Postprint: A Study on Scientific Data Citation Characteristics of Journal Papers in Library and Information Science in China

Authors: Ding Wenyao, Li Jian, Han Yi

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Exploring the characteristics and patterns of scientific data citation in journal papers not only helps to describe the utilization of scientific data within disciplinary domains, but also reveals data citation patterns in academic output.

[Method/Process] This study takes papers published in the first issues of 2017 and 2018 from six journals in China' s library and information science field as samples. Using citation elements from the national standard "Information Technology - Scientific Data Citation" and employing content analysis, the scientific data citation behavior of sample papers was coded across nine dimensions. Statistical methods were then applied to describe the characteristics of scientific data citation in library and information science journal papers and to explore relationships among features across different dimensions.

[Results/Conclusion] Journal papers in the library and information science field extensively cite statistically compiled scientific data from both domestic and international sources, with a relatively high volume of citations to personal research scientific data in journal papers. There exists a certain correspondence between scientific data citation annotation methods and scientific data types, but the diverse annotation methods lack uniformity. The phenomenon of secondary citation is relatively prominent, and the degree of secondary citation is correlated with the type of scientific data creator.

Full Text

Research on the Characteristics of Scientific Data Citation in Journal Articles in Library and Information Science in China

Ding Wenyao, Li Jian, Han Yi College of Computer and Information Science, Southwest University, Chongqing 400715

Abstract

[**Purpose/Significance**] Exploring the characteristics and patterns of scientific data citation in journal articles not only helps describe the utilization of scientific data within a discipline but also reveals data citation models in academic output. [**Method/Process**] Taking articles published in the first issues of six library and information science journals in China during 2017 and 2018 as samples, and combining citation elements from the national standard *Information Technology—Scientific Data Citation*, this study employed content analysis to code scientific data citation behaviors across nine dimensions. Statistical methods were then applied to describe citation characteristics and explore relationships between different dimensional features. [**Results/Conclusions**] Journal articles in library and information science widely cite organized statistical scientific data from both domestic and international sources, with substantial citation of personal research data from journal articles. There exists a certain correspondence between citation styles and scientific data types, but the diverse annotation methods lack uniformity. Secondary citation is prominent, and its degree correlates with the creator type of scientific data.

With the rise of data-intensive scientific research paradigms, scientific data has become an increasingly important force driving scientific research. Not only does natural science research rely on in-depth data analysis and mining, but humanities and social sciences research also shows a clear trend toward datafication. The generation and citation of scientific data in practical research have become common phenomena in data-era scientific activities.

Many studies consider scientific data as important as academic papers [1]. Scientific data citation plays a vital role in identifying data ownership, promoting data sharing, measuring data impact, and advancing data metrics research. Consequently, proper and standardized data citation practices are necessary. Meanwhile, scientific data citation standards have received continuous attention from relevant institutions and organizations both domestically and internationally. The 2018 release of the State Council' s *Measures for Scientific Data Management* and the national standard *Information Technology—Scientific Data Citation* (GB/T 35294-2017) [2] fully demonstrate China' s commitment to standardizing scientific data citation.

Scientific data citation represents a significant and universal information behavior [3]. Analyzing this behavior not only helps describe data utilization within

disciplines but also reveals the characteristics and patterns of researchers' citation practices, thereby providing a reliable basis for improving citation standards. Current research on scientific data citation behavior has accumulated certain achievements, primarily employing content analysis to explore citation characteristics across disciplines. Research dimensions include citation frequency, data types, metadata elements, citation formats, citation locations, and citation standardization. However, most studies either focus on citation standardization or describe the distribution characteristics of independent dimensions, with few examining both data content and citation methods simultaneously. Moreover, no scholars have yet explored correlations between citation characteristics across different dimensions.

This raises important questions: Do relationships exist between scientific data types and citation methods? If so, what citation characteristics might these relationships reveal? Based on these questions, this study examines scientific data citation in Chinese library and information science journals, focusing on describing citation characteristics and exploring correlations between different dimensional features to reveal citation patterns from both data type and behavioral perspectives.

2. Data and Methods

2.1 Conceptual Definitions (1) Scientific Data. While definitions vary across documents such as the *OECD Principles and Guidelines for Access to Research Data from Public Funding* [12], *Scientific Data Sharing Engineering Technology Standards Research Report* [13], and *Information Technology—Scientific Data Citation* (GB/T 35294-2017) [2], they share common elements. Scientific data should have three characteristics: First, in terms of origin, it derives from objective records of real-world phenomena and constitutes raw materials generated during scientific research activities. Second, in content and form, it comprises objective factual records reflecting human socio-technical activities or the natural world, including both qualitative descriptions and quantitative data, serving as original source material. Third, in functionality, it serves as scientific evidence while being reusable and valuable for reasoning, discussion, or computation in other research.

Based on these definitions, this study defines scientific data as: Evidence-based original recorded information or processed data materials that objectively reflect the nature, characteristics, and change patterns of human socio-technical activities or the natural world, as recognized by the scientific community.

Scientific data manifests in various forms, broadly categorized as non-textual descriptive and textual descriptive types. This study focuses on non-textual descriptive scientific data, specifically including: quantitative data, datasets, models, charts, formulas, algorithms, images, and audio/video materials.

(2) Scientific Data Citation. Currently, there is no unified conceptual description. Major definitions include: researchers' need to cite used resources in

bibliographies or footnotes, with scientific data citation referring to the description and source identification of data resources through specific identification technologies and mechanisms [3, 14]; and researchers' practice of citing scientific data to support paper arguments, providing data references through references, footnotes, or in-text citations [15].

These definitions share similar conceptual connotations, emphasizing the annotation of cited data content and sources. Therefore, this study defines scientific data citation as: The information behavior of researchers annotating the sources of scientific data cited in papers through references, footnotes, or in-text citations.

2.2 Data Sources Data were obtained from the CNKI database, with the basic requirement of covering as many journals as possible while ensuring sample quality. Based on CNKI' s journal impact factor data for the library and information science field, the top six journals by composite and comprehensive impact factors were selected (as shown in): *Journal of Library Science in China*, *Journal of Academic Libraries*, *Library and Information Service*, *Library and Information Science Knowledge*, *Journal of National Library of China*, and *Library and Information Work*. These six journals represent the research level of the field. Using systematic sampling, articles from the first issues of 2017 and 2018 were selected. After removing editorials, conference reports, briefings, and retrospective essays, 175 sample papers were obtained.

Since no standardized format or reliable automatic identification method currently exists for data citation, this study employed manual identification. The coding process involved: analyzing each sentence, phrase, presentation form, and relevant context involving scientific data in each paper, focusing on chart data and quantitative data in text; identifying data citation records while distinguishing them from literature citations, factual statements, data provision, and numerical descriptions (such as dates and natural numbers); and coding each record according to the content analysis category system.

2.3 Research Methods **(1) Content Analysis.** This method was used to code the content and presentation forms of scientific data citations in sample papers, with descriptive statistical analysis applied to coding results. The category system was developed based on the national standard *Information Technology—Scientific Data Citation* (GB/T 35294-2017) and research objectives, constructing a content analysis framework across nine dimensions (as shown in): journal name (X1), journal issue (X2), paper number (X3), scientific data type (X4), data source region (X5), data creator type (X6), data disseminator type (X7), data presentation type (X8), and citation style (X9). Each citation record was coded across multiple dimensions, with results stored in Excel files. After cleaning and organizing, frequency statistics and cross-tabulation analysis were used to examine citation characteristics across dimensions.

(2) Correspondence Analysis. Also known as association analysis, this mul-

tivariate statistical technique reveals relationships between variables by analyzing contingency tables composed of qualitative variables. It can visually present relationships between seemingly unrelated data sets through acceptable positioning maps, revealing differences within variable categories and correspondences between categories of different variables [16].

Based on the quantified coding data, IBM SPSS Statistics correspondence analysis module was used to analyze relationships between citation characteristic variables, exploring citation pattern features through correspondences between different variables and their elements.

3. Results

3.1 Scientific Data Citation Characteristics Through manual identification, preliminary statistical coding, and two rounds of data verification, 632 scientific data citation records were compiled from 175 sample papers. Descriptive statistics were applied to characterize citation patterns from overall distribution and six specific dimensions.

3.1.1 Overall Distribution of Scientific Data Citations. As shown in and , among 89 papers in the 2017 first issues, 70 contained scientific data citations with an average of 4.64 citations per paper. Among 86 papers in the 2018 first issues, 65 contained citations with an average of 4.72 per paper. These figures indicate that scientific data citation is relatively common in library and information science journals, with stable citation volumes in recent years.

3.1.2 Distribution of Scientific Data Types. As shown in [Figure 1: see original paper], researchers in library and information science cite organized statistical data most frequently, accounting for approximately 78% of total citations, followed by research results data (17%) and original record data (4%). Organized statistical data primarily comprises survey statistics from various industries and purposes—secondary data processed and organized by humans. Research results data refers to non-statistical data generated in scientific activities, including formulas, algorithms, models, scales, and empirical data. Original record data refers to recorded data formed without human statistical processing, primarily involving original data generated by internet platforms (web clicks, dates) in sample papers.

3.1.3 Distribution of Scientific Data Source Regions. As shown in [Figure 2: see original paper], Chinese library and information science researchers cite domestic (47.47%) and international (41.14%) scientific data in roughly equal proportions, indicating that research activities widely require data from different countries, with both showing equal importance. Additionally, 11.39% of data sources could not be identified, primarily because papers lacked source annotations and contextual information was insufficient to determine origin.

3.1.4 Distribution of Scientific Data Creators. As shown in [Figure 3: see original paper], personal research data is cited most frequently (28.96%),

followed by social group data (11.55%)—data formed by non-profit academic, industry, professional, and joint associations. Commercial enterprise data ranks third (7.44%), followed by data from school libraries (7.12%), databases (6.65%), government agencies (6.17%), public libraries (5.22%), schools (3.96%), research institutions (0.63%), and archives (0.47%). Approximately 16.46% of citation records could not identify creators, indicating that researchers' awareness of citing data creators needs improvement, with non-citation being relatively prominent.

3.1.5 Distribution of Scientific Data Disseminators. The data disseminator refers to the citation source, distinct from the data creator. While the creator indicates where data originated, the disseminator indicates where it was cited from. As shown in [Figure 4: see original paper], citation channels are diverse in library and information science, with journal papers being the most popular channel (25.32%), followed by internet platforms provided by social groups (9.65%), enterprises (8.86%), and educational institutions (8.23%). Notably, many cases could not identify dissemination channels. Combined with creator distribution results, this shows that researchers' awareness of annotating both creation and dissemination sources is somewhat lacking, with non-citation of dissemination sources being more prominent.

3.1.6 Distribution of Scientific Data Presentation Forms. Presentation type refers to how citations appear in papers, such as textual discussion, charts, formulas, and models. As shown in [Figure 5: see original paper], over 80% of citations appear in textual form, indicating this is the preferred method. Creating tables based on cited data (11.55%) and directly citing figures (6.49%) are also relatively common.

3.1.7 Distribution of Scientific Data Citation Styles. As shown in [Figure 6: see original paper], citation styles are diverse. Annotating references is the most common method (60%), while other methods—including chart name references, in-chart references, URL annotations, chart footnotes, page footnotes, in-text source descriptions, and acknowledgments—account for smaller proportions. Notably, non-citation is prominent at approximately 22%.

3.2 Correlations Between Scientific Data Citation Characteristics Descriptive statistics reveal distribution characteristics across dimensions from both data attribute and citation behavior perspectives. But do relationships exist between these features? Do different journals show preferences for different data sources? Are there associations between data types and presentation methods? What relationships exist between creator types and dissemination channels?

Based on these questions, five hypotheses were proposed and tested using correspondence analysis. SPSS analysis results show all five hypotheses have P-values below 0.05, indicating high correlations between variables (as shown in). Correspondence analysis reduces dimensions to plot two-dimensional scatter

diagrams showing relationships between variable categories—closer distances indicate stronger relationships. The following sections explore these relationships.

3.2.1 Correlation Between Journal Type and Data Creator Type. [Figure 7: see original paper] shows the correspondence between six journal categories and 13 creator categories. Journal and creator variable points are distributed across different quadrants with varying distances, indicating certain correspondences and citation preferences. Specifically, *Library and Information Service* shares the first quadrant with government agencies, social groups, and research institutions, indicating it primarily cites these data types, with closer proximity to government and social group data suggesting more frequent citation of these sources. *Library and Information Work* shows dense citation of personal research data and some from databases and archives. *Journal of Academic Libraries* mainly cites data from public libraries and schools. *Journal of National Library of China* primarily cites data from school and public libraries. *Journal of Library Science in China* and *Library and Information Science Knowledge* show similar patterns, mainly citing enterprise and other public service data. These preferences likely relate to journal nature, sponsors, manuscript topics, and article content.

3.2.2 Correlation Between Data Type and Presentation Form. [Figure 8: see original paper] shows that data types correspond to different presentation forms. In the first quadrant, organized statistical data shows dense variable points, indicating multiple citation methods, primarily textual discussion, followed by table/chart creation or direct table citation. The second quadrant shows original record data closely associated with direct figure citation. The third quadrant shows research results data corresponding mainly to formulas and models. Uncertain data types, resulting from non-standard or non-citation, show weak correlations with any variables.

3.2.3 Correlation Between Data Type and Citation Style. [Figure 9: see original paper] reveals that different data types occupy different quadrants, showing distinct citation style correspondences. Research results data in the first quadrant corresponds closely to: chart footnotes with references, chart name references, in-text source descriptions, and paragraph-level source descriptions with references. Organized statistical data in the third quadrant shows dense citation styles, with correspondence strength in descending order: reference list annotation, non-citation, URL annotation in parentheses, chart footnotes, and acknowledgments. Original record data in the fourth quadrant primarily corresponds to page footnotes, followed by in-chart references and page footnote references.

3.2.4 Correlation Between Presentation Form and Citation Style. [Figure 10: see original paper] shows eight presentation forms correspond to 11 citation styles. Creating tables based on data corresponds to chart footnotes, in-chart references, and acknowledgments. Textual discussion, models, and formulas overlap at the boundary of quadrants 2 and 3, showing minimal differences and primarily using reference list annotation and URL annotation. Direct

figure citation, table citation, and chart creation share similar citation styles in quadrant 4, with direct figure citation favoring chart footnotes, table citation favoring page footnotes with references, and chart creation favoring chart footnotes with references.

3.2.5 Correlation Between Data Creator Type and Disseminator Type. Scientific data citation emphasizes annotating source information. According to the national standard, source elements include both creators and disseminators, positioning data origin from both original creation and dissemination channels. From the perspective of citation consistency, creator and disseminator types should correspond.

Hypothesis 5 tested this consistency through correspondence analysis, with results shown in [Figure 11: see original paper]. Overall, 12 creator types and 10 disseminator types show relatively consistent correspondence. For instance, "government department" and "government agency" are close in quadrant 1, indicating government-created data is mainly cited from government platforms. "Personal research" and "journal articles" overlap in quadrant 2, showing personal research data is primarily cited from journal articles. However, varying distances between variable categories indicate differences in consistency across situations.

Cross-tabulation analysis further explored consistency (as shown in). Among 632 citation records, 423 showed creator-disseminator consistency, 41 showed inconsistency, and 168 were uncertain. The overall consistency rate is 66.93%, indicating most citations come directly from creators. However, 26.58% of cases could not determine consistency, mainly due to non-standard or non-citation. Comparing consistency rates across creator types reveals significant variation: personal research and school library data show high consistency (93.99% and 93.3%, respectively), while enterprise, archive, and school data range 60-70%. Database and research institution data show lower consistency (54.76% and 25%, respectively), indicating researchers' citation awareness for these sources needs improvement and "secondary citation" is common.

4. Discussion and Conclusion

This study examined 175 papers from six library and information science journals to reveal citation characteristics and inter-feature relationships, yielding the following conclusions:

(1) Scientific data citation is relatively common in library and information science journals. Papers with data citations account for 77.14%, with stable citation volumes between 2017 and 2018 (4.64 vs. 4.72 citations per paper). Previous research showed that from 2003-2013, only 49.89% of papers in *Journal of Library Science in China*, *Journal of Intelligence*, and *Journal of Academic Libraries* contained data citations, with per-paper citations between 1-2 times during 2003-2008, gradually increasing to 2.7 by 2013 [4]. Compared with these findings, both the proportion of papers citing data and per-paper

citation counts have significantly increased, indicating data-based empirical research is gaining prominence and reflecting the data-intensive research paradigm trend. In comparison, sociology journals from 2003-2014 averaged over 6 citations per paper [5], suggesting higher data dependency in sociology and room for improvement in library and information science.

(2) Researchers in library and information science most frequently cite organized statistical data, with less citation of research results and original record data. This aligns with previous social science research showing survey data from ICPSR [1] and census/yearbook data in Chinese sociology journals [5] as most common. Association analysis reveals two key findings: First, organized statistical data is primarily cited through textual discussion, with some table/chart creation, while original record and research results data are mainly cited through direct figure, formula, and model citation. Second, citation styles are diverse and lack uniformity—researchers use different annotation methods not only across data types but even within similar types. Overall, traditional literature citation methods serve as the general approach for data citation, with reference list annotation being most common for organized statistical data. However, citation styles and positions for figures, tables, and formulas are chaotic.

(3) Scientific data citation sources are broad in library and information science. Regionally, domestic and international data are cited equally. In terms of creators, data from personal research, social groups, enterprises, and government agencies are widely cited, with personal research data being most common. Association analysis shows different creator types correspond to different dissemination channels. The most-cited personal research data primarily comes from journal articles, while institutional data mainly comes from official internet platforms. Since scientific data is formed in specialized research activities and has attribution to creators, citing from original sources should be standard practice. However, analysis reveals varying degrees of inconsistency—"secondary citation." Comparison shows secondary citation is less common for personal research and school library data but more prevalent for public library and social group data. Since open sharing of data from creators is a key factor affecting citation [1], secondary citation likely relates to the openness of data dissemination channels. Personal research data is typically included in published papers or open platforms (e.g., CNKI, Wanfang) with high accessibility, reducing secondary citation. Social groups, public service departments, and government agencies also publicly release data, with internet platforms providing new dissemination channels. However, archives and research institutions have lower openness, weaker accessibility, and thus more secondary citation.

(4) Non-citation behavior is prominent among researchers in library and information science. Statistics across dimensions show that 22%-26% of cases could not be identified through citation annotations or context, indicating researchers lack awareness comparable to literature citation and that no mature citation standards guide behavior.

These conclusions offer insights for scientific data citation management:

(1) Citation distribution reflects domain data structure and needs, providing references for planning data repositories and developing citation channels. First, personal research data accompanying published papers should be uniformly managed with shared channels and standardized citation formats. Additionally, scientific data released by important institutions and social groups should be widely collected to establish unified access platforms or citation links, improving data visibility and open sharing to reduce secondary citation.

(2) Unified scientific data citation standards are needed. This study finds that citation standards are not effectively implemented in library and information science. Most authors still follow traditional literature citation methods, with diverse but non-uniform alternative annotation methods. This may result from insufficient awareness, incomplete standards, or lack of researcher consciousness. Traditional literature citation may suit current habits but only identifies data sources without indicating content, creators, or creation times. If combined with traditional citation, reference elements should be supplemented according to data type differences. Additionally, citation styles and positions for charts are chaotic and require clear guidelines for different data types.

This study has limitations: the sample of 175 papers from six journals in CNKI' s "library and information science" category is limited, and conclusions' generalizability requires verification with more domains and data. The study only analyzed citation records to reveal behavioral characteristics and patterns without exploring deep causes or influencing factors. While manual coding can identify phenomena, the process is inefficient and cumbersome, necessitating development of automatic citation identification tools.

References

[1] Qu Yajie, Wang Yanan. Analysis of the Current Situation and Characteristics of Scientific Data Citation in Social Sciences [J]. Digital Library Forum, 2017(6): 25-31.

[2] General Administration of Quality Supervision, Inspection and Quarantine of the People' s Republic of China, Standardization Administration of the People' s Republic of China. National Standard of the People' s Republic of China GB/T 35294-2017 *Information Technology—Scientific Data Citation* [EB/OL]. [2018-08-20]. http://www.gb688.cn/bzgk/gb/newGbInfo?hcno=A495CA355BAF00D962AA8DD84C3B2C16.

[3] Wang Xue, Ma Shengli, She Zengli, et al. Research on Scientific Data Citation Behavior and Its Impact [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(11): 1132-1139.

[4] Ding Nan, Ding Ying, Yang Liu, et al. Analysis of Data Citation Behavior in Library and Information Science Journals in China [J]. Journal of Library Science in China, 2014(6): 105-114.

[5] Ding Nan, Yang Liu, Ding Ying, et al. Research on Data Citation Behavior in Chinese Sociology Journals [J]. Library and Information Service, 2014(6): 88-93.

[6] WILLIAMS S C. Data practices in the crop sciences: a review of selected faculty publications [J]. Journal of agricultural & food Information, 2012, 13(4): 308-325.

[7] PARK H, YOUS, WOLFRAM D. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields [J]. Journal of the Association for Information Science and Technology, 2018, 69(11): 1346-1354.

[8] Zhao Ruihan. Research on the Reuse Status of Scientific Data Papers—Citation Analysis Based on Data Journal "Earth System Science Data" [J]. Information Studies: Theory & Application, 2017, 40(11): 52-57, 72.

[9] Meng Xiangbao, Qian Peng. Research on Characteristics of Humanities and Social Sciences Data from the Perspective of Data Lifecycle [J]. Library and Information Science Knowledge, 2017(1): 76-88.

[10] CHEN R S, DOWNS R R, SCHUMACHER J A. Analyzing data citations to assess the scientific and societal value of scientific data [EB/OL]. [2018-09-26]. http://academiccommons.Columbia.edu/catalog/ac:157135.

[11] HENDERSON T, KOTZ D. Data citation practices in the CRAWDAD wireless network data archive [C]//Chengdu: IEEE international conference on communications software & networks. 2015.

[12] PILAT D, FUKASAKU Y. OECD principles and guidelines for access to research data from public funding [J]. Data science journal, 2007, 6: 4-11.

[13] Li Huijia, Ma Jianling, Wang Nan, et al. Research Progress on Organization and Management of Scientific Data at Home and Abroad [J]. Library and Information Service, 2013, 57(23): 130-136.

[14] Li Dandan, Wu Zhenxin. Research on Research Data Citation [J]. Library Journal, 2013, 32(5): 65-71.

[15] Huang Guobin, Liu Xinran, Jiang Ying. Analysis of External Factors Affecting Scientific Data Citation [J]. Digital Library Forum, 2017(6): 2-8.

[16] Du Qiang, Jia Liyan. *SPSS Statistical Analysis from Entry to Mastery* [M]. Beijing: People' s Posts and Telecommunications Press, 2009: 419-422.

Author Contributions

Ding Wenyao: Data collection, analysis, and drafting; Li Jian: Participated in paper revision; Han Yi: Overall research design and planning, paper revision and improvement.

Research on the Characteristics of Scientific Data Citation in Journal Articles in Library and Information Science in China

Ding Wenyao, Li Jian, Han Yi
College of Computer and Information Science, Southwest University, Chongqing 400715

Abstract: [**Purpose/significance**] Exploring the characteristics and laws of scientific data citation in journal articles not only helps describe data utilization within disciplines but also reveals data citation models in academic outputs. [**Method/process**] Articles from the first issues of six library and information science journals in China during 2017 and 2018 were selected as samples. Based on citation elements from China' s national standard *Information Technology—Scientific Data Citation*, content analysis was used to code scientific data citation behaviors across nine dimensions. Statistical methods were applied to describe citation characteristics and explore relationships between different dimensional features. [**Result/conclusion**] Results show that: scientific data citation is relatively common in library and information science; organized statistical data from domestic and international sources is widely cited, with substantial citation of personal research data from journal articles; certain correspondences exist between citation styles and data types, but diverse styles lack uniformity; secondary citation is prominent and its degree correlates with scientific data creator types.

Keywords: library and information science; journal articles; scientific data citation; citation characteristics

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