ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00637

Research on the Creation and Use of Social Science Data: Postprint on the Application of Secondary Matching Data Collection Rules

Authors: Chen Xin, Cao Chaojin, Ye Chunsen, Wang Chuanlei

Date: 2023-04-01T16:02:49+00:00

Abstract

[Purpose/Significance] Within the data lifecycle framework, this study innovatively proposes a method for collecting information on the creation and use of social science data from academic papers, and conducts an in-depth investigation of its fundamental characteristics, thereby providing new insights for social science data research.

[Method/Process] Taking CSSCI-indexed papers from 2015-2020 in the highly interdisciplinary logistics research field as samples, we construct a "generalization-precision keyword lexicon" based on the data lifecycle through an iterative approach, collect relevant information on social science data, and conduct a comprehensive study on the creation and use of social science data by incorporating external environmental information.

[Results/Conclusions] Regarding the collection of information on the creation and use of social science data from papers, secondary matching data collection rules demonstrate feasibility and high efficiency. The Internet has become the primary data collection method for social science research. Different research topics exhibit distinct data usage preferences, and the adoption rate of data analysis tools remains relatively low.

Full Text

Preamble

Volume 65, Issue 10, May 2021

ChinaRxiv 中X

Research on the Creation and Use of Social Science Data: Application of Twice-Matching Data Acquisition Rules

Chen Xin[1], **Cao Chaojin**[2], **Ye Chunsen**[1], **Wang Chuanlei**[1] [1] School of Business, Anhui University, Hefei 230009 [2] School of Management, Hefei University of Technology, Hefei 230009

Abstract: [Purpose/Significance] Under the framework of the data lifecycle, this study innovatively proposes a method for collecting information related to the creation and use of social science data from academic papers, and conducts an in-depth investigation of its basic characteristics, providing new insights for social science data research. [Method/Process] Taking CSSCI-indexed papers from 2015–2020 in the highly interdisciplinary logistics research field as the sample, this study constructs a "generalized-precision keyword thesaurus" based on the data lifecycle through an iterative approach, collects relevant information on social science data, and conducts a comprehensive study of its creation and use in combination with external environmental information. [Result/Conclusion] The twice-matching data acquisition rules demonstrate feasibility and efficiency in collecting information on the creation and use of social science data from papers. The Internet has become the primary data collection method in social science research, different research topics exhibit varying preferences for data use, and the adoption of data analysis tools remains relatively low.

Keywords: social science data; generalized-precision thesaurus; twice-matching data acquisition rules; Python; bibliometrics

Classification Number: G203
DOI: 10.13266/j.issn.0252-3116.2021.10.010

Data is considered a primary resource for promoting innovation across various fields [1], especially in this era of explosive data growth where both business and academic domains are experiencing the impact of big data. Scientific data constitutes an essential component of the scientific research process, serving as both an outcome and a foundation for scientific inquiry. Scholars increasingly emphasize data-driven research, particularly in natural sciences such as life sciences, earth sciences, and geographical sciences [2]. To standardize and mature scientific data management, the General Office of the State Council issued the "Measures for the Management of Scientific Data" (hereinafter referred to as the "Measures") on March 17, 2018 [3]. However, the "Measures" primarily target natural sciences, engineering, and technical sciences, without explicitly addressing scientific data management in social science research.

In recent years, social science scholars have increasingly employed large-scale data analysis methods, complex mathematical models, and diverse data analysis tools [4], highlighting the significant value and role of scientific data in this domain. Social science data management therefore warrants serious attention. Social science data can be broadly defined as any data related to social science fields, such as social survey data, government statistics, and commercial open

data, or more narrowly as various data generated from social science research activities, including textual records, numerical statistics, and image data [5]. The diversity of research methods and data formats in social sciences—not limited to numerical data but also encompassing textual data, archival data, compiled data, PDF-format data, as well as micro- and macro-scale data—has resulted in a lack of uniform standards. This has led to poor utilization of scientific data in this field, with data scattered among individual researchers and organizations, reflecting the diversity and uncertainty of social science research [6]. Moreover, current research on social science data remains limited, with insufficient understanding of its characteristics, creating a lack of practical basis for formulating social science data management policies and hindering data management and services.

This study focuses on the logistics research field within social sciences, using CSSCI-indexed papers as samples to analyze the external environment of social science data from a bibliometric perspective. Based on the data lifecycle framework, we employ twice-matching rules to collect relevant information on social science data from papers and analyze the specific characteristics of its creation and use. Combined with the external environment, we discuss the usage preference relationships of social science data across different publishing institutions and research hotspots, and analyze the patterns of its creation and use.

2 Research Status of Social Science Data

2.1 Management and Service Research on Social Science Data

Current research on social science data management and services is relatively scarce compared to natural science data, which has more distinct characteristics and thus more management policies both domestically and internationally, such as China's "Measures" [3], NASA's Data & Information Policy [7], and the UK Biotechnology and Biological Sciences Research Council's BBSRC Data Sharing Policy [8]. Consequently, China's open scientific data sharing resources remain relatively homogeneous, predominantly comprising natural science data [9]. Current practices in social science data management in China suffer from several issues, including oversimplifying data management as routine documentation work [5]. Therefore, research on the creation and use of social science data is essential, as only through thorough understanding of its characteristics can we develop effective scientific data management and service policies tailored to social science research.

2.2 Research on Characteristics and Nature of Social Science Data

Research on the characteristics and nature of social science data has primarily focused on specific databases or metadata repositories (Data Citation Index, DCI). For instance, Luo Pengcheng et al. analyzed scientific data characteristics across temporal and spatial dimensions based on DataCite [10]; Meng Xiangbao et al. examined structural features of scientific data in five disciplines including

history and education within DCI [11]. Some scholars have used journal papers as samples, manually conducting content analysis to collect scientific data-related information. For example, Shen Tingting analyzed papers published in *Social Sciences in China*, examining researchers' data acquisition channels and data types, and proposed recommendations for library scientific data services [12]. However, existing research has limitations: first, social science research rarely utilizes scientific databases like DataCite or DCI, with data sharing being even rarer. Most social science researchers store data on personal computers and share data through informal channels, with only 46% using repositories for data sharing [13]. Therefore, analyzing Chinese social science data characteristics using scientific databases may lack representativeness. Second, while papers serve as representations of research outcomes from which social science data information can be extracted, existing paper-based studies predominantly rely on manual content analysis, which limits sample size and extraction accuracy, presenting methodological constraints.

2.3 Application of Python in Text Analysis Research

Python, as one of today's most popular programming languages, is frequently used for text analysis in scientific research. For example, Tan Chunlin et al. used Python programming to mine textual content from journal papers [14]; Zhang Na et al. employed Python's snowNLP module for opinion mining, classifying textual data into positive and negative categories [15]; Liu Yulin et al. utilized Python for text sentiment analysis of e-commerce online reviews [16]. These studies demonstrate that Python's text analysis capabilities have become increasingly sophisticated. Therefore, based on Python's characteristics and applications, this study proposes a logical approach for collecting social science data-related information: first constructing a "generalized-precision keyword thesaurus" based on the data lifecycle through an iterative method, then designing a Python-based twice-matching data acquisition rule that efficiently collects relevant information from papers by combining the thesaurus.

3 Research Design

3.1 Research Framework

The research framework is illustrated in Figure 1 [Figure 1: see original paper]. This study first retrieved logistics field literature from CSSCI within the specified timeframe as research samples. Under the data lifecycle-based data collection framework, we constructed a thesaurus and used it with twice-matching data acquisition rules to obtain data on the creation and use characteristics of social science data. Bibliometric analysis was employed to analyze the external environment of social science data, including publishing institutions, authors, publication dates, and research hotspots. Statistical analysis was then used to examine creation and use characteristics, specifically covering six aspects: creation subjects, creation methods, data types, data analysis methods, data analysis tools, and fuzzy word usage. When analyzing these characteristics,

we incorporated external environmental perspectives to comprehensively study social science data features across creation and use dimensions.

3.2 Sample Sources

This study selected CSSCI-indexed logistics research papers as samples based on the following considerations: (1) As representations of research outcomes, papers generally contain descriptions of the entire process from data collection to use, facilitating the extraction of information related to scientific data creation and use; (2) CSSCI, as a representative academic database in China, is frequently chosen by scholars as a sample source [17-19]; (3) The logistics research field exhibits strong interdisciplinary and cross-industry characteristics within social sciences, making it representative of social science research. Using logistics research as a pilot study lays the foundation for subsequent research on scientific data creation and use in other social science disciplines.

To improve recall and precision rates, we conducted subject searches in CNKI, Wanfang, and Chongqing VIP databases using the seven functional elements of logistics ("transportation, warehousing, loading/unloading, packaging, distribution processing, distribution, and information processing") plus "logistics" and "supply chain." Other search conditions are shown in Table 1 . After removing duplicate and irrelevant papers, we obtained 4,114 sample papers.

3.3 Data Collection

3.3.1 Data Lifecycle-Based "Generalized-Precision Keyword Thesaurus" Collecting scientific data-related information from different research fields and collection units requires different thesauri. To ensure completeness, this study proposes an iterative thesaurus construction method, illustrated in Figure 2 [Figure 2: see original paper].

(1) **Initial Thesaurus Construction:** Based on preliminary content analysis of scientific data-related information in logistics research papers published in CSSCI by Anhui universities over the past decade, combined with prior research on each collection unit, we initially identified generalized keywords. Using Python, we extracted sentences containing generalized keywords from sample papers, performed word segmentation, and identified frequently co-occurring terms to determine precision keywords. Both generalized and precision keywords were stored in the thesaurus to form the initial version.

(2) **Pre-retrieval Validation and Thesaurus Updating:** Using the initial thesaurus and twice-matching rules for pre-retrieval yielded three paper categories: (1) papers without generalized keywords, (2) papers with only generalized keywords, and (3) papers with both generalized and precision keywords. Systematic sampling of 20% from each category was conducted. For category (1), content analysis was performed to identify and supplement new generalized keywords; for category (2), context analysis was

used to supplement precision keywords; for category (3), context analysis assessed precision keyword effectiveness, deleting those with matching accuracy below 90%. The iteration concluded when no new keywords could be added and all precision keywords demonstrated effectiveness, outputting the final thesaurus for the data collection unit.

The final data lifecycle-based generalized-precision keyword thesaurus is presented in Table 2 . The thesaurus contains two keyword types: "generalized keywords" refer to common terms in papers describing research samples, materials, or data (used beyond these contexts), serving to narrow search scope; "precision keywords" further pinpoint social science data-related information based on generalized keyword retrieval. For example, to determine if a paper's data originates from statistical bureaus, the generalized keyword would be "统计局" (statistical bureau), but its presence alone doesn't confirm data sourcing. Through literature review, we found that when context includes both "统计局" and precision keywords like "数据来源" (data source), "获取" (obtain), or "查阅" (consult), the paper's data source can be confirmed as a statistical bureau.

3.3.3 Python-Based Twice-Matching Data Acquisition Rules The twice-matching data acquisition rules for scientific data are illustrated in Figure 3 [Figure 3: see original paper]:

Step 1: Select the thesaurus for the first data collection unit.
Step 2: Conduct the first search across all 4,114 papers using "generalized keywords" from the thesaurus.
Step 3: Determine if papers contain generalized keywords. If yes, extract relevant sentences and record them in Excel; otherwise, proceed to the next paper.
Step 4: Search using "precision keywords." If present, mark the paper as containing the assumed scientific data information and write the mark to the database; otherwise, mark it for manual judgment. Manually verify whether the paper contains the assumed information, marking and recording it in the database if confirmed; otherwise, proceed to the next paper.
Step 5: Determine if all 4,114 papers have been searched. If yes, select the thesaurus for the next data collection unit and return to Step 2; otherwise, continue searching the next paper.

4 Analysis of the External Environment of Social Science Data

Analyzing the external environment of social science data—examining publishing institutions, authors, publication dates, and keywords—helps understand the basic conditions of logistics research and, when combined with the external environment, facilitates deeper understanding of social science data creation and use.

4.1 Publishing Institution Analysis

Analysis of first-author affiliations revealed that universities account for 97% of publications, with other organizations (companies, research institutes, government departments) comprising only 3%, indicating that universities dominate social science research. Further analysis of the top 20 productive institutions (Figure 4 [Figure 4: see original paper]) shows Beijing Jiaotong University ranking first, followed by Chongqing University, Shanghai Maritime University, Southwest Jiaotong University, Central South University, Dalian Maritime University, Renmin University of China, Beijing Wuzi University, Northeastern University, and Xi'an Jiaotong University. These productive institutions' research represents the frontier of domestic logistics development and will be analyzed in conjunction with their scientific data usage characteristics.

4.2 Author Analysis

The sample papers involved 10,138 authors, with an average of 2.47 authors per paper. As shown in Table 3 , 80% of papers were completed by 2-4 authors, indicating that logistics research tends toward collaboration, particularly 2-3 author teams. Further analysis of productive authors (Table 4) shows Song Hua (Renmin University of China) leading in publications, focusing on supply chain finance; Tang Jianrong (Jiangnan University) ranking second, with research emphasizing regional logistics and performance evaluation; and Liang Wen (Anhui University) third, focusing on rural logistics and coordinated development. Other productive authors include Wang Daoping, Dan Bin, Wang Wenbin, Zhang Jianjun, Pu Xujin, Li Xinran, and Yan Bo.

4.3 Publication Date Analysis

A discipline's development speed and level can be observed through the temporal distribution of its literature. With the search cutoff date of March 1, 2020, and university-affiliated papers accounting for 97% of the total, we focus on university publications (Figure 5 [Figure 5: see original paper]). The evolution of paper counts between 2015-2019 can be divided into three phases: (1) 2015-2017 saw an overall decline in both the number of logistics papers published by Chinese universities and the number of publishing institutions; (2) 2017-2018 witnessed a rebound due to major policy releases such as the "13th Five-Year Plan for Commercial Logistics Development," "Guiding Opinions on Actively Promoting Supply Chain Innovation and Application," and "Notice on Carrying Out Supply Chain Innovation and Application Pilots," which drove research hotspots like supply chain innovation and logistics cost reduction, increasing institutional attention; (3) 2018-2019 experienced another decline due to high-level journals tightening publication quotas, intensive early-stage research on logistics hotspots, and a shift of research outcomes to international journals. Overall, industry policy releases, CSSCI publication constraints, and the transfer of research results to international journals collectively contributed to the fluctuating publication numbers.

4.4 Keyword Burstiness Analysis

Examining the rise and fall of keywords over time reveals past trends and future directions within a research field, helping explain fluctuations in 2015-2019 publication numbers. Using CiteSpace's burst detection algorithm, keywords with stronger burst intensity indicate greater academic attention during specific periods [28] (Table 5).

Influenced by economic conditions and policies, 2015 saw the emergence of multiple logistics-related research themes such as e-commerce, food safety, and green logistics, resulting in the highest publication volume in recent years. From 2016, research themes concentrated on logistics information, Internet of Things, and "Internet Plus," dropping to 12 keywords, with 8 not persisting into 2017, further reducing that year's publication count. In 2018, research objects like logistics industry and logistics service supply chains gained prominence, while keywords like carbon trading and carbon tax from 2017 continued into 2018, boosting publication numbers. In 2019, high-burst-intensity keywords were all continuations from 2017-2018, with no new high-burst keywords emerging, leading to a publication decline.

4.5 Logistics Research Hotspot Analysis

Keywords provide concise summaries of a paper's themes and content. Co-occurrence analysis of keywords in a research field can identify hotspots [29-32]. Using CiteSpace's Pathfinder and Pruning the merged network pruning algorithms, we conducted co-occurrence analysis of logistics research keywords from 2015-2019 and created a visual knowledge map to explore hotspot areas [33] (Figure 6 [Figure 6: see original paper], Table 6).

Combined with productive authors and their research directions, the past five years' logistics research hotspots can be summarized as: (1) cross-border logistics; (2) closed-loop supply chain; (3) supply chain finance; (4) agricultural product logistics; and (5) green logistics.

5 Analysis of Social Science Data Creation and Use

Based on the proposed data acquisition method, we extracted information on scientific data creation and use from papers and conducted in-depth analysis combined with the external environment.

5.1 Data Creation

5.1.1 Scientific Data Source Analysis (1) Overview of Data Sources: Data sources refer to the channels through which scientific data is obtained in research. Based on sample analysis, we categorized data sources into three channels: scientific investigation, government disclosure, and commercial disclosure (Table 7). Usage frequency from low to high was commercial disclosure, government disclosure, and scientific investigation data.

In terms of acquisition difficulty, government and commercial disclosure data are relatively easy to obtain due to high openness and consistent formats. Scientific investigation data is more difficult to acquire due to varying survey methods, objects, and purposes, featuring diverse formats and characteristics of being numerous, small-scale, and scattered. Despite acquisition challenges, scientific investigation data is more popular among researchers due to its strong autonomy and research content specificity.

(2) Data Source Analysis of Productive Institutions: Research from productive institutions often reflects a field's development frontier and is representative of scientific data usage characteristics. Based on the institutional analysis above, we examined the top 10 productive institutions (Figure 7 [Figure 7: see original paper]). Each institution used scientific investigation data in over half of its research, with Northeastern University using it in 95% of papers, confirming scientific investigation as the primary data source method in logistics and broader social science research. Notably, three Beijing-based institutions—Beijing Jiaotong University, Renmin University of China, and Beijing Wuzi University—used more government and commercial disclosure data, indicating closer university-government collaboration and easier government data access in Beijing, as well as greater emphasis on utilizing government disclosure data and higher policy sensitivity.

5.1.2 Scientific Data Collection Method Analysis (1) Overview of Collection Methods: Data collection methods refer to how scholars gather scientific data. We categorized methods into online search and non-online search. Online search primarily involves searching professional databases, industry reports, publicly listed company data, statistical yearbooks, government documents, and other websites for logistics-related information. Non-online search mainly includes simulation experiments, questionnaires, field investigations, and interviews. Statistical results show that 56.52% of sample papers used online search, while 50.58% used non-online methods, indicating that online collection has become the primary data acquisition approach in logistics research.

(2) Analysis of Method Selection Across Years: To understand evolving researcher preferences, we analyzed the proportion of online versus non-online methods by year (Figure 8 [Figure 8: see original paper]). Non-online methods remained relatively stable over the years, while online search showed an increasing trend of nearly 15%. This suggests that with advancing computer network technology, more researchers are leveraging the Internet for data collection in scientific research, making e-Science one of the primary research environments in social sciences. Meanwhile, traditional social science survey methods maintain important positions in current research.

5.1.3 Scientific Data Type Analysis (1) Overview of Data Types: Existing research offers various data classification schemes. Based on data generation purpose, we categorized scientific data into primary and secondary data.

Primary data refers to data personally collected and processed by researchers through experiments, interviews, or questionnaires for the first time. Secondary data originates from others' investigations or scientific experiments [34]. By format, data can be divided into text, numerical, image, audio, etc. [11] (Table 8).

Overall, secondary data usage slightly exceeds primary data (75% vs. 65% of sample papers). Most data formats are text and numerical. Among primary data, model parameter data, example data, and simulation data constitute large proportions. Model parameter data represents prerequisite conditions set during model construction. Example data refers to data used to validate proposed models or conclusions, partially from real enterprises or departments and partially set by researchers according to model conditions. Simulation data comprises parameters used in simulation experiments. Other common primary data include questionnaire and interview data. Additionally, nearly all studies utilized journal text data, which is therefore not listed in Table 8.

(2) Data Type Preferences Across Research Hotspots: Examining papers on five hotspots—cross-border logistics, closed-loop supply chain, supply chain finance, agricultural product logistics, and green logistics—reveals distinct preferences (Figure 9 [Figure 9: see original paper]). Cross-border logistics papers used secondary data far more frequently than primary data, while closed-loop supply chain research favored primary data. Supply chain finance, agricultural product logistics, and green logistics showed relatively balanced usage. Regarding specific data types (Figures 10 [Figure 10: see original paper] and 11 [Figure 11: see original paper]), cross-border logistics preferred questionnaire data for primary data and government document data for secondary data. Closed-loop and green logistics research favored model parameter, example, or simulation data for primary data. Supply chain finance preferred enterprise disclosure data for secondary data. These varying preferences stem from differences in research foundations, objects, and methods, which directly influence data usage preferences that dynamically evolve with research depth.

5.2 Data Use

5.2.1 Scientific Data Analysis Method Analysis (1) Overview of Analysis Methods: Data analysis methods refer to techniques used for data processing and analysis in research. Statistical analysis of sample papers (Table 9) shows that example analysis, experimental methods, and statistical methods are the three most frequently used approaches, indicating that logistics research emphasizes real-world cases over pure theoretical studies, with quantitative research outpacing qualitative research. Big data modeling methods refer to machine learning algorithms developed for big data analysis, such as decision trees, support vector machines, and artificial neural networks [35]. Other methods include content analysis, social network analysis, and bibliometric analysis [36-37].

(2) Analysis Method Preferences Across Logistics Research Hotspots: Analysis of the five hotspots (Figure 12 [Figure 12: see original paper]) reveals that statistical methods are commonly used in supply chain finance, cross-border logistics, and agricultural product logistics, but rarely in closed-loop supply chain research, which prefers example analysis, experimental methods, and game theory analysis. This aligns with closed-loop supply chain characteristics, which emphasize theoretical research requiring example analysis for validation, while other hotspots focus more on empirical studies.

5.2.2 Scientific Data Analysis Tool Analysis (1) Overview of Analysis Tools: Analysis tools assist and accelerate research work. Statistical analysis of tools mentioned in papers (Table 10) shows that MATLAB, SPSS, and AMOS are among the most frequently used simulation and statistical analysis software.

(2) Analysis Tool Selection Across Years: As shown in Table 11 , MATLAB usage increased annually over the past five years. SPSS, STATA, and EViews—functionally similar statistical software—showed divergent trends: SPSS and EViews declined, while STATA rose. SPSS's high usage among beginners due to its ease of use may be declining due to limitations in handling cutting-edge statistical processes and data management scope. STATA, also user-friendly, has strengthened its data processing and management capabilities through recent updates, gaining ground on SPSS. EViews, despite its simplicity, suffers from poor extensibility and insufficient support for programming-intensive analyses, leading to its continued decline. ArcGIS and Python are emerging tools in logistics research. Notably, most analysis tools used in logistics research originate from science and engineering disciplines (e.g., MATLAB from mathematics, computer science, and electronics; SPSS, STATA, and EViews from econometrics; ArcGIS from geography and tourism; Python from computer science), demonstrating strong interdisciplinary characteristics in logistics and social science research.

5.2.3 Fuzzy Word Usage Analysis (1) Overview of Fuzzy Word Usage: Fuzzy words are terms with vague conceptual meanings used when describing scientific data. To investigate whether their usage stems from lack of precise data support, we analyzed eight types of fuzzy words. Table 12 shows that 70% of papers used fuzzy words, with "大多" (mostly) and "很多" (many) appearing most frequently. Further analysis of total and per-paper usage frequency (Figure 13 [Figure 13: see original paper]) reveals that "大量" (a large amount) and "很多" (many) are most widely used. While "很少" (few) has higher total usage than "若干" (a certain number), its per-paper usage is lower, indicating more concentrated usage of "若干." "差不多" (almost) and "无数" (countless) are significantly less popular—"差不多" being the most "fuzzy" among fuzzy words, and "无数" having overly absolute meaning, explaining their low usage in today's rigorous academic environment.

(2) Fuzzy Word Usage Across Years: Calculating the annual proportion

of papers using fuzzy words (Figure 14 [Figure 14: see original paper]) shows a slow decline over the past five years.

6 Conclusions and Implications

This study's analysis of the logistics research field within social sciences yields the following findings:

(1) **Methodologically**, this study innovatively proposes a method for collecting information on social science data creation and use from sample literature. This includes constructing a "generalized-precision keyword thesaurus" based on the data lifecycle through an iterative method, ensuring collection accuracy and comprehensiveness. The Python-based twice-matching rule effectively improves collection efficiency. Applying this method to 4,114 CSSCI-indexed logistics papers and comparing results with prior research demonstrates its feasibility and efficiency. The analysis reveals logistics research's interdisciplinary characteristics and the diversity of research objects, which leads to complex scientific data and reflects the complexity of social science research to some extent. This method's applicability in this field provides a reference for studying scientific data creation and use in other social science disciplines, enabling future large-scale studies and promoting understanding of social science data to inform more effective policy development.

(2) **Data source selection** shows descending usage frequency: scientific investigation data, government disclosure data, and commercial disclosure data, created by research institutions/researchers, government agencies, and industry associations/enterprises respectively. Scientific investigation data is most popular due to its research specificity obtained through questionnaires, interviews, and surveys, despite complex organization, storage, and acquisition challenges. Social science research also highly values government disclosure data, which is more accessible and authoritative than commercial disclosure data, evidenced by significantly higher usage frequency. Analysis of productive institutions reveals that Beijing-based universities use more government disclosure data, demonstrating greater policy sensitivity and emphasis on government data utilization.

(3) **Social science data types** are diverse, including primary data (simulation, questionnaire) and secondary data (government documents, statistical yearbooks), with secondary data usage exceeding primary data. This relates to changing data collection methods. Analysis shows that with advancing information technology, papers using online search for data collection are increasing. Notably, social science research exhibits strong interdisciplinary characteristics, with logistics studies frequently mixing multiple data types. Data collection and analysis have become increasingly important, leading scholars to prefer collaborative research in 2-3 person teams.

(4) **Analysis of different research hotspots** reveals varying data preferences related to analysis method selection. Cross-border logistics, supply chain finance, agricultural product logistics, and green logistics emphasize statistical methods, while closed-loop supply chain prefers example analysis, experimental methods, and game theory. This suggests that future scientific data organization and management could be categorized by research field and topic.

(5) **Social science data analysis tools** are numerous but overall usage remains low (only ~30% of studies). Notably, most tools are developed by foreign institutions or scholars, indicating high foreign dependency in China's academic research that requires strengthening independent development to reduce risks of foreign academic monopoly.

(6) **Fuzzy word analysis** shows that only four fuzzy words—"大量" (a large amount), "很多" (many), "很少" (few), and "若干" (a certain number)—are frequently used. Since fuzzy words are inherently imprecise, their frequent use in a discipline might suggest inadequate precise data support. However, statistical results show stable annual usage rates without significant fluctuation due to the big data and information era, indicating that frequent use of certain fuzzy words doesn't necessarily reflect poor data quality but may relate to linguistic and cultural conventions [12].

(7) **Bibliometric analysis** serves as an external environment analysis method. Combined with the proposed twice-matching method for collecting social science data creation and use information from papers, this study examines characteristics from both internal and external perspectives.

This study addresses the challenge of extracting social science data creation and use information from papers, using logistics research for analysis. Future research should improve thesaurus completeness, develop English versions based on the Chinese version, and adapt them across disciplines. Additionally, this method should be applied to large-scale studies across social science disciplines and extended to international English journals for comparative analysis of creation and use characteristics across disciplines, research paradigms, and fields, providing a comprehensive picture of Chinese social science data features.

References

[1] Peters I, Kraker P, Lex E, et al. Zenodo in the spotlight of traditional and new metrics[J]. Frontiers in research metrics and analytics, 2017, 2(1): 1-14.
[2] He L, Nahar V. Reuse of scientific data in academic publications[J]. Aslib journal of information management, 2016, 68(4): 478-494.
[3] General Office of the State Council. Notice of the General Office of the State Council on Issuing the Measures for the Management of Scientific Data[EB/OL]. [2020-03-20]. http://www.gov.cn/zhengce/content/201804/02/content_{5279272}.htm.
[4] Sun Jianjun. How should humanities and social sciences develop in the big

data era[N]. Guangming Daily, 2014-07-07(11).
[5] Xia Yikun. Realistic dilemmas and countermeasures for humanities and social sciences data management[J]. Information Science, 2020, 38(9): 14-22.
[6] Bollacker K, Housos N, Manghi P, et al. Data as "first-class citizens"[EB/OL]. [2020-08-13]. http://www.dlib.org/dlib/january15/01guest_{editorial}.html.
[7] NASA. Data & information policy[EB/OL]. [2021-01-23]. http://science.nasa.gov/earth-science/earth-science-data/data-information-policy/.
[8] BBSRC. BBSRC data sharing policy[EB/OL]. [2021-01-23]. http://www.bbsrc.ac.uk/about/policies/policy-foi/policy/data-sharing-policy/.
[9] Li Zhifang, Deng Zhonghua. Analysis of the distribution and characteristics of domestic open scientific data[J]. Information Science, 2015, 33(3): 45-49.
[10] Luo Pengcheng, Cui Haiyuan, Zhao Jingru. Research on scientific data characteristics based on DataCite[J]. Library and Information Science, 2019(3): 101-112, 80.
[11] Meng Xiangbao, Qian Peng. Research on humanities and social sciences data characteristics from the data lifecycle perspective[J]. Library and Information Science, 2017(1): 76-88.
[12] Shen Tingting. Analysis of scientific data usage characteristics in humanities and social sciences—empirical study based on sample papers from Social Sciences in China[J]. Journal of Academic Libraries, 2015, 33(3): 101-107.
[13] Meadows A. To share or not to share? That is the (research data) question[EB/OL]. [2020-05-21]. http://www.scholarlykitchen.sspnet.org/2014/11/11/to-share-or-not-to-share-that-is-the-research-data-question/.
[14] Tan Chunlin, Liu Qinghai. Text mining and analysis of journal editors' publication patterns[J]. Acta Editologica, 2019, 31(4): 407-410.
[15] Zhang Na, Liu Yunchang, Wang Ruonan. Social media data mining based on text sentiment analysis[J]. Journal of Henan University of Urban Construction, 2019, 28(5): 74-79.
[16] Liu Yulin, Jian Lirong. E-commerce online review data mining based on text sentiment analysis[J]. Statistics & Information Forum, 2018, 33(12): 119-124.
[17] Ren Heng. Knowledge mapping of domestic think tank research: current status, hotspots, and trends—bibliometric analysis based on CSSCI journals (1998-2016)[J]. Information Science, 2018, 36(9): 159-166.
[18] Feng Yafei, Hu Changping, Li Shuangshuang. Knowledge mapping and hotspot themes of domestic academic resource research[J]. Information Science, 2019, 37(10): 3-7, 19.
[19] Yu Liping, Wang Bing, Zhang Zaijie. Research on the application of diachronic diffusion factors and diachronic relative diffusion factors—taking CSSCI library, information, and documentation science journals as an example[J]. Journal of Intelligence, 2020, 39(3): 156-162.
[20] Shi Ronghua, Liu Xiwen. Research on library scientific data services based on the data lifecycle[J]. Library and Information Service, 2011, 55(1): 39-42.
[21] Ding Ning, Ma Haoqin. Comparative study of foreign university scientific data lifecycle management models[J]. Library and Information Service, 2013, 57(6): 18-22.

[22] Wu Tong. Research on scientific data open sharing services in US research libraries based on the data lifecycle[J]. Library and Information, 2019(1): 135-144.
[23] CEOS. Data lifecycle models and concepts[EB/OL]. [2020-04-21]. http://www2.lib.virginia.edu/brown/data/.
[24] Starr J, Willett P, Federer P, et al. A collaborative framework for data management services: the experience of the University of California[EB/OL]. [2020-05-17]. https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1014&context=jeslib.
[25] Pouchard L. Revisiting the data lifecycle with big data curation[J]. International journal of digital curation, 2016, 10(2): 176-192.
[26] DCC. Curation lifecycle model[EB/OL]. [2020-07-12]. http://www.dcc.ac.uk/resources/curation-lifecycle-mo.
[27] UKDA. Research data lifecycle[EB/OL]. [2020-07-12]. http://www.data-archive.ac.uk/create-manage/life-cycle.
[28] Liu Minjuan, Zhang Xuefu, Yan Yun. Research on discipline theme evolution methods based on core words, burst words, and new words[J]. Journal of Intelligence, 2016, 35(12): 175-180.
[29] Xiao Ming, Chen Jiayong, Li Guojun. Visual analysis of scientific knowledge mapping based on CiteSpace[J]. Library and Information Service, 2011, 55(6): 91-95.
[30] Hou Jianhua, Hu Zhigang. Review and prospect of CiteSpace software application research[J]. Modern Information, 2013, 33(4): 99-103.
[31] Chen Yue, Chen Chaomei, Liu Zeyuan, et al. The methodological functions of CiteSpace mapping knowledge domains[J]. Studies in Science of Science, 2015, 33(2): 242-253.
[32] Wang Faming, Zhu Meijuan. Bibliometric analysis of domestic blockchain research hotspots[J]. Journal of Intelligence, 2017, 36(12): 69-74, 28.
[33] Chen Yue, Chen Chaomei, Hu Zhigang, et al. Principles and applications of citation space analysis: a practical guide to CiteSpace[M]. Beijing: Science Press, 2014.
[34] Hox J J, Boeije H R. Data collection, primary vs. secondary[J]. Encyclopedia of social measurement, 2005, 1: 593-599.
[35] Li Huajie, Shi Dan, Ma Limei. Economic research based on big data methods: frontier progress and research review[J]. Economist, 2018(6): 96-104.
[36] Zhang Chengzhi, Zhang Yingyi. Research on automatic identification of research method entities based on full-text academic papers[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(6): 589-600.
[37] Wang Fang, Wang Xiangnü. A quantitative analysis of research methods in Chinese information science: taking *Journal of the China Society for Scientific and Technical Information* (1999-2008) as an example[J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(4): 652-658.

Author Contributions:

Chen Xin: Provided guidance and important suggestions on topic selection, framework, writing, and revision.

Cao Chaojin: Responsible for rule design, programming, initial drafting, and

paper revision.
Ye Chunsen: Provided directional revision suggestions on content.
Wang Chuanlei: Provided suggestions on paper structure modification.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.