Research on Metadata in the Era of Big Data Based on Bibliometric Analysis by CiteSpace

Shimin Yan(✉)

School of Public Administration, Sichuan University, Chengdu, People's Republic of China
yanshimin@stu.scu.edu.cn

Abstract. Metadata is an indispensable element for the development of big data. In order to explore the construction of metadata in the context of big data and to provide reference for future metadata research, this paper uses 989 papers under this topic included in the Web of Science core database between 2001 and 2021 as the data source, and examines the external characteristics and content characteristics of the literature with the help of bibliometric methods and the visual analysis software CiteSpace. Through the study, this paper concludes that the current cooperation is relatively fragmented and there is less cross-regional cooperation; the research heat of metadata is on the rise and will still continue for some time, and the hot spots of metadata research in recent years are in the fields of data mining and machine learning, and there is a trend of developing to finer branches such text mining.

Keywords: Metadata · Big Data · Bibliometric Analysis · Citespace

1 Introduction

In *The Third Wave* Alvin Toffler refers to Big Data as "the colorful music of the third wave" [1]. The mining and application of big data heralds the arrival of a new wave of productivity growth and consumer surplus [2]. *Nature* launched a special issue on big data in 2008, followed by *Science* in 2011. Big Data has become a hot spot of continuous concern for global academia and industry, and has also received widespread attention from countries around the world, and big data strategies have been launched at the national level one after another [14]. Metadata is the data that describes the data, which provides a description of the data and enables machines to interpret and use the data accurately. For big data analysis, the existence of metadata as a form of data attached to the data is necessary and indispensable. The splitting, reorganization, analysis and mining of data require the participation of metadata [4, 9]. Metadata research is an essential part of big data research, and metadata innovation should be developed together with the progress of data science.

The history of the development of metadata shows that it is itself a continuous innovation. In its early days, the concept of metadata referred to the use of a data element to describe or represent some characteristic of another data element [6, 7] As the term "metadata" spread, experts and scholars defined it in a variety of ways, with

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 246–255, 2023.
https://doi.org/10.2991/978-94-6463-034-3_25

Greenberg describing metadata as "structured data about objects that support functions associated with a specified object"; Pomerantz argues that "metadata is a statement about potential information objects". In addition, there are more targeted definitions of metadata in specific domains. The change of definitions indicates the change of experts and scholars' understanding of the functional role of metadata, which is gradually being recognized for its functions in data research and analysis, in addition to describing data. In terms of metadata standards, general metadata standards such as Dublin Core metadata terms, DataCite metadata framework, Data Catalog Vocabulary (DCAT). Professional domain standards such as getty vocabularies, JATS, ISO 19115 geographic information-metadata, Darwin Core (DwC). These metadata standards have played an important role in resource construction and data management. The "Chinese Modern Documentary Image Database" built by Nanjing Library refers to the DC metadata terms and the actual condition of the resources for metadata production; Deng et al. produced metadata for UAV remote sensing data by referring to the standards such as Geographic Information Metadata [12, 13]. These research examples will demonstrate how metadata can manage and describe a large amount of data, especially how to unify data structure and description in the case of diverse data sources. In addition, the research of scholars such as Gan Xiyu illustrates that metadata has an important role in reducing redundant data, reducing data maintenance costs, improving the data life cycle, and maximizing the value of data [10].

Big data has the 3V characteristics of volume, variety, and velocity. For the development and innovation of metadata, the 3Vs bring challenges. For the design of metadata standards: the large volume of data and the wide range of data sources require metadata standards to be fully extensible and reusable; the diversity of data types brings the need to use multiple metadata standards for description, but this faces the problem of poor interoperability among metadata standards. For the management of metadata: the rapid transformation of data causes problems for real-time capture, update and management of metadata; big data implies big metadata, which brings difficulties for storage and processing of metadata; for better further analysis of large amount of data, metadata needs to give better answers in semantic association problems [5, 8, 10]. In the past, a considerable number of experts and scholars have explored these issues. And this paper hopes to explore the research topics and research hotspots based on bibliometrics and through the analysis of previous studies, and provide references and suggestions for future research in the field of metadata.

2 Materials and Methods

2.1 Data Resource and Retrieval Conditions

The data source for this paper is the Web of Science database, a large multidisciplinary core journal citation database covering 8,500 scholarly journals in the natural sciences, engineering, social sciences, arts and humanities. In this paper, the time range was set from 2001-01-01 to 2021-12-31, and the search formula was: TS = ((big data) AND metadata). The number of search results was 1002, and 989 articles were obtained by removing the data of news, books and letters, and removing the documents with obvious deviation from the topic. All of these 989 articles were exported in plain text format, and all record items provided by WoS were exported.

2.2 Research Methods

This paper uses bibliometrics as a methodological theoretical guide and CiteSpace for data visualization to analyze the cooperative network of metadata-related literature and keywords, respectively, in the context of big data.

3 Results and Analysis

3.1 Issuance Trends

The results of the literature publication volume statistics on the topic of metadata in the context of big data are shown in Fig. 1. The results show that: from 2001 to 2009, the number of published articles showed an upward trend, but the overall level was low, with an average annual number of published articles of about 5. After the sudden decline in 2010, the number of published articles in the field surged from 2011 to 2018; After reaching the peak of 148 articles in 2018, the number of published articles began to decline, but remained at a high level by 2021.

In terms of the countries where the papers were published, USA was the country with the highest number of published papers, with a total of 314. The top three countries are the USA, China, and Germany. And the top eleven countries published a total of 974 articles, accounting for 97.2% of the total. This indicates that the vast majority of metadata research in the context of big data is concentrated in these countries (Fig. 2).

3.2 Cooperation Network Analysis

(1) Analysis of country cooperation networks.

In CiteSpace analysis software, select the parameter "country", and the co-occurrence network diagram of metadata-related big data research is shown in Fig. 3. The size of the nodes in the graph represents the number of articles issued in the country, the larger the node, the more the number of articles issued. As far as the country collaboration network is concerned, the three countries that have the highest number of co-occurrence are USA, ENGLAND, and AUSTRALIA. Among them, the USA ranks first in terms of influence and is far more central than other countries, twice as much as the second

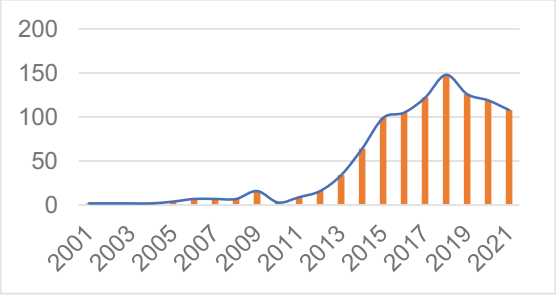

Fig. 1. Annual volume of publications on Big Data & Metadata search topics on Web of Science.

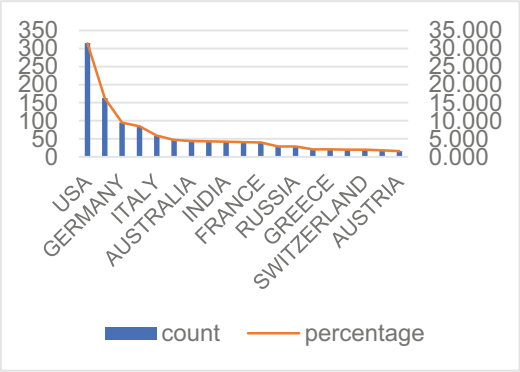

Fig. 2. National distribution of publications.

Fig. 3. Country Cooperation Network Map.

ranked UK.US scholars and research institutions have made outstanding contributions to metadata research in the context of Big Data, and are the core region for research in this area. It is worth mentioning that although China ranks second in terms of the number of publications (156), it is less central and less influential than the UK and Australia, which have more publications. Table 1 shows the top ten countries in terms of centrality.

(2) Analysis of institutional cooperation network.

Figure 4 shows the collaboration between institutions studying metadata in the context of big data. The thickness of the connecting lines represents the intensity of the collaboration between institutions, the thicker the inter-institutional collaboration the more. Among them, Univ Michigan is an important institution, and it is a pivot node for the cooperation network of European and American institutions. In addition, Australian Nalt Univ, UNIV Calif San Diego, and Chinese Acd Sci are more central, and they are also quite important institutions in the whole cooperation network. We found a total of 1620 institutions contributing to the field of metadata in the context of big data. This reflects that there is indeed a broad need for metadata research in the era of big data, and a considerable number of institutions are working on it. However, in terms of institutional cooperation co-existence, the geographical nature of inter-institutional cooperation is obvious, the connection between different sub-networks is not strong, and the value of

Table 1. Country intermediary centrality results.

Rank	Centrality	Year	Countries
1	0.24	2001	USA
2	0.12	2006	ENGLAND
3	0.10	2013	AUSTRALIA
4	0.07	2007	GERMANY
5	0.07	2007	PRCHINA
6	0.06	2009	FRANCE
7	0.06	2007	BELGUIM
8	0.05	2014	NETHERLANDS
9	0.05	2009	MALAYSIA
10	0.05	2009	CANADA

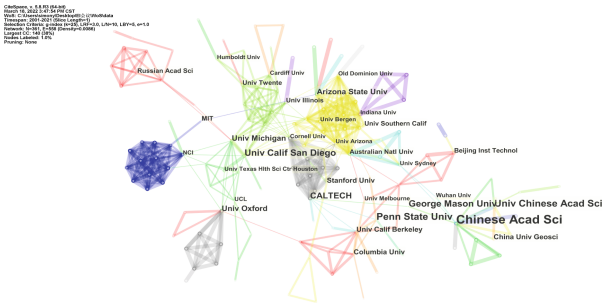

Fig. 4. Organizational cooperation network diagram.

network density is only 0.0096. For metadata research in the context of big data, a unified standard and management model will make it easier to interoperate and analyze data at a later stage. Therefore, we hope that future research into different regions will be more related in the future. Table 2 shows the ten institutions with strong influence.

(3) Analysis of Author Collaboration Network.

Authors are affiliated with the institutions they belong to, so author co-occurrence networks will have similarities with institutional co-occurrence networks. The author co-occurrence network as shown in Fig. 6 also suffers from the lack of communication between sub-networks. The team led by CHAOWEI YANG, who focuses on computer science and physical geography, is the most prominent in the network diagram. Their research on metadata focuses on the problem of metadata description for large-scale data, and the utilization of metadata in retrieval (Table 3 and Fig. 5).

Table 2. Most Productive Institutions.

Rank	Count	Centrality	Institute
1	30	0.05	Chinese Acad Sci
2	14	0.01	Penn State Univ
3	13	0.00	Uni Chinese Acad SSci
4	10	0.06	Univ Calif San Diego
5	9	0.01	George Mason Univ
6	8	0.03	CALTHCE
7	8	0.08	Univ Michigan
8	7	0.04	Univ Oxford
9	6	0.02	Arizona State Univ
10	6	0.02	Univ Calif Berkeley

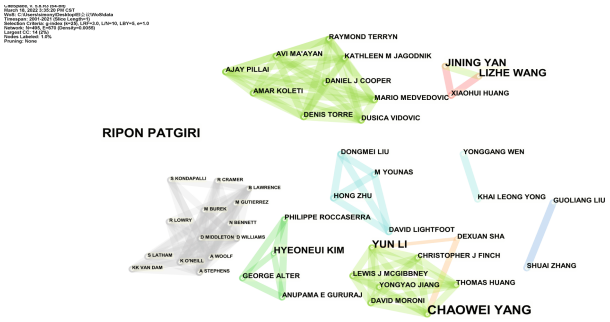

Fig. 5. Author Collaboration Network.

4 Keyword Analysis

For an article, keywords usually describe the core content. Often, keywords also address cutting-edge developments in related fields. If each word appears frequently in a certain period, it can be judged that the word reflects the significant present content of the field in that period [14].

Citespace can extract the frequency of keywords involved in a paper by analyzing the frequency of word occurrences using the statistical principles of metrics, and display the keywords or clustering relationships in the form of graphical labels. In citespace, the node type is set to keywords, and the merged network is cropped using the pathfinder network algorithm (Pathfinder) to adjust the node color, layout, etc. To generate Fig. 6. The size of the nodes indicates the frequency of keyword occurrences, and the color of the graph from light to dark indicates the time from far to near. In general, big data appears most frequently, followed by model, system, and the three are at the top of the list, and are important objects for metadata research in the context of big data. From

Table 3. Most prolific writers.

Rank	Count	Name	Year
1	6	RIPON PATGIRI	2003
2	6	CHAOWEI YANG	2014
3	4	NONG XIAO	2014
4	4	LIZHE WANG	2017
5	4	FANG LIU	2013
6	4	YUN LI	2009
7	4	MARIO JOSE DIVAN	2017
8	4	ALBERTO ABELLO	2011
9	3	HYEONEUI KIM	2013
10	3	CHRISTOPHER J FINCH	2015

the color of nodes and links, the budding period of metadata research in the context of big data is 2001–2012, which is studied in big data, semantic web, data management, database, etc.; the development period is 2013–2017, which is studied in data analysis, information system, cloud computing, natural language processing, etc.; the deepening period is 2018 -2021 where metadata is explored in data mining, machine learning, artificial intelligence, etc. In addition, words such as classification, network are also impression node of keyword co-occurrence graph, indicating that these nodes are also the focus of research in these literatures. In the timeline network, the horizontal axis mainly reflects time, and the vertical axis shows the names of keyword clusters. The larger the node, the more frequently the keyword appears, and the keyword is the hot spot of that time period. The darker the color of the node, the closer the research is to the present, and it may be the future research direction. As shown in Fig. 7, big data, management, modle, science and system have a high centrality. Meanwhile, archive management, open system, digital library, and bibliographic system are the first keywords that appeared in 2005. We can conclude that the need for metadata research in the big data perspective has evolved from library and archival management to web-based knowledge organization to data mining, machine learning, and other computer technologies for processing and utilizing data (Table 4).

The keyword burst not only reflects the shift in research focus, but also demonstrates the frontiers of research in the field. Figure 7 shows the top ten keywords with the

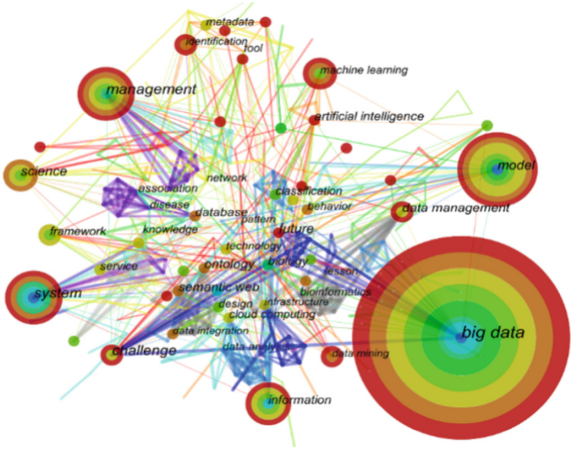

Fig. 6. Keyword co-occurrence network diagram.

Table 4. Top 10—keyword co-occurrence.

Rank	Count	Centrality	Keyword
1	188	0.8	big data
2	37	0.11	model
3	27	0.13	system
4	23	0.1	management
5	20	0.04	information
6	17	0.06	science
7	16	0.04	data management
8	14	0.04	machine learning
9	12	0.02	framework
10	12	0.03	cloud computing

strongest citation bursts. The top three strongest citation bursts are framework (2018–2019), linked data (2016–2017), and metadata (2018–2021). Metadata is also the keyword with the longest burst and closest to the present time. From the ten keywords with the strongest citation outbreak, the correlation between research keywords from 2001 to 2010 is low and there is no obvious citation. The keywords citation bursts started in 2011, but none of the outbreak lasts long, indicating that the research hotspots change quickly and there may be branch studies that are not explored and analyzed. Meanwhile, the citation of metadata as a keyword appears to explode in 2018–2021, indicating that metadata research fervor has increased significantly in recent years, the importance of metadata is gradually recognized, researchers treat metadata as a key to research, and metadata may be a continuous research hotspot in the future (Fig. 8).

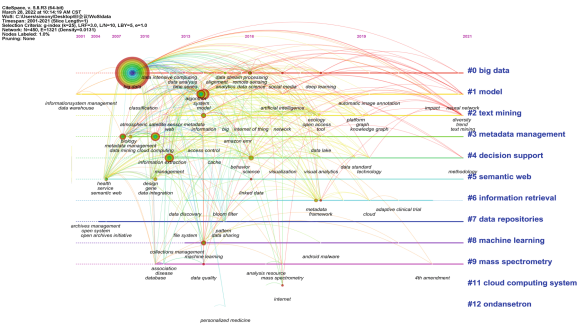

Fig. 7. Keyword time zone line graph.

Top 10 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2001 - 2021
database	2001	2.37	**2011**	2012	
semantic web	2001	2.78	**2015**	2017	
data sharing	2001	2.38	**2015**	2016	
big data analytics	2001	2.29	**2015**	2016	
linked data	2001	3.63	**2016**	2017	
file system	2001	3.02	**2017**	2018	
analytics	2001	2.25	**2017**	2018	
framework	2001	4.59	**2018**	2019	
metadata	2001	3.18	**2018**	2021	
science	2001	2.47	**2018**	2019	

Fig. 8. Keyword reference bursts.

5 Conclusion

Metadata, as a form attached to data, is extremely important for enhancing the value of data. In this paper, by reviewing the research articles in the field of metadata in the context of big data from 2001–2021, we summarize some conclusions through bibliometric and visualization methods: (1) Metadata research is gradually becoming a hot topic and will still continue for some time. Research related to metadata in the context of big data has shown growth over the past two decades, and although the number of studies has fallen back since 2018, the citation burst of metadata appearing as a keyword has continued for four years. There is no doubt that the number of articles in related fields will remain considerable in the future. (2) The United States leads the research in this field, ranking first in both the number of articles and intermediary centrality. It can be predicted that the U.S. will continue to lead the research in this field in the future. In terms of collaborative networks, there is less institutional cross-regional and cross-national collaboration in this field, and enhanced collaboration may be a breakthrough point for new results to emerge in the future. (3) Metadata is being studied by scholars in many fields, especially computer science, engineering, information science, and library science. And judging from the time line of research hotspots and the depth of research, metadata research is gradually emerging in finer-grained segments. This trend is likely to continue in the future.

References

1. Alvin Toffler. (2006). *The Third Wave*, CITIC Press. Beijing.
2. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh &
3. Angela Hung Byers. (2011). Big data: The next frontier for innovation, competition, and productivity. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.
4. Jane Greenberg. (2017). Big Metadata,Smart Metadata,and Metadata Capital:Toward Greater Synergy Between Data Science and Metadata. *Journal of Data and Information Science*(03),19–36.
5. Lancheng Wang, Xiaoliang Liu, & Yongqin Huang. (2019). Research on metadata construction and integration technology in archival social media information integration. Archival Research (05), 102-107.
6. Mayernik, MS. (2020). Metadata. J. Knowledge Organization(47), 696–713.
7. Norman, J. (2019). Philip Bagley coins the term metadata. http://www.historyofinformation.com/detail.php?entryid54241.
8. Patgiri, R, Dev, D & Ahmed, A. (2018). dMDS: Uncover the Hidden Issues of Metadata Server Design. Advances in Intelligent Systems and Computing(518),531–541.
9. Qianqian Yu, Jianyong Zhang & Yongwen Huang. (2018). Analysis of the design characteristics of document metadata standards in the big data environment. Library Journal (11), 35–39+46.
10. Siyu Gan, Pinjue Che, Tianshun Yang & Junwei Wu. (2018). Big data governance system. Computer Applications and Software (06), 1–8+69.
11. Synergy Between Data Science and Metadata. *Journal of Data and Information Science*(03),19–36.
12. WJ Ding. (2018). The practice of image digital resource metadata storage construction - taking Nanjing Library as an example. New Century Library(03),64–68.
13. Xiaoming Deng, Xiaohan Liao, Huanyin Yue, Chenchen Xu & Huping Ye. (2020). Design and practice of metadata for UAV remote sensing data sharing. Remote Sensing Information (06), 99-104.
14. Yu, DJ, Xu, ZS, Pedrycz, W & Wang, WR. (2017). Information sciences 1968–2016: A retrospective analysis with text mining and bibliometric. Information Science(418), 619–634.Zhiyan Feng,
15. Xunhua Guo, Dajun Zeng, Yubo Chen & Guoqing Chen. (2013). Some frontiers of business management research in the context of big data. Journal of Management Science (01), 1-9.