DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.2, MAY. 2022

Studies on the characteristics of scientific data citation in Chinese researchers: Case studies of twelve academic journals

Wenyao Ding[a], Yi Han[b*]

a. Guizhou Minzu University Library, Guiyang, China
b. Business College, Southwest University, Chongqing, China

ABSTRACT

Scientific data citation is a common behavior in the process of scientific research and writing academic papers under the context of data-intensive scientific research paradigm. Standardized citation of scientific data has received continuous attention from academia and policy management departments in recent years. In order to explore the characteristics and the correlation of scientific data citations of Chinese researchers, based on the results of scientific data citations presented in academic papers, this study use CNKI as the data source to extract 771 papers in 12 academic journals during 2017 to 2019. Combining with the Chinese national standard *Information Technology-Scientific Data Citation (GB /T 35294-2017)*, a set of variables were given to reflect the reference characteristics. First, 4992 citation records of scientific data were manually identified and coded one by one, and the citation characteristics were presented with the statistical distribution of data frequency. Then, the chi-square test, log-linear model analysis, and correspondence analysis methods were used to analyze and explore the significant correlation among the characteristic variables. The study found that in general, the phenomenon of scientific data citations in Chinese researchers is widespread, and the number of citations has increased year by year, but there are also a large number of irregular citations. At present, there are roughly two modes of citation labeling behavior, and the traditional document citation mode is still the mainstream citation method for data citation. Furthermore, distributor type of scientific data may affect the reference in marked way. In addition, the completeness of the labeling elements differed in different bibliographic elements of scientific data. Irregular references to Unique Identifiers and parsing addresses are particularly prominent, which may be related to the type of distributor.

KEYWORDS

China's mainland; Journal articles; Scientific data citation; Citation characteristic; Characteristic relevance

1 Introduction

With the development of data-intensive scientific research paradigm, scientific data has increasingly become an important force driving scientific research (Chao et al, 2021). At pre-

* Corresponding Author: hanyi72@swu.edu.cn

sent, scientific research in various disciplines has gradually relied on in-depth analysis of scientific data. These practices not only promote the output of scientific data, but also accelerate a large number of data citations. Researchers reuse scientific data to analyze the feasibility and reliability in the past experiments, which can greatly reduce the repeating data collection and improve the efficiency of scientific research (Zhang, 2013).

Scientific data citation, as the main type of scientific data sharing, refers to the practice which the authors used references, footnotes or in-text annotations to mark the source of the scientific data used in the paper (Ding et al, 2014b). Numerous studies believe that scientific data and academic papers are equally important (Qu & Wang, 2017), and it is necessary to standardize the presentation and citation of scientific data, which can improve the visibility of scientific data, promote scientific data sharing, and effectively play the value of scientific data in scientific research. With the massive reuse of scientific data and the development of data-level measurement theory (Gu, 2015), the normative standards for scientific data citation have gradually become a necessary research point. In order to standardize the management and citation of scientific data in China, in 2018, the State Council of China issued the national *standard Information Technology-Scientific Data Citation (GB/T 35294-2017)*, which especially focused on elements and formats for scientific data citation.

In terms of academic research, scientific data citation in empirical researches is gradually enriched. Related researches usually take specific disciplines as samples, which explores the relationship between scientific research and scientific data based on citation marks and fields in academic papers, and reveals the characteristics of the citation behavior of researchers. The dimensions of research concern are shown in the Table 1.

Table 1 Scientific Data Citation Characteristics Research Dimensions

Author	Research object	Research dimension					
		Citation frequency	Scientific data characteristics	Citation method	Citation location	Citation integrity	Citation presentation form
(Mo, 2004)	Economic Papers of *Xinhua Digest* from 2001 to 2003					√	
(Ding, et al, 2014c)	Papers of *Sociological Research* and *Chinese Population Science* from 2003 to 2014	√				√	
(Ding et al, 2014a)	Papers of *Journal of the Library Science in China, Journal of Information, Journal of Academic Libraries* from 2003 to 2013	√				√	
(Shi & Si, 2019)	Papers of *Journal of Ecology, Biodiversity, Journal of Geophysics, Acta Petrologica, Economic Research, Chinese Population Science journal papers* from 2013 to 2017		Type, Granularity level	√		√	
(Liu et al, 2019)	Funded project papers of 20 journals in the Natural Sciences, Humanities and Social Sciences in China from 2015 to 2016		Creator, Title, Time, Distributor, Acquisition address,		√	√	
(Ding et al., 2019)	Papers of 6 journals in the Library and Information in China from 2017 to 2018	√	Type, Country, Creator, Distributor	√		√	√

In summary, scientific data citation has received continuous attention from all cycles of life. Previous studies have shown that, in order to promote the development of scientific data citation standards, it is necessary to pay more attention to the actual citation needs and citation behavior characteristics. However, the current empirical research on the characteristics of scientific data citation behavior in China is relatively fragmented, and there are still expandable aspects in terms of research dimensions and research samples. Therefore, this study combines the national standard of *Information Technology-Scientific Data Citation* to expand and refine the research dimensions of citation characteristics, and expand the scope of sample disciplines, aiming to reveal the scientific data citation characteristics in Chinese researchers in depth.

Scientific data citation takes text as the expression carrier, and the citation content in the paper directly presents the objective result of the citation behavior. Based on the objective distribution of data citations in twelve Chinese academic journal articles, this research focuses on two aspects:

(1) Citation characteristics of scientific data.

(2) The correlations between the citation characteristics.

2 Data and Methods

2.1 Data Acquisition

Sampling strategies. Taking CNKI as the data source, the sample selection is based on the Discipline Navigation. In the horizontal dimension, the principles of stratified sampling and discriminative sampling are adopted to select a journal that is representative of the subject from each of the 10 subject albums. In the time dimension, the sampling time window is set to 3 years (2017~2019), and all full texts of the first issue of the sample journals from 2017 to 2019 are sampled as content analysis samples.

After pre-investigation and supplementary sampling, 12 journals were selected as samples, and the full text of the first issue from 2017 to 2019 was downloaded. Excluding non-academic papers such as editor's message, conference report, newsletter, etc., 771 valid papers were obtained in total, and the distribution is shown in Table 2.

Table 2 Sample distribution

Discipline	Journal Title	Number of papers sampled in 2017	Number of papers sampled in 2018	Number of papers sampled in 2019	Total
Natural Science and Engineering Technology	*Acta Ecologica Sinica*	35	36	38	109
	Scientia Agricultura Sinica	48	49	49	146
	Journal of Software	22	25	16	63
	Food Science	17	18	16	51
	Chinese Journal of Vaccines and Immunization	28	26	26	80
	Chinese Journal of Rock Mechanics and Engineering	20	16	16	52

Discipline	Journal Title	Number of papers sampled in 2017	Number of papers sampled in 2018	Number of papers sampled in 2019	Total
Social Science	*Economic Research Journal*	15	15	14	44
	Sociological Studies	11	10	9	30
	CASS Journal of Political Science	12	14	10	36
Humanities	*China´s Borderland History and Geography Studies*	19	21	21	61
	Studies in World Religions	17	19	21	57
	Philosophical Research	17	13	12	42
Total		261	262	248	771

2.2 Research Methods

The citation behavior of scientific data usually takes the text as the expression carrier, and the citation content fragment in the paper directly records the objective result of the citation. First, the content analysis method is used to manually code the content of scientific data citation, and statistical data distribution by citation frequency. Secondly, in order to explore the possible correspondences between the citation characteristics of scientific data, EXCEL and IBM SPSS Statistics are used to further analyze the association between the citation characteristics of scientific data based on the distribution of coded data.

Data encoding. The categories for content analysis are the coded variable set, which is composed of 15 variables in three dimensions, as shown in Table 3, based on the scientific data citation description elements contained in the national standard *Information Technology-Scientific Data Citation (GB/T 35294-2017)*, as well as relevant empirical research experience.

3 variables of Sample Characteristics (A) reflect the objective data of sample attributes, and 12 variables of Citation Behavior Characteristic (B) and Scientific Data Characteristic (C) collectively reflect scientific data citation characteristics. Citation Behavior Characteristics (B) mainly reflect the actual performance of citation behavior in the text of the paper, including citation expression, citation mark, and citation labeling. Scientific Data Characteristic (C) include the attribute characteristics of the cited scientific data mined from the citation annotation and description related information.

This article is coded by manual identification, analysis, and classification, and the coding details are formed by the pre-coding and post-coding, and the quantization standard for classification is adopted.

Table 3 Set of coding variables

Dimensions	Number	Variable item	Instruction	Reference
Sample Characteristics (A)	A1	Journal Title	Sample journal name	
	A2	Discipline	Academic disciplines of the sample journals	
	A3	Years	Sample journal year	

Dimensions	Number	Variable item	Instruction	Reference
Citation Behavior Characteristic (B)	B1	Citation Label Mode	The way of indicating the source or related information	(Shi & Si, 2019; Ding et al., 2019)
	B2	Citation Presentation Form	The form in which scientific data is cited in the text	(Ding et al., 2019)
	B3	Citation Location	Chapter position of scientific data citation in the text	(Liu et al, 2019)
	B4	Citation Mark Position	The location of the citation cue mark	Incidental indicators based on Citation Label Mode
	B5	Citation Source Information Label Position	The location of the specific information of the scientific data citation source	
Scientific Data Characteristics (C)	C1	Production Time	The year the scientific data was created	
	C2	Producer	The creator or institution of scientific data	
	C3	Distributor	Dissemination agency or channel of scientific data	Information Technology –Scientific Data Citation (GB/T 35294–2017)
	C4	Unique Identifier	Unique Identifier for scientific data	
	C5	Resolution Address	Used to parse the Unique Identifier of scientific data and return the access address of the scientific data URL, XML or other formats	
	C6	Scientific Data Form Type	Form type of scientific data	(Shi & Si, 2019: Qu & Wang, 2017)
	C7	Scientific Data Content Type	Content type of scientific data	

Statistical Analysis. Based on 4992 variable data, chi-square test, logarithmic linear model, and multiple correspondence analysis methods are used to analyze the correlations among characteristic variables in-depth.

3 Scientific Data Citation Characteristics

After data sorting and statistics, there are 639 papers with scientific data citations and 132 papers without scientific data citations. A total of 4992 clear scientific data citation records are identified.

3.1 Frequency of Scientific Data Citation

The overall situation of scientific data citations of 12 journal articles is shown in Table 4. The average number of citations per scientific data is the ratio of the number of scientific data citation records to the number of sampled articles.

Table 4 Citation of scientific data from 2017 to 2019

Discipline	Journal Title	Sample size (pieces)	Number of cited papers	Number of uncited papers	Total citations	The average number of citations per
Natural Science and Engineering Technology	*Acta Ecologica Sinica*	109	108	1	1083	9.94
	Scientia Agricultura Sinica	51	48	3	414	8.12
	Journal of Software	52	50	2	415	7.98
	Food Science	146	143	3	1034	7.08
	Chinese Journal of Vaccines and Immunization	80	77	3	476	5.95
	Chinese Journal of Rock Mechanics and Engineering	63	60	3	338	5.37
	Total	501	486	15	3760	7.50
Social Science	*Economic Research Journal*	44	39	5	419	9.52
	Sociological Studies	30	21	9	129	4.30
	CASS Journal of Political Science	36	19	17	96	2.67
	Total	110	79	31	644	5.85
Humanities	*China´s Borderland History and Geography Studies*	61	48	13	395	6.48
	Studies in World Religions	57	24	33	191	3.35
	Philosophical Research	42	2	40	2	0.05
	Total	160	74	86	588	3.68
Summation		771	639	132	4992	6.47

In general, the average level of citations per article is about 6.47 articles per article. In 2017, 2018, and 2019, there were 5.77, 6.40, and 7.29 respectively. It can be shown that, in general, the extent to which Chinese researchers cite scientific data in academic papers has been increasing year by year. However, the changes in different disciplines are different. The average amount of scientific data in the field of natural sciences and engineering technology is about 7.5, the field of social sciences is 5.85, and the field of humanities is about 3.68. It is inferred that researchers in the fields of natural sciences and engineering technology have a high degree of citation of scientific data.

3.2 Characteristics of Scientific Data Citation Behavior

3.2.1 Citation Label Mode

Using references is the most common way for Chinese researchers to cite and annotate scientific data. Approximately 81% of scientific data are annotated with their source information in a formatted citation description which is similar to literature citations. About 10% directly explain the source of scientific data in natural language. About 2% indicated the

source by marking the URL (Uniform Resource Locator) of the scientific data network access address. In addition, there are 7% no citations.

3.2.2 Citation Presentation Form

The Citation Presentation Form of scientific data refers to the one in which scientific researchers express the content of the cited scientific data in the text, such as text descriptions, tables, formulas, images, biological gene sequences, computer language codes, etc. About 82% of scientific data citations use text descriptions (including mixed descriptions of numbers and text) to express the content of scientific data citations, which is the most common way for Chinese researchers to present scientific data citations. Other cases of citing scientific data in the form of tables, formulas, and images accounted for 7.7%, 5.3%, and 4.7% respectively.

3.2.3 Citation Mark Position

Scientific data citation marks are signs indicating scientific data citations. Common marks include numerical superscripts, such as square brackets serial number superscripts ([1], [2], [3]), circle serial number superscripts (①, ②, ③)), and parentheses. The citation mark exists with the citation of scientific data. The common positions of the mark include: in the text, in the chart (cell), in the name of the chart, below the chart, indication in the figure, and footer.

About 85% of scientific data citation marks are located in the main text, that is, the position of the scientific data cited in the text. There are relatively few other reference marks, about 3.5% are in the chart (cell), about 3% are in the chart name, and about 1% are below the chart.

3.2.4 Citation Source Information Label Position

The citation source information of scientific data generally includes the detailed content of the data title, creator, distributor, time, etc., which should be displayed in the paper.

About 69% of the detailed source information of scientific data is described in the reference list after the paper, about 13% is in the footer, and about 10% is in the text. The survey once again shows that Chinese researchers still tend to use literature citation methods, and the source information of data citations can be found from the reference list.

3.2.5 Citation Location

The scientific data citation location is the chapter in the paper where the scientific data is cited. Generally, academic papers are similar in structure, including Introduction, Summary, Data/Methods, Results, Discussion, and Conclusions in order, as well as areas such as Acknowledgments, Appendices, and Footer Notes. It shows that scientific data is cited most frequently in the "Data/Methods" section of the papers, which contains 47% of the citation records. What's more, 38% in the "Discussion", and 11% in the "Introduction".

3.3 Scientific Data Characteristics

3.3.1 Production Time

It is of great importance to identify the production year of the cited data in the paper. The scientific data cited in the paper has a large time span. Although the earliest scientific data was produced in 1567, most of the scientific data cited in the paper are produced after 2000, and about 46% of the scientific data was formed after 2010. At the same time, there are a large number of scientific data cited in the papers during the Tang and Song Dynasties and the Ming and Qing Dynasties (Ancient Chinese Dynasty). In addition, about 14% of citation records cannot determine the year of scientific data creation.

3.3.2 Producer

Scientific data creators can be individual researchers, teams or organizations that produce scientific data, and they also include some database platforms that integrate scientific data[1]. About 70% of scientific data is produced by single/team researchers. The other 16% come from various types of organizations, such as the National Bureau of Statistics, Ministry of Health, People's Government and other government administrative agencies (5%), integrated databases (3%), science and technology, culture, education and other public institutions (2.7%), academic, industry, civil joint social organizations (2%), commercial enterprises (2%), international organizations such as the World Health Organization, the World Bank, and The Nature Conservancy (1.3%). Furthermore, 14% of the citation records cannot identify the scientific data creation agency.

3.3.3 Distributor

Distributors are the dissemination and distribution institutions of scientific data and the channels through which researchers obtain scientific data. Traditional information resource carriers are the main citation channels for scientific data: about 63% of scientific data are cited through academic papers and reports, among which journals [J] are the main dissemination carriers, including collections of papers [C], dissertations [D] and scientific reports [R], about 12% cited monographs [M], and about 1.5% cited technical standards [S]. Internet public data resources are also an important part: about 4% of scientific data is quoted through online Internet web platforms, such as the official website of government departments (.gov), non-profit organization websites (.org), industrial and commercial financial enterprise websites (.com), Educational institution websites (.edu), etc., about 2% of scientific data are referenced through database platforms.

In addition, a small amount (about 1.7%) of scientific data is obtained through archives, libraries, herbariums and other actual resource preservation venues.

In total, 15% of citations cannot identify the distributor of scientific data, which is slightly higher than the case where the producer of scientific data cannot be identified.

3.3.4 Unique Identifier

According to the national standard for scientific data citation (GB/35294-2017), the current Unique Identifier for scientific data is the OID (Object Identifier) specified in GB/26231-2017. In addition, for STRI (Science & Technology Resource Identifier), DOI (Digital Object Identifier) and other digital content identifiers, there are corresponding identifier preparation rules in GB/26231 and ISO 26324.

The survey shows that 77% of scientific data references are not marked with any Unique Identifiers. 18% marked the DOI number of the source document of the scientific data, rather than a Unique Identifier for the scientific data. About 5% are marked with other types of identifiers related to scientific data sources, such as standard number, file number, work report number, image number, etc., as shown in Table 5. This shows that Chinese researchers do not have the habit of labeling Unique Identifiers for scientific data, and they still remain at the level of citing the source of scientific data.

3.3.5 Resolution Address

The purpose of parsing the Resolution Address is to analyze the Unique Identifier and return the access address of the scientific data, such as URL (Uniform Resource Locator), XML, etc. In this study, both the resolution of the address and the direct identification of the URL access address are identified.

1 Note: When the scientific data is quoted from the database thematic data set and the name of the other accurate creation or–ganization cannot be identified, the database name is used as the creation organization.

Table 5 Distribution of Unique Identifiers related to scientific data

Scientific data types	Identifier	Amount
Standard Specification	Standard number	168
Text– (file) Administrative records	File number	49
Text–Interview Record	Interview record number	12
Image–Manual drawing	Image number	5
Text–Cultural relics inscriptions/inscriptions	Cultural relic number	4
Text–Administrative Record	Administrative Record Number	4
Concept model/index system	Work report number	3
Numerical–Survey and research data (set)	Work report number	2
Product patent	Patent No.	1
Formula—Formulas and Algorithms	Work report number	1

It is shown that 96% of the citations did not identify any Resolution Addresses related to scientific data. In fact, only 5 citations marked the Resolution Addresses corresponding to the DOI numbers of the scientific data source documents. Approximately 3.9% of scientific data citations directly mark the access address of the World Wide Web URL (Uniform Resource Locator) of the source of scientific data. It can be inferred that the researchers ignored the labeling of the Resolution Address of the scientific data.

3.3.6 Scientific Data Types

The scientific data content of 4992 citation records was uniformly classified, and a total of 74 types of scientific data were counted.

Based on the different forms of expression, scientific data is roughly divided into 8 types: numeric, text, image, table, formula, data file, patent, and others. Based on the different content of scientific data, it can be roughly divided into four types: scientific research data, fact description data, survey statistical data, and standard specification data.

(1) Form Type

The distribution of different types of scientific data is shown in Figure 1. Numerical scientific data has the largest number of citations, accounting for about 50%, followed by text

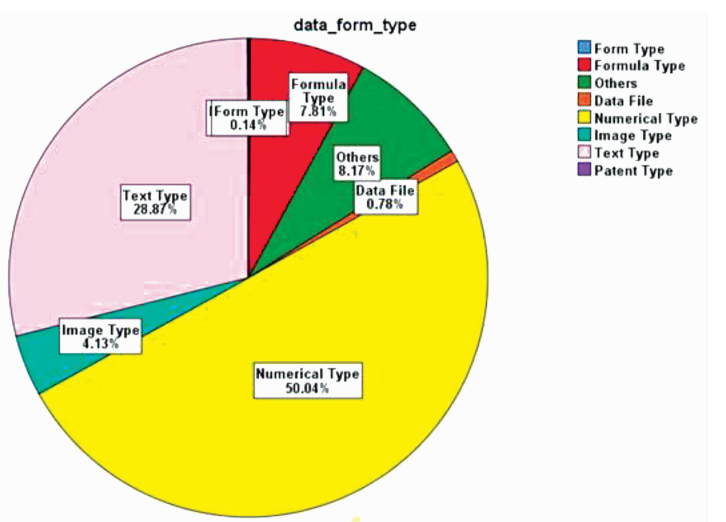

Figure 1 Distribution of scientific data form types

(about 29%), formula (about 7.8%), image (about 4%), and data files (about 1%), the number of citations to tabular and patented scientific data is relatively small. In addition, scientific data in the form of citing standard specifications, conceptual models/indicator systems, biological gene sequences, computer codes, etc. account for about 8%.

(2) Content Type

According to the differences in the content of scientific data, scientific data is roughly divided into 4 types.

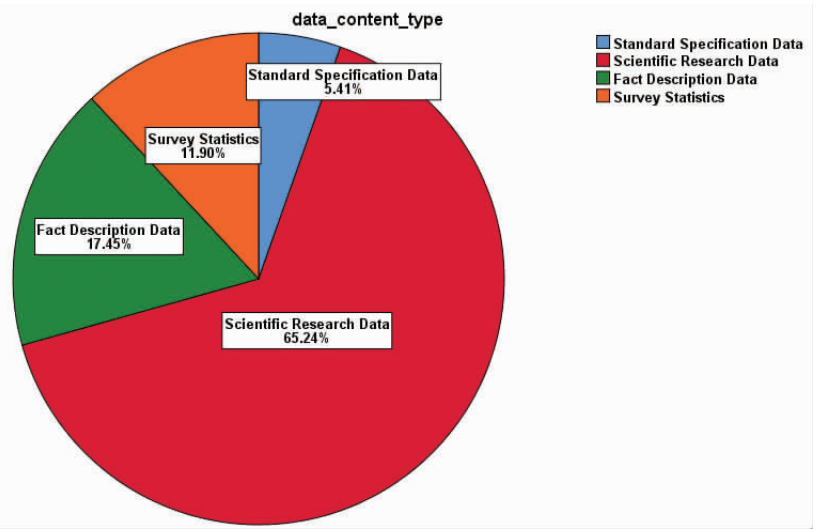

Figure 2 Distribution of scientific data content types

First, scientific research data: based on scientific research questions, process data and result data obtained by scientific research methods such as questionnaires, interviews, experiments, and inductive deductive analysis. Second, fact description data: objective measurement data and descriptive data reflecting natural phenomena, natural environment, biological behavior, and social events, such as natural observation data, administrative record data, etc. Third, survey statistical data: the original data and descriptive data collected through survey statistics. Most of the statistical data reflect the overall status quo, such as national or regional census statistics, targeted statistical survey data and directories of a certain industry, institution, or department. Fourth, standard specification data: the parameters and specification data contained in national and industry standards, the specification parameter data of software and patented products, etc.

The distribution of scientific data content types is shown in Figure 2. Among the scientific data cited by researchers, about 65% are scientific research data, 18% are fact description data, 12% are survey statistics, and 5% are standard specification data.

4 The Correlations among Scientific Data Citation Characteristics

The frequency distribution of data in each dimension reveals the citation characteristics of scientific data in academic papers by Chinese researchers. This study points out the question from the perspective of behavioral data regularity measurement: Are there any correlations among the data distributions of behavioral variables? In other words, is there any influence effect or correlation law between the cited characteristics?

Based on this question, the study uses statistical methods to focus on the correlation among scientific data citation characteristic variables:

1. Analyze the correlation between characteristic variables.
2. Construct correlation hypotheses based on variables with significant correlations.
3. Verify the hypothesis and explore the significant influence effects or correlation laws between the citation characteristics of scientific data.

4.1 Correlation Test and Hypothesis of Characteristic Variables

We quantitatively encode 4992 scientific data citation records, and call the chi-square test operation in SPSS; Choose the chi-square statistic and Monte Carlo method, the confidence level (Confidence level) and the number of samples (Number of samples) are 95% and 5000 respectively; Carry out the correlation significance test on different variable combinations, and decide to measure the Cramer's V coefficient to quantify the degree of correlation between the variables.

The value range of Cramer's V coefficient is between 0 and 1. The larger the value, the stronger the correlation between variables, that is, the greater the corresponding difference between different categories of variables. It is generally believed that there is a strong correlation between variables when 0.5≤ Cramer's V≤ 1.

This study uses Cramer's V≥ 0.6 as the threshold, focusing on variable combinations with significant correlations, as shown in Table 6.

Table 6 Chi-square test Cramer's V≥ 0.6 characteristic variable combinations

Dimension	Number	Variable combination	Cramer’s V value
Citation behavior characteristics (B)	1	Citation Source Information Label Position * Citation Label Mode	0.816
	2	Citation Source Information Label Position * Citation Mark Position	0.664
	3	Citation Mark Position * Citation Label Mode	0.603
Citation behavior characteristics (B) *scientific data characteristics (C)	1	Citation label mode * Distributor	0.724
	2	Resolve address * Distributor	0.707

As shown in Table 6, Citation Label Mode, Citation Mark Position, and Citation Source Information Label Position are the key elements of citation behavior characteristics (B), which are the direct manifestations of scientific data citation labeling behavior. The value of Cramer's V coefficient infers that the three variables may have a close correlation, and hypothesis 1 is proposed.

H1: There is a significant correlation between the Citation Label Mode, Citation Mark Position, and Citation Source Information Label Position.

Hypothesis 2 and Hypothesis 3 are proposed based on the combination of variables with strong correlation between the scientific data characteristics (C) and the citation behavior characteristics (B).

H2: The Citation Label Mode is significantly related to the Scientific Data Distributor.

H3: Distributors of scientific data are significantly related to address resolution.

4.2 Explanation of the Correlation of Characteristic Variables

Since hypothesis 1 needs to test the correlation of three variables, this study introduces log-linear model analysis and multiple correspondence analysis specifically for the study of the relationship between multi-categorical variables, aiming to explain the interaction effects between the three variables.

For Hypothesis 2 and Hypothesis 3, only the correspondence analysis method is used for testing, and the correlation between the variable characteristics is visualized through "dimensionality reduction" calculations and graphical methods. The hypothesis test results are shown in Table 7.

Table 7 The results of the hypothesis test of scientific data citation characteristic variables

Number	Hypothetical Description	Hypothetical Description	Results Interpretation
H1	There is a significant correlation between the Citation Label Mode, Citation Mark Position, and Citation Source Information Label Position.	Significant (LR chi –square value = 0.000)	There is no cross effect of 3 variables. Detailed explanation: Citation Source Information Label Position * Citation Label Mode (Effect value 3143.4) > Citation Source Information Label Position * Citation Mark Position (698.1) > Citation Mark Position * Citation Label Mode (37.8)
H2	The Citation Label Mode is significantly related to the Scientific Data Distributor.	Significant (Chi–square adjoint probability p=0.000)	The Citation Label Mode is strongly correlated with the type of Scientific Data Distributors.
H3	Distributors of scientific data are significantly related to address resolution.	Significant (Chi–square adjoint probability p=0.000)	The type of Scientific Data Distributors is highly correlated with the characteristics of the resolved address.

The multiple correspondence analysis method uses a method similar to factor analysis to classify variables and reduces the dimensionality, and uses a method similar to multi-dimensional scales to reflect the variable points and the relationship between the variables in a two-dimensional scatter plot.

The analysis process makes the closely-connected variable category points more concentrated, and the distantly connected points more scattered, thereby quantitatively describing and visually displaying data that does not have a clear connection. Observing the distance between the variable points in the corresponding analysis result graph, we can infer the strength of the connection between the variable elements (Du & Jia, 2009).

The SPSS multiple correspondence analysis module is used to draw the correlation of the variable characteristics of hypotheses 1 to 3, and the relevant results are shown below.

4.2.1 Correlation Analysis of Citation Label Mode, Citation Mark Position, and Citation Source Information Label Position

The log-linear model analysis results prove that there is a significant pairwise correlation between the three variables: citation labeling method, citation mark labeling position, and citation source information labeling position. Among them, the effect value of "Citation labeling method*Citation source information labeling position" is 3143.4, which is much larger than the effect of cross-influence between the other two variables. It not only shows that the

relationship between the characteristics of the two is very close, but also the cross effect of the two is representative for explaining the characteristics and differences of the citation mark behavior.

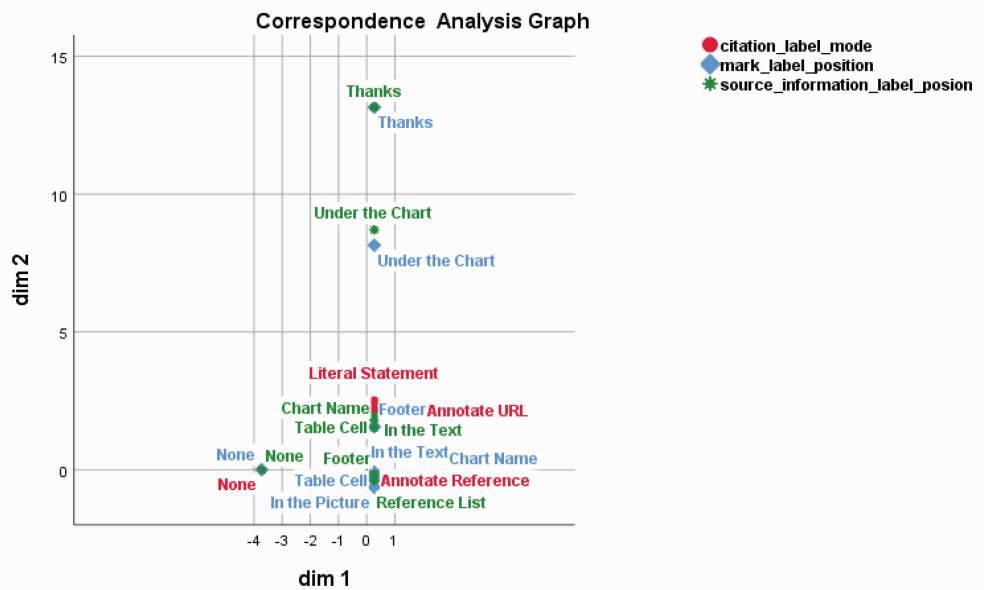

Figure 3 H1 Correspondence Analysis Graph

Combining the corresponding analysis diagram of the three (Figure 3), it can be seen that:

On the one hand, the "Literal Statement" and "Annotate URL" are relatively close, indicating that the two corresponding situations are similar, forming a type of citation and labeling behavior. These two citation labeling methods usually do not have special citation prompt marks (such as number superscript), but directly declare the source or URL address of the data in unformatted natural language at the place of citation. Such citation labeling situation is common in the text, chart names, and table cells.

"Annotate references" is another type of citation label behavior performance. Similar to traditional literature citations, researchers often use citation prompts (such as number superscripts) in texts or charts to declare citations, then take the bibliographic description format to mark the source information of the scientific data in the reference list at the end of the text, or in the footer.

4.2.2 Correlation Analysis of Citation Label Mode and Scientific Data Distributors

The corresponding analysis result of the citation labeling method and the Scientific Data Distributor is shown in Figure 4. The results show that the type of distributor (citing channel) of scientific data corresponds to the way of Citation Label Mode

With the origin (0, 0) as the center, " Annotate Reference " is the closest, indicating that it is the most common mode of citation label, and it mainly corresponds to research papers and reports, technical standards, patents, monographs, and newspapers. In addition, from a distance point of view, when citing scientific data from resource preservation venues, "Annotated Reference" tend to be used to declare the source. When citing scientific data from databases, Literal Statement are often used, and, relevant source information is often provided by "Annotating URL" for citing from Internet webpages.

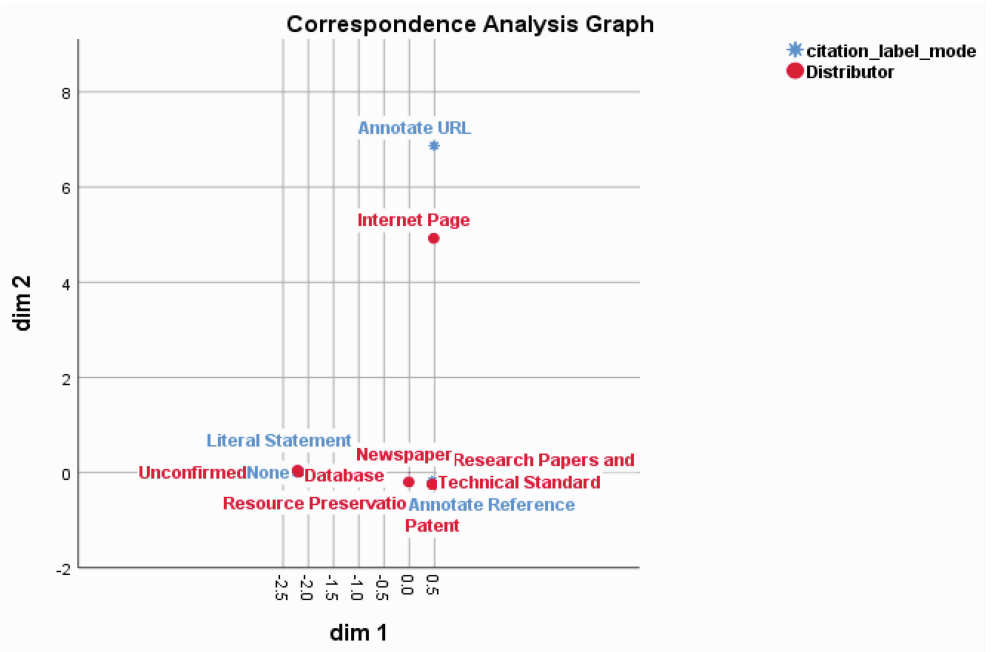

Figure 4 H2 Correspondence Analysis Graph

4.2.3 Correlation Analysis between Scientific Data Distributors and Resolution Address

The analysis result of the correspondence between the Scientific Data Distributor and the Resolution Address is shown in Figure 5. In general, the point where the Resolution Address is "none" is basically close to the origin, indicating that the unlabeled behavior of the re-

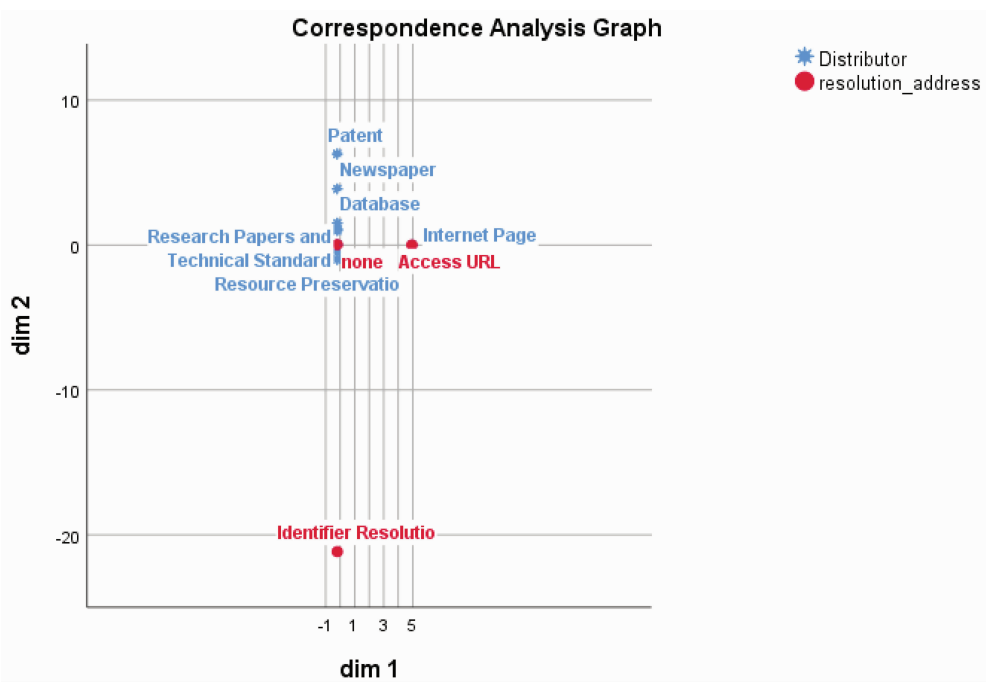

Figure 5 H3 Correspondence Analysis Graph

solved address of scientific data is a common phenomenon for any communication channel. The figure clearly shows that only in the case of citing scientific data from Internet webpages, researchers identify the URL address of the resource in order to achieve the purpose of indicating the source of the data. At the same time, corresponding to research papers and reports as the source of scientific data, the DOI identifier and Resolution Address of the source document are also very rare. However, according to the national standards for scientific data citation, the above are not standard scientific data citations.

5 Discussion and Conclusion

5.1 Summary

From the perspective of objective results of scientific data citations written by researchers in academic papers, the study took 4992 valid scientific data citation records identified in 771 Chinese academic journal papers from 2017 to 2019 as data samples.

By introducing the Chinese scientific data citation national standard *Information Technology-Scientific Data Citation (GB/T 35294-2017)* and related research results, 15 characteristic variables from the two aspects of citation label behavior characteristics and scientific data characteristics were statistically analyzed, and in-depth correlation analysis of variables was carried out through statistical methods.

The purpose of this study is to reveal the citation characteristics of scientific data by researchers and the significant correlation of characteristics.

Through various analyses, this study mainly draws the following conclusions:

(1) The overall citation level of scientific data

In general, scientific data citations in Chinese academic journal papers are common. About 83% papers have scientific data citations, and average citation data is about 6.47 per article. The average citations of the articles in 2017, 2018, and 2019 were 5.77, 6.40, and 7.29, respectively, inferring that the number of citations of scientific data is increasing. This result is similar to the results of existing studies (Ding et al, 2014a; Qu & Wang, 2017; Shi & Si, 2019). Consistency shows that the importance of scientific data in various disciplines has become increasingly prominent, and also reflects the development trend of data-intensive science paradigm.

(2) Standardization of scientific data citation

Compared with the scientific data citation description standard proposed by the national standard (GB/35294-2017), currently, the citation label modes are inconsistent, and the labeling of different citation elements of scientific data is quite different. This research shows that the order of the complete labeling degree of each element is (from high to low): Creator>Distributor>Production Time> Unique Identifier>Resolution Address. It can be inferred that Chinese researchers have not implemented the scientific data citation norms advocated by the national standards and norms, and the non-citing behavior is obvious. Among them, the phenomenon of unlabeled Unique Identifiers and resolved addresses of scientific data is particularly prominent, with unlabeled cases 77% and 96%, respectively.

(3) Characteristics of Scientific data citation label behavior

Previous studies have investigated the data citation forms of researchers in the fields of life sciences, geophysics, and social sciences, pointing out that there are multiple forms of scientific data citations coexisting (Shi & Si, 2019) .

Through the correlation analysis of citation feature variables, this study found that the Ci-

tation Label Mode, Citation Mark Position, and Citation Source Information Label Position are closely related, which together reflect the coherence performance of citation label behavior of scientific data. Among them, the Citation Source Information Label Position and the Citation Label Mode are more closely, and the correlation effect between the two is the key to distinguish different citation label behaviors.

At present, the citation and label behavior of scientific data of Chinese researchers roughly shows two typical behaviors: *Marking citation & Using references to indicate source and No mark & Text indicating source.* In the two performances, the position of the citation mark and the source information has formed a relatively fixed collocation, as shown in Table 8.

Table 8 Performances of scientific data citation label behavior

Citation Behavior	Citation Mark Position	Reference source information location
Marking citation & Using references to indicate source	In the text >Chart name >Table cell>In the picture	Reference List>Footer
No mark & Text indicating source	None	In the Text>Chart Name>Table Cell>Footer>Under the Chart>Thanks

Notes: The connection with ">" in the table indicates the order of the cumulative frequency of occurrence of different situations

Marking citation & Using references to indicate source is more common in scientific data citations (approximately 81%), indicating that Chinese researchers are still accustomed to using traditional literature citation formats for citing scientific data. At the same time, there are more inconsistencies in scientific data citation behaviors, such as Literal Statement and Annotations. URL, which account for about 12%, and some non-citing phenomena.

(4) Significant correlation of scientific data citation characteristics

The inference conclusions of the significant correlation analysis between the scientific data citation characteristics are as follows:

①The citation labeling method of the scientific researcher may be related to the type of distributor of the scientific data. Combined with the verification of Hypothesis 1 and Hypothesis 2, the situation of obtaining and citing scientific data through traditional communication media (such as academic journal articles and reports, monographs, technical standards, patents, newspapers) and resource preservation venues often corresponds to *Marking citation & Using references to indicate source*. The case of citing from databases and Internet pages mainly corresponds to *No mark & Text indicating source*.

②Only a few citation record samples are marked with scientific data resolution addresses, and the characteristics of the Resolution Addresses are related to the type of Scientific Data Distributor. For example, marking the DOI identifier and its resolution address basically correspond to the situation of citing scientific data from academic papers and reports. Most scientific data cited from Internet web pages directly identify the source URL address of the resource. However, in a strict sense, the current citation labeling of scientific data parsing addresses is almost in a "zero citation" state.

5.2 Enlightenments

Based on the above conclusions, this research provides some enlightenment for the related work of Chinese scientific data citation:

(1)At present, there are almost no scientific data citations that meet national standards among Chinese researchers. In general, the citation methods of scientific data tend to use traditional literature citations, and the citation and labeling behaviors are inconsistent. The reason may be that scientific data citation standards are not as well-received as traditional literature citations, leading to lack of citation awareness among researchers, or that citation standards cannot meet the citation needs of various types of scientific data in practice. Therefore, on the one hand, it is necessary to improve the formulation of universal and flexible scientific data citation standards, which is a necessary condition to support the interoperability and unification of citation in various disciplines. According to the characteristics of the user's citation behavior pattern, we should raise the awareness of correct citation, and put forward some ways to improve the comprehensiveness and convenience of data citation specifications. On the other hand, it is necessary to promote in-depth cooperation among relevant stakeholders in scientific data citation. Studies have proved that the regulations of the data warehouse play an important role in the standardization of scientific data citation (Liu et al, 2019). Hence, it is of great importance to plan an integrated citation platform for scientific data, and formulate appropriate data citation rules from the data distributors, actively provide users with standardized citation formats, and cooperate with relevant agencies (such as the journal editorial department) for review and supervision, which can improve the normative citation behavior of researchers to a certain extent.

(2)Unique Identifiers and Resolution Addresses are currently the most neglected elements in the citation label of scientific data, which may be due to insufficient understanding of the two by researchers, or low practicality in citation identification. Therefore, it is necessary to make a more accurate definition of the concept, scope, form of the scientific data parsing address and its association with the unique identification degree. Admittedly, this issue needs to be further explored and improved.

Finally, this research has some limitation:

(1) The sample is only 771 papers, and the sample depth and sample size are relatively limited. The universality of the conclusions needs to be further verified by more data.

(2) This study uses manual coding methods to analyze the objective results of scientific data citation. The manual coding process is cumbersome and low in efficiency, and is limited by personal knowledge and judgment ability, which may cause data coding errors.

(3) This study only focuses on the relevance of citation characteristic variables with strong correlation, which has limitations. The follow-up research needs to broaden the perspective and carry out more in-depth analysis.

References

Borjigin, C., Zhang, C., Sun Z., & Yi, Ni. (2021) . Theoretical Data Science: bridging the gap between domain–general and domain–specific studies. *Data Science And Informetrics, 1* (1) , 01–28.

Ding, N., Ding, Y., Yang, L., Ling, C., & Pan, Y. (2014a) . Data citation behavior in library and information science in China. *Journal of Library Science in China, 40* (6) , 105 –114. https://doi.org/10.13530/j.cnki.jlis.146010

Ding, N., Li, J., Li, Y., Bai, J., & Pan, Y. (2014b) . Scientific data evaluation based on data citation. *Library and Information, 5*, 95–99. https://doi.org/10.3969/j.issn.1003–6938.2014.05.019

Ding, N., Yang, L., Ding, Y., Ling, C., & Pan, Y. (2014c) . Data citation behavior in the journal papers of sociology in China. *Library and Information, 6*, 88–93.

Ding, W., Li, J., & Han, Y. (2019) . Research on the characteristics of scientific data citation in journal articles

in library and information science in China. *Library and Information Service, 63* (22) , 118–128. https://doi.org/10.13266/j.issn.0252–3116.2019.22.013

Du, Q., & Jia, L. (2009) . *SPSS statistical analysis from entry to proficiency*. People´s Mail Publishing House.

Gu, L. (2015) . Data level metric: Its concepts and progress. *Journal of Library Science in China, 2*, 56–71. https://doi.org/10.13530/j.cnki.jlis.150008

Liu, Y., Liu, J., Xiao, M., & Yu, J. (2019) . Research on the citation of scientific research data in domestic fund –sponsored papers. *Library Tribune, 39* (7) , 75 –83. https://doi.org/10.3969/j.issn.1002 –1167.2019.07.009

Liu, S., Ding, Y., & Zhang, C. (2015) . New stage of citation analysis: From citation description analysis to citation context analysis. *Document, Informaiton & Knowledge, 3*, 25 –34. https://doi.org/10.13366/j.dik.2015.03.025

Mo, Y. (2004) . Academic norms for data citation. *Edit Journal, 2*, 68 –69. https://doi.org/10.13530/j.cnki.jlis.150008

N*ational Standard of the People´s Republic of China (GB/T35294–2017) Information Technology–Scientific Data Citation.* (2017, December 29) . http://www.gb688.cn/bzgk/gb/newGbInfo?hcno =A495CA355BAF00D962AA8DD84C3B2C16

Qu, Y., & Wang, Y. (2017) . Research on the citation present situation and characteristics of scientific data in social science. *Digital Library Forum, 6*, 25–31. https://doi.org/10.3772/j.issn.1673–2286.2017.06.004

Shi, Y., & Si, L. (2019) . Analysis on data citation behavior characteristics of Chinese researchers. *Information Studies*: *Theory & Application, 42* (6) , 36–41. https://doi.org/10.16353/j.cnki.1000–7490.2019.06.007

Statistics How To. *Chi–Square Statistic: How to Calculate It / Distribution.* (2019, January 11) . http://www.statisticshowto.com/what–is–a–standardized–residuals/.

Zhang, L., & Li, J. (2014) . Analysis of the stakeholder of data citation. *Information Studies: Theory & Application, 37* (7) , 44–47. https://doi.org/10.16353/j.cnki.1000–7490.2014.07.014

Zhang, X. (2013) . Open access, open knowledge, and open innovation pushes for open knowledge services: 30 convergence and a new paradigmatic shift for research libraries. *Data Analysis and Knowledge Discovery, 2*, 1–10.