Journal of Global Change Data & Discovery. 2020, 4(3): 299-313
DOI:10.3974/geodp.2020.03.13
www.geodoi.ac.cn

Progress Analysis on International Geoscience Data Management Based on Bibliometrics

Wang, S. Q.[1] Wang, J. L.[1,3,5*] Li, Y.[2,4*] Wang, J.[1] Wang, Y. J.[1] Li, H. Y.[2]

1. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;
2. National Science Library, Chinese Academy of Sciences, Beijing 100190, China;
3. China-Pakistan Joint Research Center on Earth Sciences, Islamabad 45320, Pakistan;
4. Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China;
5. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China;

Abstract: Scientific data is an important resources for science and decision making. With the advent of the big data era, countries around the world have incorporated scientific data management into national development strategies. Facing the urgent needs of geoscience data management, the development trend of geoscience data management research should be deeply understood. In this study, publications related to geoscience data management from 1900 to 2018 were obtained from Web of Science databases, including Science Citation Index (SCI), Social Science Citation Index (SSCI), and Conference Proceedings Citation Index (CPCI). Considering the number as well as the citations and cooperation of publications, a general analysis was performed on the basis of bibliometric indicators at the levels of countries/regions, institutions, and disciplines. A knowledge mapping analysis revealed that geoscience data management has developed into six research subject areas. Among them, a better theory-and-method system of geospatial data management has been formed in the last 10 years. “Research on open-access policy of scientific research data” has rapidly attracted attention. Geospatial data management research will play a driving role in geoscience big data research, data management decision modeling, and other research fields in the future.

Keywords: geoscience; data management; bibliometrics; hot topic; progress analysis

1 Introduction

Scientific data—the original and basic data that reflect the nature, characteristics, and change laws of the objective world—are obtained via scientific and technological activities. Scientific data provide an indispensable basic scientific and technological support for scientific and technological innovation, economic development, and relevant decision making.

Received: 10-08-2020; **Accepted:** 20-09-2020; **Published:** 25-09-2020

Foundations: National Natural Science Foundation of China (41842061); Ministry of Science and Technology of P. R. China (2020WT22); Chinese Academy of Sciences (XXH13505-07).

***Corresponding Author:** Wang, J. L., Institute of Geographical Sciences and resources, CAS, wangjl@igsnrr.ac.cn
Li, Y., National Science Library, liyang@ mail.las.ac.cn

Authors ID: Wang, S. Q. 0000-0002-2432-8161; Wang, J. L. 0000-0002-5641-0813; Li, Y. 0000-0002-2890- 9000 Wang, J. 0000- 0002-9669-358X; Wang, Y. J. 0000-0002-7531-2880; Li, H. Y. 0000-0002-1520-516X

Citation: Wang, S. Q., Wang, J. L., Li, Y., *et al*. Progress analysis on international geoscience data management based on bibliometrics [J]. *Journal of Global Change Data & Discovery*, 2020, 4(3): 299–313. https://doi.org/10.3974/geodp.2020.03.13.

They constitute an important national strategic resource and are recognized as the third type of resource after materials and energy[1–2]. Geoscience research is typically data-intensive, requiring a large amount of scientific data to support the process of solving scientific and application problems. Additionally, it constantly produces new derived data and products through scientific research activities[3–4]. Therefore, research on geoscience data management is of considerable strategic significance for promoting the development of geoscience, as well as the discipline of scientific data management, in China.

There are two main sources of data in the field of geoscience: 1) scientific research data obtained directly through geoscience research and practice and 2) professional data collected and managed by government departments over a long period, such as the geological survey data of the Land and Resources Management Department of China, the hydrological data of the Ministry of Water Resources of China, and the meteorological and climatic data of the Chinese Meteorological Bureau. Geoscience data management involves using computer hardware and software technology to collect, store, process, and apply geoscience data effectively. Its objectives are to fully exploit the data, effectively manage the two types of data, promote their wide sharing, and maximize their value[5]. Researchers should not only deal with difficult or untouchable scientific problems via real-time and dynamic monitoring and analysis of data but also perform data-based scientific research[6], which leads to a series of scientific data management problems in the field of geoscience.

The management and sharing of geoscience data has been gradually attracted the attention of international science communities, since the middle of the 20th century. The World Data Center was established in 1957, focusing on geoscience, space science, and astronomy data under the organization of the Council of the International Federation of Science[7]. The National Center for Atmospheric Research was established in the United States in 1960, which began the modeling, collection, and preservation of geoscience data[8]. In 1969, White[9] answered the question of why geophysical science data management should be conducted. Data management research for geoscience has become an important scientific platform to drive scientific discovery and decision support, and relevant research issues include data storage, sharing, and management policies, as well as information mining. The open sharing of scientific data has opened a channel for the wide dissemination and reuse of scientific research achievements. Under the organization of the International Council of Science, the United States and developed countries in Europe have established national scientific data center groups and data sharing service networks, such as the Distributed Active Archive Centers and the Global Change Master Directory, which are hosted by the National Aeronautics and Space Administration[10–12]. Diversified data forms pose considerable challenges to management. With the rise of the data-intensive scientific research paradigm, data publishing, data repository, and data hubs, which take entity data as the core, have attracted the attention of many scientific research institutions and scholars. Earth System Science Data has been published since 2009 and has completed data storage in cooperation with several data centers, such as Pangaea[13]. In 2019, the American Geophysical Union (AGU) launched the journal data repository program, requiring its academic journals to publish the original data associated with papers, and the data must be archived in 226 data selected repository centers identified by the AGU[14]. As the largest and most authoritative and influential intergovernmental organization in the field of earth observation, the Group on Earth Observations initiated and promoted the development of the Data Hub, which is expected to bring together open-access data, papers, algorithms, models, and computing power through a cloud-based platform. The content of geoscience data management research in China is similar to that of international research, but it started late. It was not until 1981 that Li *et al*.[15] introduced the concept of the American geoscience STATPAC data management system to

tem to China, which was significantly promoted the research on geoscience data management in China. In 1996, Li *et al*. [16] reported that the establishment of a metadata system for geoscience data was helpful for the development and utilization of geoscience data and explained the application of metadata in geoscience data management. In 2002, Sun *et al*.[17] reported that revolutionary progress has been made in earth information science owing to computer and remote-sensing technology. The problem that the daily-obtained geoscience data size in TB and PB levels cannot be effectively used perplexes the majority of geoscientists. The introduction of grid technology to the geoscience data storage and sharing system will help to solve this problem. In 2003, an updated version of China's Geochemical Data Management Information System was released[18]. Subsequently, Du *et al*.[19] established a conceptual model of the China coastal zone scientific data platform based on an analysis of many information features, aiming to satisfy the urgent needs of national spatial data infrastructure and application based on multi-source information from space, the conventional coastal zone, and offshore. On the basis of this conceptual model, the logic structure, ArcSDE storage of remote-sensing image data, metadata storage of remote-sensing data, and other models are designed. According to the needs of geoscience data sharing, Wang *et al*.[20] analyzed the framework mode and method of general geoscience metadata. The metadata framework constructed via this method includes three levels: core metadata, pattern metadata, and application domain-specific metadata. Aiming to solve the problems of data storage organization, data throughput processing, and data integration application faced by the current domestic geographic spatiotemporal big data production management and application (from the perspective of the whole process management of geographic entity generation and death and geographic data production service), Xiao *et al*.[21] studied the related methods for the whole-life cycle management and application of geographic spatiotemporal big data. Following the general trend of global scientific data publishing and storage development, *China Scientific Data*, *Journal of Global Change Data & Discovery*, *Big Earth Data* and other data journals have taken the lead in establishing relatively complete data paper review, storage, and peer review processes to quickly promote domestic data publishing. Among them, the "Global Change Scientific Research Data Publishing and Repository" (in both Chinese and English) has taken practical steps towards protecting data intellectual property rights and promoting data sharing. Achievements in data property rights certification, data quality standards, peer review by experts, long-term data preservation, open data sharing, and international qualification networking significantly affect the value of data[22]. Since 2018, the Chinese Astronomical Data Center; World Data Center for Renewable Resources and Environment; Geo-scientific Data & Discovery Publishing Center; WDC for Geophysics, Beijing; and National Space Science Data Center have successively become internationally recognized centers for data repositories and publishing or data hubs[23]. The Standing Committee of the Political Bureau of the CPC Central Committee held a meeting on March 4, 2020, pointing out that it is necessary to accelerate the construction of new infrastructure such as data centers. The construction of big data centers—the hubs in the era of the digital data economy—has become an inevitable trend.

The research on geoscience data management at home and abroad has experienced several decades of development and resulted in achievements, but most of them involve advancements in technical methods in specific fields, and there is a lack of comprehensive analysis of geoscience data management research from a bibliometric perspective. To satisfy the development needs of big data technology and data management standardization in the field of geoscience in the current era of big data we analyzed the development trend and research progress of international geoscience data management. Our objective was to provide a decision-making guidance for promoting and developing China's geoscience data management.

2 Data Sources and Research Methods

2.1 Data Sources

Geoscience, which takes the whole earth as the object of study, is a fundamental natural science for human beings to rationally develop natural resources; fully exploit natural conditions; avoid and mitigate natural disasters; adapt to natural laws; coordinate populations, resources, and environments; and achieve sustainable development[24]. Geoscience data management includes not only natural science but also humanities, society, and management science. In this study, publications related to geoscience data management from 1900 to 2018 were retrieved from Web of Science (WoS) core collection, including SCI, SSCI, and CPCI database that had wide coverage and influence.

2.2 Retrieval Principles and Strategies

The subject words of "geoscience data management" can be divided into "geoscience" and "data management". Therefore, a combined retrieval strategy for geoscience subjects and data management topics is suitable for this study. The retrieval words of geoscience include 20 branches: ecological environment science, geochemistry and geophysics, geology, remote sensing, astronomy and astrophysics, meteorology and atmospheric science, public environment and occupational health, water resources, agricultural physical geography, oceanography, mining and mineral processing, forestry, fishery, geography, mineralogy, urban research, regional research, biodiversity protection, imaging science & photography technology. Data management subject retrieval words are defined according to the elements in Table 1 and data management-related laws, regulations, and policies.

2.3 Data Processing and Evaluation Indices

A total of 3,202 publications were obtained via the foregoing retrieval strategy on June 20, 2019. After expert identification and the exclusion of irrelevant works, 2,391 publications were retained. To obtain more accurate quantitative statistics, we cleaned the information of institutions and keywords. Then, we analyzed the comprehensive development trend of global data management research and the progress of the research field quantitatively using software tools such as DDA, Microsoft Excel, CiteSpace[25], and VOSviewer[26].

The evaluation indices used in this study included the number of papers published, total citation frequency, and average citation frequency. The number of published papers refers to the number of papers published by scientific researchers, scientific research institutions, or countries within a certain time period. The total citation frequency refers to the number of citations of all the literature in a certain field within a certain period of time. The average citation frequency refers to the ratio of the number of citations to the total number of records retrieved in a certain field within a certain period of time.

3 Results and Analysis

3.1 Global Comprehensive Situation Analysis Based on Articles and Citation Records

3.1.1 Analysis of Research Interest and Overall Influence

The annual number of publications in a certain research field can reflect the research enthusiasm in the field to a certain extent. Figure 1 shows the trend of publications in the global geoscience data management research field. The earliest literature can be traced back to 1953, titled "Sources of legal information in Poland", published in *Law Library Journal*.

From 1953 to 1974, the total number of publications worldwide was less than 10. From 1975 to 1997, the number of publications fluctuated, but there was an increasing overall trend. From 1997 to 2001, the number of publications decreased slightly. Since 2002, it has been increasing rapidly. In 2014, 99 articles were published, and the research interest reached the highest point in history. In recent years, the number of publications has remained high.

The influence of research results can be judged by citations to a certain extent. Geoscience data management research spans 66 years from 1953 (when the first paper is published) to 2018. In this study, the average number of citations is calculated by five-year span segments (the last segment is six years). As shown in Figure 2, the average citation frequency fluctuated between 0 and 3.5 before 2002. Since 2003, the average citation frequency has increased rapidly, reaching a maximum of 10.85 citations per article in the time zone from 2008 to 2012, when the influence reached its peak.

However, the influence of geoscience data management was lower than the overall research influence of geoscience. Table 2 presents a comparison between the average of citations in the geoscience data management field in a decade and the citation base-lines of ESI. From table 2 and Figure 3, it is clear that in the past 10 years, the average number of citations in the field of geoscience data management only exceeded the 50% baseline of ESI in 2010 and 2014.

Table 1 List of data management topics

1. **Data governance**	7. **Data repository and business intelligence management**
Data assets	Business intelligence
Data governance	Data mart
Data specialist	Data mining
2. **Data architecture, analysis, and design**	Data movement (extract, transform, load)
Data analysis	Data repository
Data architecture	8. **Document, record, and content management**
Data model	File management system
3. **Database management**	Records management
Database management	9. **Metadata management**
Database management system	Metadata
Data maintenance	Metadata discovery
4. **Data security management**	Metadata publishing
Data access	Metadata registration
Data erasure	10. **Contact data management**
Data confidentiality	Business continuity planning
Data security	Market operation
5. **Data quality management**	User data integration
Data cleaning	Identity management
Data integrity	Identity theft
Data richness	Data theft
Data quality	ERP software
Data quality assurance	CRM software
6. **Reference data and master data management**	Location
Data integration	Postcode
Master data management	E-mail
Reference data	Telephone number

Note: This table comes from https://encyclopedia.thefreedictionary.com/data+management.

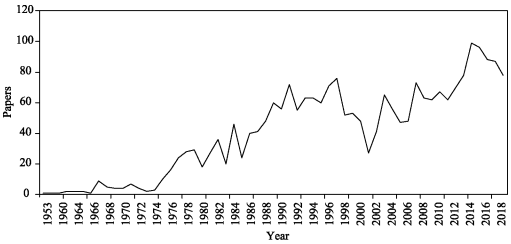

Figure 1 Trend of publications in geoscience data management research

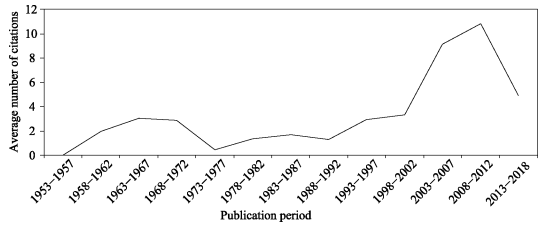

Figure 2 Average number of citations in each time zone

The distribution of publishing journals can also reflect an aspect of literature influence. According to the statistics of the literature types, 2,391 papers in the field of geoscience data management were published in 342 journals. Among the journals, 249 can be found to have impact factors in the latest version of the 2018 JCR, including six Chinese journals. The interval distribution of the impact factors of these journals is presented in Table 3. There are five journals

with an impact factor (IF) greater than 7: *Environmental Health Perspectives* (8.309), *Frontiers in Ecology and The Environment* (8.302), *Bulletin of the American Meteorological Society* (7.804), *Conservation Letters* (7.279), *Water Research* (7.051). Impact factors (IF) of most journals are greater than or equal to 1 but less than 4. There were 378 articles published in the journals which IFs were greater than or equal to 1 but less than 2. ranking the first place. While, 327 articles published in the journals which IFs were greater than or equal to 2 but less than 4, ranking the second.

Table 2 ESI field baselines and average citations of geoscience data management research in a recent 10-year period

Subject	Baseline	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Geoscience	0.01%	2,028	1,401	1,211	696	759	541	567	256	110	43
	0.10%	609	530	448	372	328	239	174	108	55	22
	1.00%	195	163	153	131	111	85	63	44	24	9
	10.00%	59	52	48	41	35	28	22	15	9	3
	20.00%	37	33	30	27	23	19	15	10	6	2
	50.00%	15	13	12	11	9	8	6	4	3	1
average citations		12.82	18.07	6.74	9.24	8.24	9.04	5.74	3.52	2.13	0.88

3.1.2 Analysis of Scientific Research Strengths of Countries/Regions and Institutions

(1) Country/region ranking of publications

Table 3 Numbers of journals and publications in each impact-factor interval

Impact-factor interval	Number of journals	Number of papers
IF ≥ 7	5	11
4 ≥ IF < 7	29	237
2 ≥ IF < 4	82	327
1 ≥ IF < 2	74	378
IF < 1	59	258

The authors who published their articles in the field of geoscience data management are from 88 countries/regions. As shown in Figure 3, there is a significant difference in the number of articles among the TOP20 countries (referred to ones with the TOP20 publications). Among them, the United States takes the lead, with 655 published papers, accounting for 27.39% of the total published papers. The United Kingdom and China rank the second and the third, with 177 and 123 papers, respectively. The combined proportions of the two countries (United Kingdom, 7.40%; China, 5.14%) are less than half of that of the United States. They are followed by Germany (89), Canada (75), and so on.

As shown in Figure 4, the United States has conducted research in this field since 1966 and was the first among the TOP20 countries to do so. It was followed by Britain, Italy, and France, in 1972; Belgium in 1975; and Canada, Australia, and India in 1977. Compared with other countries, China started late, but its research has developed rapidly. The first international journal paper was published in 1998, and the total of published articles has increased rapidly after 2004, indicating that China has had a high level of research interest in this field in recent years. The United States, which holds a leading research position, has been in a low-speed and unsteady development state since the first paper was published in 1966. Until the 1990s, the amounts of articles and research interest remained large.

(2) Analysis of scientific research strengths of key institutions

Based on the statistics of the first authors from all the papers in the research field of geoscience data management, there are 25 institutions with 6 papers or more (hereinafter referred to as the TOP25 key institutions), as shown in Table 4. The key institutions are concentrated in the United States (17), the United Kingdom (5), China (2), and Canada (1). These countries account for 68%, 20%, 8%, and 4%, respectively, of the total number of institutions. The two Chinese institutions in the TOP25 are Wuhan University (11 papers) and Peking University (7 papers).

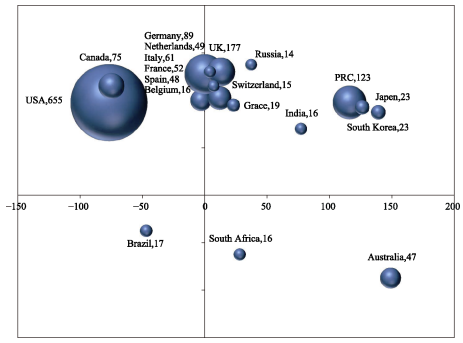

Figure 3 Distribution of papers in TOP20 countries

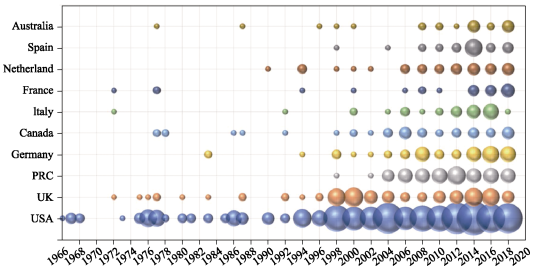

Figure 4 Paper output-time matrix of the TOP10 countries

Loughborough University, which ranked first with regard to the number of papers, published four papers for the first time in 1998. Although the papers were published relatively later, the number of papers was relatively large. 1998–2009 was the most active period of research at Loughborough University. By 2010, the number of papers published per year was reduced, and no papers were published in the past three years. The active period for publishing papers at the University of Illinois at Chicago was 1998–2007. After 2008, the number of papers published per year decreased significantly. The active research period at the National Oceanic and Atmospheric Administration of the United States was 2002–2010. The active research period at Wuhan University in China was 2003 to the present. The University of Kentucky started publication later, but it has been in an active period of research since 2014.

Table 4 Publications and citations of the TOP25 first author's institutions

Institution of the first author	Number of papers	Number of citations	Citation percentage (%)	Average number of citations per paper
Loughborough University	29	117	1.21	4.03
University of Illinois at Chicago	16	56	0.67	3.50
US National Oceanic and Atmospheric Administration	15	19	0.63	1.27
Wuhan University	11	11	0.46	1.00
University of Kentucky	10	70	0.42	7.00
City University London	9	53	0.38	5.89
Florida State University	9	83	0.38	9.22
Indiana University	9	110	0.38	12.22
University of Illinois at Urbana-Champaign	9	56	0.38	6.22
The University of Sheffield	9	118	0.38	13.11
University of Southern California	8	58	0.33	7.25
Victoria University of Wellington	8	7	0.33	0.88
Michigan State University	7	103	0.29	14.71
Peking University	7	7	0.29	1.00
University of Michigan	7	56	0.29	8.00
University of Toronto	7	17	0.29	2.43
Columbia University	6	304	0.25	50.67
US National Optical Astronomy Observatory	6	12	0.25%	2.00
Purdue University	6	24	0.25%	4.00
State University of New York at Albany	6	186	0.25%	31.00
University College London	6	2	0.25%	0.33
University of Maryland	6	67	0.25%	11.17
University of Pittsburgh	6	68	0.25%	11.33
University of Tennessee	6	58	0.25%	9.67
University of Wisconsin	6	32	0.25%	5.33

Although Loughborough University ranks first with regard to the number of papers published, the average number of citations per paper is only 4.03, which is lower than the average number of citations in the field (6.55). Although Columbia University published fewer

papers (6 papers), the average number of citations per paper is 50.67, ranking first. The number of papers published by the State University of New York at Albany is 6, and the average number of citations per paper is 31. Among the TOP25 institutions, 11 of them have more than the average of citations, and most of these institutions are located in the United States. The average of citations per paper published by Wuhan University and Peking University in China is 1.

3.1.3 Subject Branch and Relationship Analysis

According to the subject classification standard of "Web of Science categories", we statistically analyzed the subjects of the literature in the field of geoscience data management and listed the top 20 disciplines (referred to as the TOP20 subjects) in order of the number of publications. See Table 5 for details. The statistical results indicated that the research papers on global geoscience data management involved 118 subjects. Among them, information science and library science had the most papers (1,411), followed by information systems and computer science (579 papers). The total number of papers for the two subjects was 1990, which is an absolute proportion of the whole number of papers related to geoscience data management.

The TOP15 Web of Science (WoS) subjects in the field of geoscience data management and the corresponding first-level classifications of "geoscience data sharing platform classification and cataloging system"[27] are presented in Table 6. The total number of publications in the TOP15 subjects is 1,174. Among the TOP15 subjects, the following WoS sub-disciplines have100 papers or more: environmental science, remote sensing, geology, multidisciplinary integrated geoscience, water resources, and environmental research. According to this classification, the TOP15 WoS sub-disciplines can be classified into six first-level classifications. The terrestrial surface includes 8 sub-disciplines, with 703 papers, accounting for 59.9% of the total. Remote-sensing data account for 16.0% with 188 papers. Natural resources include three sub-disciplines, with 154 papers, accounting for 13.1%. Oceanography accounts for 5.3%, with 62 papers. Atmospheric accounts for 4.0%, with 47 papers. Solid earth and ancient environment accounts for 1.7%, with 20 papers.

The subjects of WoS classification system are in multipoint system, and an article may belong to multiple subjects. To reveal this relationship, we drew the subject relationship map of the papers related to the TOP15 disciplines, and the results are presented in Figure 5. The sizes of the circles in the figure indicate the numbers of papers published (by subject), and the line thickness between the circles indicates the relative number of papers belonging to two or more subjects at the same time. As shown, the fields of environmental sciences and remote sensing not only had a large number of articles, but also closely related to other subjects, particularly remote sensing and environmental research and physical geography.

Table 5 Number of papers published in the TOP20 subjects

Subject	Number of papers	Subject	Number of papers
Information science and library science	1,411	Communication	78
Information system computer science	579	Electrical and electronic engineering	76
Environmental science	223	Physical geography	73
Remote-sensing science	188	Geography	65
Interdisciplinary applied computer science	164	Environmental engineering	62
Multidisciplinary geoscience	150	Oceanography	62
Imaging science and photographic technology	117	Law	58
Water resources	112	Astronomy and astrophysics	56
Environmental research	101	Artificial intelligence computer science	55
Electric communications	98	Management	54

Table 6 TOP15 subjects among geosciences in WoS

Number	WoS category	Number of papers	First-level classification
1	Environmental science	223	Terrestrial surface
2	Remote sensing	188	Remote-sensing data
3	Multidisciplinary geoscience	150	Terrestrial surface
4	Water resources	112	Natural resources
5	Environmental studies	101	Terrestrial surface
6	Physical geography	73	Terrestrial surface
7	Geography	65	Terrestrial surface
8	Oceanography	62	Ocean
9	Meteorology and atmospheric science	47	Atmosphere
10	Ecology	43	Terrestrial surface
11	Urban studies	26	Terrestrial surface
12	Diversity protection	22	Terrestrial surface
13	Forestry	22	Natural resources
14	Multidisciplinary agriculture	20	Natural resources
15	Geochemistry and geophysics	20	Solid earth and paleoenvironment

3.2 Research Field Division and Progress Review Based on Graph Analysis

3.2.1 Research Field Division

All the data in the field of geoscience data management were regarded as a dataset based on the results of literature retrieval, and the "Author Keyword" field of the article was cleaned manually using the DDA software. 101 keywords with a frequency greater than4 were selected from the 2,609 keywords as the analysis objects, and the data of high-frequency keywords were clustered using the VOSviewer software to generate a keyword co-occurrence relationship network map (Figure 6).

The keywords can be divided into six research areas according to the analysis of this map and expert interpretation: (1) research on the management, integration, and sharing mechanism of big data for environmental monitoring; (2) research on the main content and evolution of information policy; (3) research on management models and systems that support decision making; (4) research on the open-access policies for scientific research data; (5) research on the application of GIS as an Earth science data management tool; and (6) research on information security policy formulation.

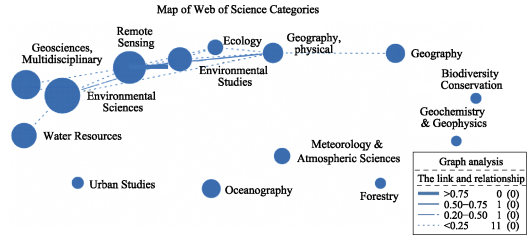

Figure 5 Relationship map of the TOP15 key subjects

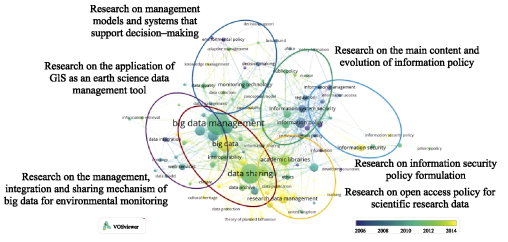

Figure 6 Time-evolution diagram of the geoscience data management research field

3.2.2 Relevance Analysis of Research Fields

Table 7 presents the clustering parameters of the foregoing research fields according to the number of core topics. The average citation frequency of the core topics in the analysis results represents the average citation frequency of the papers containing core topics. The average correlation strength represents the closeness of the connection among core topics contained in the cluster. Higher average correlation strength for a cluster corresponds to a higher co-occurrence strength between the core topics and more concentrated research. Conversely, a lower collinear strength corresponds to more scattered research. The total correlation

strength of the core subject represents the importance of the subject in the co-occurrence network. Higher correlation strength indicates that the subject is more important for the construction of the network.

"Research on the management, integration, and sharing mechanism of big data for environmental monitoring" had the highest average correlation intensity; i.e., it is the topic with the most concentrated research content and has relatively little crossover with other research contents. It is followed by "Research on management models and systems that support decision making", which mainly focuses on the policy research of decision support and the construction of related systems. "Research on information security policy formulation" has the lowest average correlation strength; i.e., it is the most divergent topic with regard to research content. It has many crossovers with other research contents, including multiple crossovers with geoscience data acquisition, management, and integration, as well as computer science and cryptography.

Table 7 Clustering parameters of research fields

No.	Research topics	Number of core topics	Mean time of occurrence	Average citation frequency	Average correlation strength
1	Research on the management, integration, and sharing mechanism of big data for environmental monitoring	25	2011	7.52	23.32
2	Research on the main content and evolution of information policy	20	2008	12.6	15.35
3	Research on the management model and system of supporting decision	16	2010	15.47	10.06
4	Research on the open-access policy for scientific research data	14	2012	16.56	20.71
5	Research on the application of GIS as an earth science data management tool	13	2010	8.32	13.54
6	Research on information security policy formulation	12	2008	17.82	8.58

3.2.3 Analysis of Research Progress in Various Fields

As shown in Table 7, among the six research fields of geoscience data management, the earliest ones are "Research on management models and systems that support decision making" and "Research on information security policy formulation", both of which appeared in 2008. The latest field to appear was "Research on open-access policy for scientific research data", which was seen in 2012 and is an emerging research topic.

(1) Research on The management, integration, and sharing mechanism of big data for environmental monitoring

The International Geophysical Year (1957–1958) and the International Biological Program (1964–1974) are the embryonic forms of big data research on the ecological environment and are referred to as "big scientific research". The objective is to obtain a large number of reliable observation data to study the Earth's spheres and problems with the ecological environment. These studies led to the development of an ecosystem research network based on long-term positioning observations, for obtaining comprehensive observation data about the ecological environment. Academic journals such as *Nature*[28] published special issues discussing big data in 2008, indicating that big data research had received worldwide attention and recognition. As indicated by Table 7, there are 25 core topics in this field, among which the main core keywords include "big data management", "data sharing", "monitoring technology", and "data archive". Among the 100 papers with the highest citation frequencies in all fields of geoscience data management, there are 15 papers related to big data of the ecological environment.

The main research content was the research on management systems, integration methods, and sharing mechanisms of big data obtained by various monitoring systems in the field of environmental science. The fields covered by big data include different large-scale sky sur-

vey observation data[29], river basin and air pollution monitoring data[30], agricultural resources and production-related data[31], meteorological data[32], and marine data[33]. The research on data management systems is closely combined with the development of management systems. The development of the data integration method also involves the construction of network infrastructure. Data sharing mechanisms need to be concerned with copyright issues, privacy issues, and collaboration systems. The mean time of occurrence of this field is 2011, and it has been a popular research topic, receiving continuous attention in the Chinese and foreign academic circles.

(2) Research on main content and evolution of information policy

There are 20 hot topics in the information policy field, and the keywords are "information policy", "public policy", "policy information", and "model". The main research focus is to analyze and summarize the information policy formulation and practices of various countries. Most of the publications are review papers[34], public opinions[35], etc. China and Europe are focused mostly, and France and the United Kingdom are also major research area. Government-level information policies include intellectual property policies, communication policies, public information dissemination policies, and information acquisition policies. The concept of information policy can be traced back to the 1990s[36] and is the earliest topic of data management research. Lately, the research on information policy and its implementation effect[37–38], technical standards of system interoperability, open-access system, information processing, were concerned, and the formulation of a complete information-technology standard policy was proposed[39].

(3) Research on management models and systems that support decision making

There are 16 hot topics in geoscience data management models field, and the keywords include "decision making", "decision support", "management model", and "environmental policy". The main research focus is on designing a management model for decision support, and the core issues include data acquisition, data quality, sustainability, and knowledge management[40]. Additionally, the development of corresponding management platforms and other key technical research should be pursued. This research on decision support is mainly applied to consulting issues related to environmental monitoring and protection[29], and there are also cases of electric vehicles and medical issues[41].

(4) Research on open-access policy for scientific research data

There are 14 hot topics in open access data field. The keywords include "research data", "research data management", "research data sharing", "data access", and "data sharing behavior". The average citation frequency is 16.56, which is only lower than that of "Research on information security policy formulation". The main research content in this field is related to research on open access of scientific research data generated or purchased by universities, research institutions, and governments. The specific issues include management support, the sharing mechanism, the implementation approach, relevant legislation, and service licensing. The scientific research information involved includes both data directly generated by subject research and resources purchased through various methods.

Many scholars have conducted multidimensional investigations on the behavior and intention of scientific data sharing[42–46]. As early as 1985, Fienberg *et al*. proposed the sharing of scientific research data. Some authors considered that a compulsory sharing policy for scientific research data was necessary in 1995[47]. In 2003, publication sharing attracted attention, and the United States issued the "Scientific Literature Public Access of Science Act" (Public Access of Science Act, commonly known as the Sabo Act) in the same year. The proposal calls for amending the current copyright protection laws in the United States to exempt copyright protection for publicly funded research results. The publication and sharing of original scientific data received attention around 2011[48]. Supporters of open access

proposed an enhanced version of the "Fair Access to Science and Technology Research (FASTR) Act" until February 2013. This memorandum was a milestone in the development of open access to scientific data[49]. Additionally, governments such as the United Kingdom and Denmark place considerable importance on the formulation of open-access policies. In the United Kingdom in particular, a national open-access policy was initially formed through RCUK (Research Councils UK), and the opinions of all parties have been incorporated into the implementation, leading to gradual improvements. The average year of formation in this field is 2012, and the topic continues to attract attention in Chinese and foreign academic circles.

(5) Research on application of GIS as earth science data management tool

There are 13 hot topics in GIS applications, and the keywords include "geospatial data", "geographic data management", "graphic database", "topology", and "data integration", mainly based on the geographic information system as the core tools or research objects unfolded. The research issues include the integration and management of geospatial data, the mechanism for sharing geospatial data among multiple subjects, the development of related management systems, and the design of geographic services. The main technical issues to be discussed include system interoperability, data standardization and integration, network technology development, infrastructure construction for spatial data acquisition, and semantic retrieval. The application fields include mineral resource exploration, water resource management, waterway data management, biodiversity protection, and architectural heritage surveying. Additionally, practical research in this field has been conducted in Europe and Central Asia. The foregoing research content can be collectively referred to as "geospatial data management". Geospatial data management links data acquisition, data modeling, data visualization, and data analysis. It makes the continuous availability and replicability of geospatial data possible. Four major achievements have been made in geospatial data management research in the past 10 years. First, GIS/BIM integration has been promoted at the data, process, and application levels[50–54], which has improved the level of geospatial data management. Second, topology is taken as a key concept of geospatial data management to construct the entity relationship model[55–56]. Third, significant progress has been made in the construction of three-/four-dimensional geospatial databases. For example, the parallelization of queries using n-dimensional space-filling curves has been verified[57]. Fourth, the GeoAI approach is used for geospatial data management to provide a more efficient solution for intensive use of data sources, including geoscience[58]. The average year of formation in this field is 2010, and the field has broad prospects for development.

(6) Research on information security policy formulation

There are 12 hot topics in the data and information security field. The keywords include "information security", "security policy", "system policy", "research framework", "information science", and "public management". The average citation frequency is 17.82, ranking first among the six research fields. The research in this field involves the formulation of network security strategies for geographic scientific information in complex network environment[59–60]. The field is characterized by multidisciplinary research, involving subjects such as cryptography, public management, and public policy formulation. The data and information holders involved include universities, governments, and commercial institutions. Additionally, research in this field has applications in water management and agricultural production in Africa.

4 Conclusion and Prospects

Since the concept of data management was proposed, the total of articles published related to earth science data management has an increasing trend. This event indicated that the hot

topic on data management has increased slowly in the early period from 1957–1997, and began to be increased rapidly since 2002. There were three stages in the development of related fields: the embryonic stage (1953–1974), the growth stage (1975–1997), and the formation and development stage (1998–present). It is obviously that the United States took the lead in establishing the first data management law (the "Freedom of Information Act" in 1966); it is true that research on geoscience data management had already begun to pay attention from more than 10 years earlier in Poland. It can be inferred that international geoscience data management research continues to lag behind the development of geoscience research and is far below the overall impact of geoscience research. China should further strengthen its research on geoscience data management and strive to become a world leader as soon as possible.

With regard to numerous comprehensive indicators, such as the number of papers published, key institutions, and citation frequency, the United States is in the leading position in the field of geoscience data management research, followed by the United Kingdom. Research in this field is on the rise in China; the number of publications is exceeded only by that of the United Kingdom, and there is room for development and improvement. Among the top 25 key institutions, Loughborough University, University of Illinois at Chicago, National Oceanic and Atmospheric Administration, Wuhan University, and University of Kentucky published the most papers. The average number of citations per paper of Columbia University, the State University of New York at Albany, and Michigan State University was larger. The results indicate that institutions with a higher level of research interest are not necessarily those with greater influence. In the future, we should focus on the overall research and development plans of institutions with greater influence.

Geoscience data management research involves non-geoscience disciplines, such as information science & library science, and information system computer science, and the fields with a large data volume or information-technology advantages are developing rapidly. This confirms that information science & library science, information system computer science, etc. have become the foundation of the main theories and methods of geoscience data management research. Among the sub-disciplines of geoscience, data management research in the fields of environmental science, remote sensing, geology, multidisciplinary geoscience, water resources, and environmental study confers clear practical advantages and has developed rapidly. Although the research on land-surface monitoring based on remote-sensing data has prominent advantages, the research on geoscience data management will have substantial potential for development in various disciplines under the current data-intensive scientific research paradigm.

Based on the graph analysis, the geoscience data management can be divided into six research fields: (1) the management, integration, and sharing mechanism of big data for environmental monitoring; (2) the main content and evolution of information policy; (3) management models and systems that support decision making; (4) open-access policies for scientific research data; (5) the application of GIS as an Earth science data management tool; and (6) information security policy formulation. Among them, geospatial data management has developed into an interdisciplinary scientific field, with scientific methods, processes, algorithms, and systems that can extract knowledge, patterns, and conclusions from unstructured and structured data. Research on geospatial data management will play a significant role in the fields of geoscience big data research and data management decision models. The key future research directions of geospatial data management are as follows: (1) semantics, geometry, and topology may become key concepts supporting geospatial data modeling and management. (2) direct application of *in situ* geographic computing of data flow libraries and objects to sensors will completely change geographic information science and geo-

spatial data management. (3) research and applications of geospatial data management based on geoAI will be further developed.

References

[1] Guo, H. Big Earth Data: a new frontier in earth and information sciences [J]. *Big Earth Data*, 2017, 1(1/2): 4–20.

[2] Guo, Hu. D. Scientific big data—a footstone of national strategy for big data [J]. *Bulletin of Chinese Academy of Sciences*, 2018, 33(8): 768–773.

[3] Boulton, G. The challenges of a big data earth [J]. *Big Earth Data*, 2018, 2: 1–7.

[4] Wang, J. L., Yang, Y. P., Zhu, Y. Q., *et al*. Data archiving progress and data types analysis of national basic research program of China (973 Program) in resource and environment field [J]. *Advances in Earth Science*, 2009, 24(8): 947–953.

[5] Si, Li., Xing, W. M. Scientific data management and sharing policies in foreign countries: investigation and inspiration to us [J]. *Information and Documentation Services*, 2013(1): 61– 66.

[6] Ding, P. Data management policy for scientific research in overseas universities [J]. *Library Tribune*, 2014(5): 103–110.

[7] Li, J. H., Yu, L. Q. Review on progress and trend of international scientific databases [J]. *E-science Technology & Application*, 2009(1): 6–13.

[8] Hou, X. G., Luo, Y. F. The new characteristics of research priorities of national Center for atmospheric research [J]. *Advances in Earth Science*, 2006(7): 751–756.

[9] White, R. M. Geophysical data management—why [J]. *Bulletin of the Amrican Meteorolgical Society*, 1969, 50(3): 143.

[10] NASA Distributed Active Archive Centres [EB/OL]. http://gcmd.gsfc.nasa.gov/. 2005.

[11] NASA's Global Change Master Directory [EB/OL]. http://gcmd.gsfc.nasa.gov/. 2007.

[12] The Canadian Earth Observation Network [EB/OL. http://www.geoconnections.org]. 2005.

[13] Wang, J. L., Shi, L., Wang, Y. J., *et al*. Analysis of the modes of aggregation of scientific data and proposals for its development in China [J]. *Advances in Earth Science*, 2020, 35(8): 839–847.

[14] Repository Finder [EB/OL]. https://repositoryfinder.datacite.org/. 2020-12-01.

[15] Vantrump, G., Miesch, A. T. The U. S. geological survey rass-statpac system for management and statistical reduction of geochemical data [J]. *Computers & Geosciences*, 1977, 3(3): 475–488.

[16] Li, J., Chen, C. C. A study on the metadata of earth Science data (Geo-metadata) [J]. *Geographical Research*, 1997(1): 31–38.

[17] Sun, J. L., Li, S. Geo-data sharing and data-grid [J]. *Earth Science*, 2002(5): 539–543.

[18] Data Processing Division of Development Research Center, China Geological Survey. The updated version of Regional Geochemical Data Management Information System (GeoMDIS 2003) comes out—a practical tool for prospecting geochemists [J]. *Geologcal Bulletin of China*, 2003(7): 547–548.

[19] Du, Y. Y., Yang, X. M., Wang, J. G. Construction and implementation of multi-sources spatial data management system of China's coastal zone and offshore [J]. *Acta Oceanologica Sinica*, 2003(5): 38–48, 57.

[20] Wang, J. L., You, S. C., Xie, C. J. Analysis and design of metadata standard structure for geosciences data Sharing [J]. *Geography and Geo-Information Science*, 2005(1): 16–18, 37.

[21] Xiao, J. H., Wang, H. Z., Peng, Q. S., *et al*. Research on the construction of cloud platform for the spatio-temporal big data management and application [J]. *Bulletin of Surveying and Mapping*, 2016(4): 38–42.

[22] Liu, C., Guo. H. D., Uhlir, P. F., *et al*. GCdataPR: Infrastructure for data publishing repository & sharing in/for/with developing countries [J]. *Journal of Global Change Data & Discovery*, 2017, 1(1): 3–11.

[23] Wang, J. L., Wang, Y., Bu, K., *et al*. Practice in the core trust seal certification of World Data Center—a case study of WDC—renewable resources and environment [J]. *Journal of Agricultural Big data*, 2019, 1(3): 71–81.

[24] Geoscience Dictionary Editorial board. Geoscience Dictionary (Basic Subject Volume) [M]. Beijing: Geological Publishing House, 2006.

[25] Chen, C. M. CiteSpaceIII [DB/OL]. http: // cluster ischool.Drexeledu/ cchen/citespace/ download/. 2016.

[26] Centre for Science and Technology Studies, Leiden University. VOSviewer Version 1.6.4 [DB/OL]. http: //www.vosviewer.com/. 2016.

[27] 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): 265–274.

[28] Frankel, F., Reid, R. Big data: distilling meaning from data [J]. *Nature*, 2008, 455(7209): 30–30.

[29] Ivezic, Z., Lupton, R. H., Schlegel, D. SDSS data management and photometric quality assessment [J]. *Astronmische Nachrichchten*, 2004, 325(6/8): 583–589.

[30] Chen, X. Y., Shao, S., Tian, Z. H. Impacts of air pollution and its spatial spillover effect on public health based on China's big data sample [J]. *Journal of Cleaner Production*, 2017, 142: 915–925.

[31] Steiner, J. L., Sadler, E. J., Chen, J. S. Sustaining the earth's watersheds-agricultural research data system: overview of development and challenges [J]. *Journal of Soil and Water Conservation*, 2008, 63(6):

569–576.
[32] Muller, C. L., Chapman, L., Grimmond, C. S. B. Toward a standardized metadata orotocol for urban meteorological networks [J]. *Bulletin of the American Meteorological Society*, 2013 94(8): 1161–1185.
[33] Costello, M. J. Distinguishing marine habitat classification concepts for ecological data management [J]. *Marine Ecology Progress Series*, 2009, 397: 253–268.
[34] Amante, M. J, Correia, A. M. R., Wilson, D. Information policy in the EU: legislative framework in Portugal (1989–1992) [J]. *Cadernos BAD*, 1994(2): 9–28.
[35] Lemke, A. A., Wolf, W. A., Hebert-Beirne, J. Public and biobank participant attitudes toward genetic research participation and data sharing [J]. *Public Health Genomics*, 2010, 13(6): 368–377.
[36] Itermon, R., Relyea, H. C. Information Policy [M]. Encyclopedia of Library and Information Science (Volumn 48), Kent, Allen. ed. New York: MaroelDekker, 1991: 176–204.
[37] Shuler, J. A. Citizen-centered government: information policy possibilities of the 108th Congress [J]. *Journal of Academic Librarianship*, 2003, 29(2): 107–110.
[38] Hardwicke, T. E., Mathur, M. B., MacDonald, K. N. G., *et al*. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal cognition [J]. *Royal Society Open Science*, 2010, 5(8): 180448.
[39] Moen, W. E. Interoperability of information access: Technical standards and policy considerations [J]. *Journal of Academic Librarianship*, 2000(2): 129–132.
[40] Michener, W. K. Ecological data sharing [J]. *Ecological Informatics*, 2015, 29: 33–44.
[41] Nakayama, T. Evidence-based healthcare and health informatics: derivations and extension of epidemiology [J]. *Journalof Epidemiology*, 2006, 16(3): 93–100.
[42] Cech, T. R., Eddy, S. R., Eisenberg, D., *et al*. Sharing publication-related data and materials: responsibilities of authorship in the life sciences [J]. *Plant Physiology*, 2003, 132(3): 19–24.
[43] Parr, C. S., Cummings, M. P. Data sharing in ecology and evolution [J]. *Trends in Ecology and Evolution*, 2005, 20 (7): 362–363
[44] Fienberg, S. E., Martin, M. E., Straf, M. L. Sharing Research Data [M]. Washington, D. C: National Academy Press, 1985.
[45] Constant, D., Kiesler, S., Sproull, L. What's mine is ours, or is it? A study of attitudes about information sharing [J]. *Information Systems Research*, 1994, 5: 400–421.
[46] Matzler, K., Renzl, B., Muller, J., *et al*. Personality traits and knowledge sharing [J]. *Journal of Economic Psychology*, 2008, 29: 301–313.
[47] McCain, K. Mandating sharing: journal policies in the natural sciences [J]. *Science Communication*, 1995, 16: 403–431.
[48] Piwowar, H. A. Who shares? Who doesn't? Factors associated with openly archiving raw research data [J]. *PLoS One*, 2011, 6(7): e18657.
[49] SPARC applauds White House for Landmark Directive Opening up Access to Scientific Research [EB/OL]. http: //www.sparc.arl.org/. 2013-08-28.
[50] Zhu, J., Wright, G., Wang, J., *et al.* A critical review of the integration of geographic information system and building information modelling at the data level [J]. *ISPRS International Journal of Geo-Information*, 2018, 7: 66.
[51] Sacks, R., Ma, L., Yosef, R., *et al*. Semantic enrichment for building information modeling: procedure for compiling inference rules and operators for complex geometry [J]. *Journal of Computing in Civil Engineering*, 2017, 31: 04017062.
[52] Irizarry, J., Karan, E. P., Jalaei, F. Integrating BIM and GIS to improve the visual monitoring of construction supply chain management [J]. *Automation in Construction*, 2013, 31: 241–254.
[53] Amirebrahimi, S., Rajabifard, A., Mendis, P., *et al*. BIM-GIS integration method in support of the assessment and 3D visualisation of flood damage to a building [J]. *Journal of Spatial Science*, 2016, 61: 317–350.
[54] Kang, T. W., Hong, C. H. A study on software architecture for effective BIM/GIS-based facility management data integration [J]. *Automation in Construction*, 2015, 54: 25–38.
[55] Ozel, F. Spatial databases and the analysis of dynamic processes in buildings [C]. In Proceedings of the Fifth Conference on Computer Aided Architectural Design Research in Asia, Singapore, 2000, 2: 97–106.
[56] Bradley, P. E., Paul, N. Using the relational model to capture topological information of spaces [J]. *The Computer Journal*, 2010, 53: 69–89.
[57] Guan, X., van Oosterom, P., Cheng, B. A parallel N-dimensional space-filling curve library and its application in massive point cloud management [J]. *ISPRS International Journal of Geo-Information*, 2018, 7: 327.
[58] VoPham, T., Hart, J. E., Laden, F., *et al*. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology [J]. *Environmental Health*, 2018, 17: 40.
[59] Wang, L. G. Reference model for creating information security policy [C]. Proceedings of Information Technology and Environmental System Sciences, 2008, 2: 279–281.
[60] Tang, Y. L., Xu, G. A., Niu, Y. X., *et al*. Information security risk analysis model based on entropy [C]. Proceedings of Information Technology and Environmental System Sciences, 2008, 4: 1146–1150.