of the atmosphere, hydrosphere, biosphere, anthroposphere, noosphere, and pedosphere	[29]
4	Atmosphere, hydrosphere, biosphere, anthroposphere, pedosphere, lithosphere	A complex open mega-system formed by interactions of the atmosphere, biosphere, anthroposphere, hydrosphere, pedosphere, and lithosphere	[24, 30–34]
5	Atmosphere, hydrosphere, biosphere, lithosphere, pedosphere	From the outermost layer of the atmosphere to the asthenosphere, including the lithosphere, hydrosphere, atmosphere, biosphere, and near-surface physical fields on and under the ground	[17]
6	Atmosphere, hydrosphere, biosphere, anthroposphere, lithosphere, centrosphere, celestial bodies	Including the troposphere, hydrosphere, land structures, as well as the biosphere and anthroposphere interacting with these layers, with the coupling of the air, water, and shell systems, along with extraterrestrial and intratelluric dynamic actions as the core of researches	[35]
7	Atmosphere, cryosphere, hydrosphere, anthroposphere, land, lithosphere	Interactions among the atmosphere, cryosphere, land, ocean, and lithosphere, covering physical, chemical, and biological processes, with human activities as part of the system's functionality	[25]

In conclusion, this paper defines the earth surface system as an open, complex mega-system with the atmosphere, hydrosphere, biosphere, anthroposphere, pedosphere, and lithosphere as its research objects. It involves elements such as the atmosphere, water bodies, ecosystem, human society, and geological structures, interacting with each other to form a mutually dependent, ever-changing holistic system.

3.3 Data Sources of the Data Thesaurus of Earth Surface System Sciences

Considering the scientific rigor and precision required by the earth surface system science, the vocabulary for the data thesaurus of earth surface system sciences primarily originates from the four types of sources: authoritative domain-specific subject headings dictionaries, authoritative monographs, domain ontologies, and comprehensive online resources (Table 3). Authoritative domain-specific subject headings dictionaries are maintained by authoritative institutions or organizations in the field of earth science, containing terms or concepts that have undergone professional review and approval, contributing to ensuring the professionalism and accuracy of the thesaurus. Authoritative monographs authored by domain experts cover rich scientific knowledge and terminology, providing reliable background information and specialized vocabulary for the thesaurus. Domain ontologies of the earth surface system represent the formalized knowledge of domain-specific concepts and relationships, aiding in a better understanding of the knowledge structure within the field and facilitating the establishment of the classification and hierarchy of terms. There are extensive comprehensive resources on the internet regarding the earth surface system, offering broad background information and classification indexes. These resources can be utilized for cross-validation, supplementation, and enrichment of the thesaurus conceptual content with other data sources, thereby enhancing the quality and coverage of the thesaurus. By integrating these data sources, the comprehensiveness, accuracy, and adaptability of the thesaurus can be ensured, providing robust support for the organization and standardization of the data related to the earth surface system sciences.

3.4 Structural Framework for the Data Thesaurus of Earth Surface System Sciences

3.4.1 Compilation Structure Design

A comprehensive thesaurus consists of a main table and auxiliary tables[43]. The main table is the core component of the thesaurus, organized in a specific order, such as alphabetical order in

Table 3 Data sources of the data thesaurus of earth surface system sciences

Data source type	Data source	Research domain covered by the data source
Authoritative domain-specific subject headings dictionaries	Global Change Master Directory (GCMD)[36]	Atmosphere, biosphere, human factors, terrestrial surface, terrestrial hydrosphere, solid earth
Authoritative domain-specific monographs	Geography Dictionary[37]	Physical geography, human geography, resource geography
	Earth System Research and Scientific Data[38]	Atmosphere, terrestrial surface, ocean, lithosphere, outer space
	Research on Earth System Scientific Data Resources[39]	Atmosphere, human-earth relations, solid earth, terrestrial surface, ocean
	Research on Integration and Sharing of Earth System Scientific Data: a Standard Perspective[40]	Atmosphere, terrestrial surface, biosphere, cryosphere, natural resources, human factors, oceans and polar regions, solid earth
Domain ontologies	Semantic Web for Earth and Environmental Terminology (SWEET)[41, 42]	Geological features, human activities, natural phenomena
Comprehensive online resources	Wikipedia[1]	Natural sciences, humanities and social sciences
	Baidu Baike[2]	Natural sciences, humanities and social sciences

English or Pinyin order in Chinese. It includes all subject terms and their related semantic relationships. The auxiliary tables reorganize the structure of the main table to meet users' diverse retrieval needs and typically include classification tables, index tables, appendices, and similar formats. To meet the research needs for the scientific data related to the earth surface system, the Earth Surface System Scientific Data Thesaurus is presented in two forms: the main table and the classification table. The classification table is organized based on the thesaurus classification system, facilitating users in analyzing the hierarchical relationships between subject terms.

3.4.2 Classification System Design

As shown in Figure 2, to effectively explore, manage, and share the scientific data related to the earth surface system, this paper, based on the structural characteristics of the earth surface system, integrating the classification principles of the Global Change Master Directory (GCMD) and features of scientific data sharing, categorizes the scientific data into three major classes: the Near-Earth Space Data, Surface Cover Data, and Surface Mechanism Data[44, 45]. The Near-Earth Space Layer encompasses fields such as atmospheric science and meteorology, focusing on various characteristics and processes within the atmospheric sphere. It aims to understand changes in meteorology, climate, and atmospheric environments. The Surface Cover Layer encompasses water bodies, soil, and activity areas of human and other creatures, covering ocean movements, interactions within ecosystem, land use, and cover types. It contributes to understanding issues related to ecosystem and resource management. The Surface Mechanism Layer includes geological and geophysical processes within the lithosphere and the crust interior, covering geological structures, volcanic activities, and rock and mineral resources, among others. It aids in understanding solid earth science and mineral resource management.

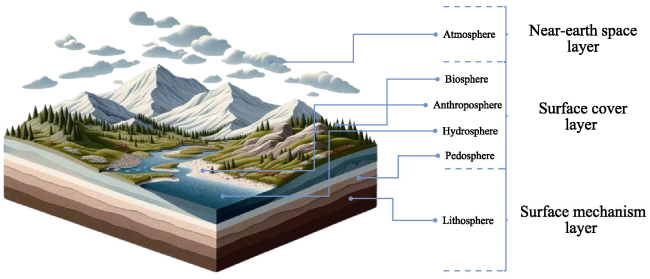

Figure 2 Sphere classification of the earth surface system sciences data

Building upon this foundation and considering the Earth's sphere structure, the three major subject classes are further subdivided into six sphere groups. Each sphere is subdivided based on its distinctive characteristics to better reflect the complexity and diversity of the

[1] https://zh.wikipedia.org.
[2] https://baike.baidu.com.

scientific data related to the earth surface system. The three-level classification system of the earth surface system scientific data thesaurus is presented in Table 4, comprising six second-level categories, and 35 third-level subcategories that directly list subject headings.

Table 4 Three-Level classification of the data thesaurus of earth surface system sciences

First level	Second level	Third level
Near-Earth space layer	Atmosphere	Atmospheric physics
		Atmospheric chemistry
		Meteorology and climate
		Weather
		Atmospheric environment
Surface cover layer	Hydrosphere	Ocean
		Polar regions
		Glaciers and permafrost
		Surface water
		Groundwater
		Hydrochemistry
	Anthroposphere	Physical geography
		Paleogeography
		Human geography
		Resource and environment
	Biosphere	Ecosystem
		Plant
		Animal
		Indigenous organism
		Bacteria
		Fungi
		Virus
	Pedosphere	Soil physics
		Soil chemistry
		Soil biology
		Soil geography
		Soil resources and environment
Surface mechanism layer	Lithosphere	Geodesy
		Rock and mineral
		Geomagnetism
		Seismology
		Geological structure
		Geological disaster
		Geotherm
		Volcano

3.4.3 Hierarchy Structure Design

The design of the hierarchical structure aims to highlight the hierarchical relationships among the subject terms. In the hierarchical structure design of the data thesaurus of earth surface system sciences, we mainly referred to the tree-like hierarchical structure of the GCMD and the classification standards in related earth science disciplines. GCMD keywords are placed under a multi-level tree structure of Category > Topic > Term > Variable > Detailed Variable to classify and associate concepts. Therefore, in designing the hierarchical structure of the earth surface system scientific data thesaurus, we followed principles of scientific rigor, systematic organization, and precision to organize the structure into levels of Basic Space > Sphere > System > Subsystem > Object > Element > Attribute, as illustrated in Figure 3. Here, Basic Space represents the geographic and spatial scope covered by earth surface system scientific data and serves as the top-level hierarchy of the thesaurus. Sphere includes the six basic sphere structures that make up the earth surface system. System represents the main domains within each sphere. Subsystem further refines the subdomains of a sphere to better represent the

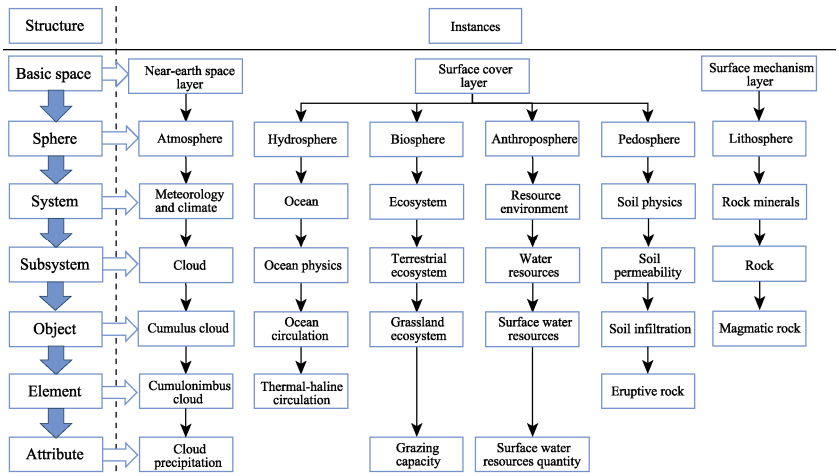

Figure 3 Hierarchy structure of the data thesaurus of earth surface system sciences

differences in various research areas. Object represents more specific entities or concepts within a subsystem. Element represents the basic components of an object, providing a more detailed description of the composition and characteristics of the object. Attribute provides a detailed description of the features and content of an element.

3.4.4 Semantic Relationship Design

The ISO 25964[46] standard specifies three fundamental semantic relationships in a thesaurus: Equivalence Relation, Hierarchy Relation, and Association Relation[8].

- **Equivalence Relation:** It indicates that two or more semantically identical or similar terms are interchangeable. This includes synonymy, abbreviation, and name evolution. Synonymy represents different terms with the same or similar meanings, such as "crustal movement" and "geological conformation", which describe deformations of the Earth's crust due to geological processes. Abbreviation refers to the relationship between the abbreviated or shortened form of a term and its complete form, for example, "CO_2" and "carbon dioxide". Name evolution signifies changes in the term's name over time, such as the replacement of geographical names in different historical periods. Analyzing the concepts of terms helps identify the existing synonym relationships.
- **Hierarchy Relation:** This denotes the hyponymy relation of terms, including Genus/Species relationships, Whole/Part relationships, and Instance relationships[47]. The generic relationship indicates a parent-child relationship between two terms. For example, "rhizobium" is a child of "bacteria". Whole-part relationships indicate that one term is a part of another term, such as "Arctic" being a part of "polar regions". Instance relationships signify that one term represents a certain entity, and the other term is an instance of that entity, for instance, "Qinghai-Tibet Plateau" is an instance of "plateau". Building hierarchical relationships between terms ensures the clarity and multi-level nature of the thesaurus.
- **Association Relation:** This indicates a relationship between terms that does not involve equivalence or hierarchy. It includes various types of relationships, as shown in Table 5.

Table 5 Main association relationships in the data thesaurus of earth surface system sciences

Relationship	Relation	Meaning
Influence Relationship	Has impact on	Indicates that one subject term impacts another
	Influenced by	Indicates that one subject term is influenced by another
Causal Relationship	Has possible cause	Indicates that one subject term may cause another
	Caused by	Indicates that one subject term is caused by another
Metric Relationship	Measures	Indicates that one subject term measures another
	Measured by	Indicates that one subject term is measured by another
Inverse Relationship	Inverse of	Indicates an inverse relationship between one subject term and another

4 Data Results and Validation

4.1 Data Composition

The data thesaurus of earth surface sciences consists of three parts:

- **"Data Thesaurus of Earth Surface Sciences Main Table" (.xlsx):** This includes the names of the thesaurus terms in both Chinese and English, synonyms, relationships, definitions, and data sources.
- **"Data Thesaurus of Earth Surface Sciencse Classification Table (Chinese Version)" (.xlsx):** This includes classification information and data sources for terms in Chinese.
- **"Data Thesaurus of Earth Surface Sciences Classification Table (English Version)" (.xlsx):** This includes classification information and data sources for terms in English.

The fields and their descriptions are shown in Table 6.

4.2 Data Products

The data thesaurus of earth surface sciences divides terms into a 7-level hierarchical tree

structure, comprising a total of 3,463 terms. Under the three major core themes, it covers 6 secondary categories and 35 tertiary categories, corresponding to the basic space, spheres, and systems in the thesaurus hierarchy. Below the tertiary categories, one term can belong to several subclasses and there are 166 subsystem terms, 589 object terms, 2,480 element terms, and 532 attribute terms. The distribution of the number of thesaurus terms in each sphere is shown in Figure 4, with the hydrosphere, anthroposphere, and atmosphere being dominant, while the number of terms in the pedosphere is relatively small. Figure 5 shows the distribution of the number of thesaurus terms referenced from each data source.

Table 6 Fields of the data thesaurus of earth surface system sciences

Entry	Description	Entry	Description
Keyword	English term name	OnProperty	Semantic relationships
ChineseName	Chinese term name	SomeValuesFrom	Object of the relationship
AltLabel	English term synonym	Comment	English term definition
ChineseAltLabel	Chinese term synonym	ChineseComment	Chinese term definition
SubClassOf	Parent class of the term	Source	Source of the term

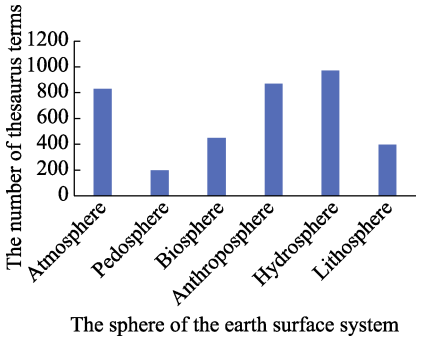

Figure 4 Distribution of thematic terms in the data thesaurus of earth surface system sciences by Sphere

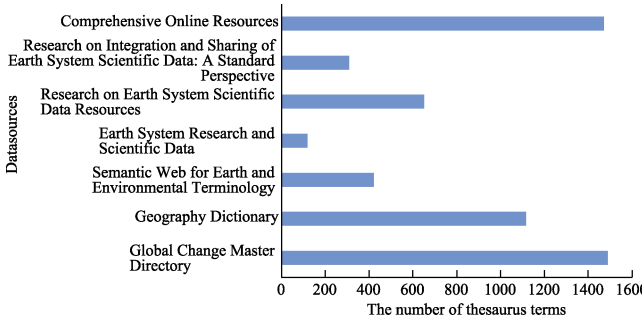

Figure 5 Distribution of thematic terms in the data thesaurus of earth surface system sciences by data source

An ontology is a formalized knowledge representation of entities, attributes, and their relationships, providing a more precise description of terms and concepts in the thesaurus. By constructing an ontology model for the data thesaurus of earth surface system sciences, the relationships between subject terms can be more intuitively expressed. In the ontology, all categories are built based on the conceptual objects in the thesaurus, organized according to the hierarchical structure of the thesaurus, and each class is augmented with its respective properties and relationships with other classes. Properties and relationships between classes are added based on their connections in the thesaurus. Properties within the ontology include shared properties and data properties. Shared properties are attributes shared among multiple categories, owned or inherited by different categories, indicating similar features or common properties. Data properties describe the basic characteristics or attributes of concepts or entities, mainly including name, definition, unique identification code, data source, etc. Relationships within the ontology are established based on the semantic relationships between subject terms in the thesaurus. The final visualization result of the data thesaurus of earth surface system sciences ontology model (the first three levels) is shown in Figure 6.

4.3 Data Validation

To validate and analyze the scale and performance of the data thesaurus of earth surface system sciences, this study refers to relevant research methods and experiences, comparing it with the Chinese Thesaurus of Geology[48], Global Change Master Directory Earth Scientific Data keywords[49], and Cultural Relics Digital Protection Thesaurus[50]. This comparison aims to provide better support for research in the field of Earth Science.

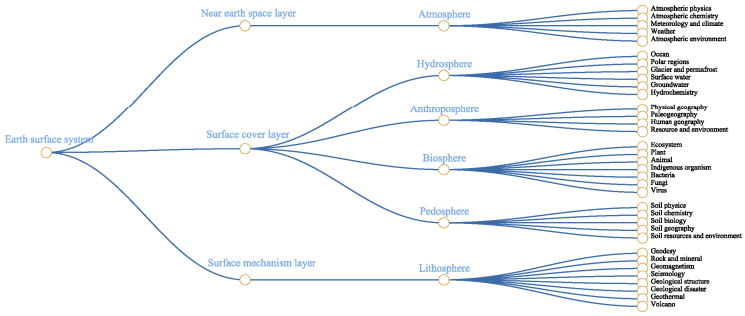

Figure 6 Visualization results of the data thesaurus of earth surface system sciences ontological model (Top three levels)

4.3.1 Thesaurus Scale Analysis

The vocabulary size refers to the volume of terms covered by a thesaurus and is a key indicator for assessing the extent of domain knowledge coverage. The vocabulary sizes for the Data Thesaurus of Earth Surface System Sciences, Chinese Thesaurus of Geology, Global Change Master Directory Earth Scientific Data keywords, and Cultural Relics Digital Protection Thesaurus are shown in Table 7. In this table, formal terms refer to vocabulary selected to represent core subjects; informal terms are vocabulary with similar or synonymous semantics to formal terms; sub-terms represent subordinate subject headings, which are more specific or detailed concepts under a broader topic; super-terms represent superordinate subject headings, which are broader or more general concepts that encompass or summarize multiple subordinate terms; and related terms are subject terms that have relevant connections, which may be associated with the subject terms in different contexts or intersect conceptually with the subject terms.

Table 7 Comparison of the scale of the Data Thesaurus of Earth Surface System Sciences, Chinese Thesaurus of Geology, Global Change Master Directory Earth Scientific Data Keywords and Cultural Relics Digital Protection Thesaurus

Thesaurus	Entries	Formal terms	Percentage of formal terms/%	Informal terms	sub-terms	super-terms	related terms
Chinese Thesaurus of Geology	10,510	8,572	81.56	1,938	\	\	\
Global Change Master Directory Earth Scientific Data keywords	1,556	1,556	100.00	0	1,541	254	0
Cultural Relics Digital Protection Thesaurus	2,605	2,468	94.74	137	407	1,648	182
Earth Surface System Scientific Data Thesaurus	3,463	3,130	90.38	333	3,460	979	354

Comparison reveals that, in terms of vocabulary size, the data thesaurus of earth surface system sciences falls into the category of medium scale, indicating its commendable coverage of domain knowledge. Additionally, its numbers of sub-terms, super-terms, and related terms also prove this characteristic.

4.3.2 Performance Analysis

In accordance with Vocabulary Control for Information Retrieval, it is stated that the performance indicators of a vocabulary include equivalence ratio, association ratio, referential density, and ancestor density[51]. The equivalence ratio is the ratio of informal terms to formal terms, and a higher equivalence ratio helps improve the retrieval effectiveness of the vocabulary, while a lower equivalence ratio shows the emphasis of the vocabulary on the accurate expression of core concepts. Association ratio and referential density are used to measure the degree of association between terms. Specifically, the association ratio is the ratio of subject terms with semantic relationships to the total number of formal terms. Referential density includes super-term referential density, related-term referential density, and total referential density. Super-term referential density is the ratio of terms with superordinate relationships to the total

number of formal subject terms, indicating the clarity in the classification and hierarchical structure of the thesaurus, related-term referential density is the ratio of terms with associative relationships to the total number of formal subject terms, indicating the extent of horizontal connection among terms, and total referential density is the sum of super-term and related-term referential densities, indicating the richness and complexity of the semantic relationships in a comprehensive manner. The performance indicators for the data thesaurus of earth surface system sciences, Chinese Thesaurus of Geology, Global Change Master Directory Earth Scientific Data keywords, and Cultural Relics Digital Protection Thesaurus are shown in Table 8.

Table 8 Performance comparison of the data thesaurus of earth surface system sciences, Chinese thesaurus of geology, global change master directory earth scientific data keywords and cultural relics digital protection thesaurus

Thesaurus	Equivalence ratio	Association ratio	Super-term referential density	Related-term referential density	Total referential density
Chinese Thesaurus of Geology	0.226	0.813	0.850	1.530	2.380
Global Change Master Directory Earth Scientific Data keywords	\	1.000	1.154	\	1.154
Cultural Relics Digital Protection Thesaurus	0.053	0.746	0.789	0.070	0.859
Earth Surface System Scientific Data Thesaurus	0.106	1.000	1.418	0.110	1.528

As shown in Table 8, the equivalence ratio of the four thesauri are generally low, which may indicate that these thesauri are not rich enough in providing synonyms or near-synonyms, thus limiting the breadth and depth of information retrieval to some extent. Particularly, the Global Change Master Directory Earth Scientific Data Keywords has an equivalence ratio of 0, which may imply that the thesaurus does not include informal subject terms, or its retrieval system does not distinguish between formal and informal subject terms, potentially affecting the flexibility and accuracy when users search data. Despite the general low equivalence ratio, the Global Change Master Directory Earth Scientific Data Keywords and the Data Thesaurus of Earth Surface System Sciences have achieved an association ratio of 1.000, demonstrating that these two thesauri have a high degree of association between terms, providing multiple related terms for each subject term, which helps to enhance the depth and accuracy of retrieval. In contrast, although the Chinese Thesaurus of Geology and the Cultural Relics Digital Protection Thesaurus have association ratio below 1, they still show a certain degree of term association, indicating that they also have certain advantages in term association. The super-term referential density and related-term referential density provide a perspective on the internal structure of the thesaurus. The data thesaurus of earth surface system sciences stands out in the super-term referential density, indicating that the thesaurus has a high degree of clarity and organization in the hierarchical structure and classification of terms, which helps users better understand the relationships between subject terms. The Chinese Thesaurus of Geology excels in the related-term referential density, showing that the thesaurus does well in ensuring horizontal connections and diversity among terms, which provides users with more retrieval perspectives and increases the coverage of retrieval. The total referential density combines the super-term referential density and the related-term referential density, reflecting the comprehensiveness of term relationships in the thesaurus. The Chinese Thesaurus of Geology and the Data Thesaurus of Earth Surface System Sciences have a higher total referential density, indicating that they are relatively superior in building term relationships, which helps to provide more comprehensive retrieval results.

In summary, comparative analysis reveals that the data thesaurus of earth surface system sciences excels in vocabulary association ratio and super-term referential density, indicating its robust hierarchical vocabulary relationships and commendable vocabulary association density, effectively reflecting the complex conceptual relationships within the earth surface system. However, the data thesaurus of earth surface system sciences shows relatively lower performance in equivalence ratio and related-term referential density as proved by a slightly inadequate

equivalence rate and related-term referential density. Therefore, further expansion of the vocabulary should be done with a focus on specific domain application requirements, such as emergency data sharing for disasters, so as to improve the retrieval effectiveness of the thesaurus.

5 Discussion and Conclusion

With the deepening understanding of the Earth's surface, surface system sciences data have become indispensable scientific resources. This paper, based on a clear definition and scope of the surface system, constructs the data thesaurus of earth surface system sciences in a combination of top-down and bottom-up approaches. The thesaurus covers various elements within the atmosphere, hydrosphere, biosphere, anthroposphere, pedosphere, and lithosphere. It integrates data from authoritative domain-specific dictionaries and comprehensive online resources, categorizing vocabulary into 3 primary categories, 6 secondary categories, and 35 tertiary categories, covering 3,463 subject terms in total. This provides robust foundational data support for data management and knowledge sharing in the field of earth sciences.

Future research will work around the data thesaurus of earth surface system sciences as a core outcome, focusing on regular updates and series application analysis mainly from the following aspects:

- **Vocabulary Expansion and Automatic Updates:** Further expanding the breadth and depth of the vocabulary, establishing an automatic update mechanism, and regularly integrating the latest scientific research results and domain knowledge from emerging and interdisciplinary fields related to the earth surface system to ensure the thesaurus's timeliness and novelty, and identifying and filling gaps in subject terms to improve the distribution of term categories.
- **Enrichment of Semantic Associations:** Enhancing the relevance between subject terms by introducing advanced deep learning and natural language understanding technologies, achieving more accurate and enriched semantic associations, and further improving the usability and effectiveness of the thesaurus.
- **Diverse Applications:** Extending the thesaurus's applications to more fields, including education, environmental protection, and disaster management, promoting the wide application of surface system sciences data, and providing greater support for social development and interdisciplinary collaboration.

Author Contributions

Zhu, Y. Q. and Duan, F. Z. designed the overall dataset development; Chen, J. collected and processed the data sources for constructing the data thesaurus of earth surface system sciences; Wang, S. designed the overall model; Chen, J., Wang, S., and Wang, B. conducted the data validation; Chen, J. wrote the data paper; Wang, S. reviewed the data paper.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Wu, C. J. On the core of geographical research: Human-environment regional systems [J]. *Economic Geography*, 1991(3): 1–6.

[2] Phillips, J. D. Earth Surface Systems [M]. Oxford: Blackwell, 1999.

[3] Qian, X. S. On the content and research methods of geographic science (Speech at the symposium on "Geographic Science" of the China Geographical Society on April 6, 1991) [J]. *Acta Geographica Sinica*, 1991(3): 257–265.

[4] Zhu, Y. Q., Sun, K., Hu, X. J., *et al*. Research and practice on the framework for the construction, sharing, and application of large-scale geoscience knowledge graphs [J]. *Journal of Geo-Information Science*, 2023,

25(6): 1215–1227.
[5] Li, X., Feng, M., Ran, Y., *et al*. Big Data in Earth system science and progress towards a digital twin [J]. *Nature Reviews Earth & Environment*, 2023, 4: 1–14.
[6] Knight, J., Harrison, S. The impacts of climate change on terrestrial Earth surface systems [J]. *Nature Climate Change*, 2013, 3(1): 24–29.
[7] Reichstein, M., Camps-Valls, G., Stevens, B., *et al*. Deep learning and process understanding for data-driven Earth system science [J]. *Nature*, 2019, 566(7743): 195–204.
[8] Martínez-González, M. M., Alvite-Diez, M. L. Thesauri and semantic web: discussion of the evolution of thesauri toward their integration with the semantic web [J]. *IEEE Access*, 2019, 7: 153151–153170.
[9] Guo, Y. Geographical Science Thesaurus [M]. Beijing: Science Press, 1995.
[10] Xue, S. S., Zhou, F., Wang, C. N., *et al*. Reconstruction of knowledge organization system based on subject headings—taking geoscience knowledge organization system as an example [J]. *Natural Resources Informatization*, 2020(3): 9–14.
[11] Shi, J. Chinese Thesaurus of Geology [M]. Beijing: Geology Press, 2010.
[12] Compilation team of the Environmental Science Thesaurus. Environmental Science Thesaurus [M]. Beijing: China Environmental Press, 1989.
[13] Editorial Committee of the Chinese Classified Thesaurus of the National Library of China. Chinese Classified Thesaurus [M]. Beijing: National Library of China Publishing House, 2017.
[14] Timmer, R. C., Mark, M., Khoo, F. S., *et al*. NASA Science mission directorate knowledge graph discovery [Z]. Companion Proceedings of the ACM Web Conference 2023. Austin, TX, USA; Association for Computing Machinery. 2023: 795–799. DOI: 10.1145/3543873.3587585.
[15] Chen, J., Wang, S., Zhu, Y. Q., *et al*. Thesaurus of scientific data for the Earth Surface System [J/DB/OL]. *Digital Journal of Global Change Data Repository*, 2024. https://doi.org/10.3974/geodb. 2024.07.10.V1. https://cstr.escience.org.cn/CSTR:20146.11.2024.07.10.V1.
[16] GCdataPR Editorial Office. GCdataPR data sharing policy [OL]. https://doi.org/10.3974/dp.policy.2014.05 (Updated 2017).
[17] Zhou, J. The second discussion on the "Earth's surface layer" [J]. *Journal of Natural Disasters*, 2004(6): 1–7.
[18] Xie, J. Z. Several issues on the view of the earth surface system [J]. *Advances in Earth Science*, 1995(5): 432–435.
[19] Ramos, V. A. Hans Keidel and Alexander du Toit's relationship and its impact on Wegener's continental drift hypothesis [J]. Geological Society, London, Special Publications, 2023, 531(1): SP531-2022-2181.
[20] Conder, J. A. An active role for the ocean in seafloor spreading [Z]. American Geophysical Union Fall Meeting 2022. Chicago, American Geophysical Union. 2022: T26B-06.
[21] Zheng, Y. F. Plate tectonics in the twenty-first century [J]. *Science China Earth Sciences*, 2023, 66(1): 1–40.
[22] Pausas, J. G., Bond, W. J. Feedbacks in ecology and evolution [J]. *Trends in Ecology & Evolution*, 2022, 37(8): 637–644.
[23] Huang, B. W. The theoretical foundation of regional sustainable development—Land system science [J]. *Acta Geographica Sinica*, 1996(5): 445–453.
[24] Wang, C. S., Cao, K., Huang, Y. J. Sedimentary record and cretaceous earth surface system changes [J]. *Earth Science Frontiers*, 2009, 16(5): 1–14.
[25] Steffen, W., Richardson, K., Rockström J., *et al*. The emergence and evolution of Earth System Science [J]. *Nature Reviews Earth & Environment*, 2020, 1(1): 54–63.
[26] Yang, S. H., Song, X. D., Wu, H. Y., *et al*. A review and discussion on the earth's critical zone research: status quo and prospect [J]. *Acta Pedologica Sinica*, 2023: 1–14.
[27] Jin, Z. J., Wang, X. M., Wang, H. J., *et al*. Organic carbon cycling and black shale deposition: an Earth System Science perspective [J]. *National Science Review*, 2023, 10: nwad243.
[28] Dou, X. C. On the ontological modal composition of the earth's surface space [J]. *Research on Development*, 1998(1): 50–51.
[29] Pu, H. X. Systems and evolution of the earth's surface [J]. *Chinese Journal of Nature*, 1983(2): 126–128.
[30] Lu, D. D. Research on the earth surface system and the development of geographic theory [Z]. Academic

Conference Commemorating the 90th Anniversary of the Establishment of the Chinese Geographical Society. Beijing, China. The geographical society of China. 1999: 8–13.

[31] Phillips, J. D. Global and local factors in earth surface systems [J]. *Ecological Modelling*, 2002, 149(3): 257–272.

[32] Zhang, M. L., Lei, X. Y. A discussion on the earth surface system [J]. *Northwestern Geology*, 2005(2): 99–101.

[33] Li, X. L., Wu, K. N., Feng, Z., *et al*. Research progress of land surface system classification: from land type to earth's critical zone type [J]. *Progress in Geography*, 2022, 41(3): 531–542.

[34] Chen, M., Qian, Z., Boers, N., *et al*. Iterative integration of deep learning in hybrid Earth surface system modelling [J]. *Nature Reviews Earth & Environment*, 2023, 4(8): 568–581.

[35] Ma, Z. J., Gao, X. L., Du, P. R. Pondering over the study on the outermost sphere system of the earth [J]. *Earth Science Frontiers*, 2006(6): 96–101.

[36] Parsons, M. A., Duerr, R., Godøy, Ø. The evolution of a geoscience standard: an instructive tale of science keyword development and adoption [J]. *Geoscience Frontiers*, 2023, 14(5): 101400.

[37] Tan, J. A. Geography Dictionary [M]. Beijing: Chemical Industry Press, 2008.

[38] Lin, S. J. Earth System Research and Scientific Data [M]. Beijing: Chemical Industry Press, 2009.

[39] Bao, L. S. Research on Earth System Scientific Data Resources [M]. Beijing: Science Press, 2010.

[40] W, J. L. Research on Integration and Sharing of Earth System Scientific Data: A Standard Perspective [M]. Beijing: China Meteorological Press, 2015.

[41] Haribabu, S., Kumar, P. S. S., Padhy, S., *et al*. A novel approach for ontology focused inter-domain personalized search based on semantic set expansion [Z]. 2019 fifteenth international conference on information processing (ICINPRO). Bengaluru, India; IEEE. 2019: 1–5

[42] Whetzel, P. L., Noy, N. F., Shah, N. H., *et al*. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications [J]. *Nucleic Acids Research*, 2011, 39(suppl_2): W541–W545.

[43] Chen, R., Zeng, J. X. Research on thesaurus integration systems and the promotion of their application [J]. *Journal of the China Society for Scientific and Technical Information*, 2022, 41(4): 401–411.

[44] Wang, J. L., Lin, H., Ran, Y. Y., *et al*. A study of earth system science data classification for data sharing [J]. *Advances in Earth Science*, 2014, 29(2): 265–267+273–274.

[45] Wang, J. L., Wang, M. M., Shi, L., *et al*. The situation of scientific data management and its enlightenment to earth sciences of China [J]. *Advances in Earth Science*, 2019, 34(3): 306–315.

[46] ISO. ISO 25964-2:2013 Information and documentation-Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies [EB/OL]. (2013-03-04) [2016-03-20]. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53658.

[47] Jing, X. Q. Multilingual marine thesaurus construction based on the category system of Wikipedia [D]. Tsingtao: Ocean University of China, 2016.

[48] Bao, X. L., Wu, W. N. Overview on the revision status of Chinese thesaurus in recent 40 years [J]. *Library and Information Service*, 2013, 57(2): 109–113.

[49] Global Change Master Directory (GCMD). GCMD Keywords, Version 17.3 [Z]. Greenbelt, MD: Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center, NASA. 2023. URL (GCMD Keyword Forum Page): https://forum.earthdata.nasa.gov/app.php/tag/GCMD+Keywords.

[50] Luo, W. Establishment and study of cultural relics digital protection thesaurus [D]. Beijing: Beijing University of Chemical Technology, 2018.

[51] Hider, P. A survey of the coverage and methodologies of schemas and vocabularies used to describe information resources [J]. *Knowledge Organization*, 2015, 42: 154–163.