Postprint: A Study on the Usage Characteristics of Scientific Datasets Combining Bibliometric and Content Analysis

Authors: Yang Ning, Zhang Zhiqiang

Date: 2023-04-01T15:51:26+00:00

Abstract

[Purpose/Significance] This study investigates the usage characteristics of scientific datasets from both bibliometric and content analysis perspectives, quantitatively evaluates their impact on disciplinary development, and provides references for scientific data management services and policy research. [Method/Process] By comprehensively employing text mining and bibliometric methods to analyze full-text articles from PubMed Central, this research examines dataset usage across seven dimensions including temporal distribution and usage intensity, and subsequently assesses the actual impact of scientific datasets on disciplinary development. [Results/Conclusion] The results indicate that the influence of scientific datasets on biomedical research is growing increasingly significant; data publishing and high-level journals promote the openness and sharing of scientific datasets; dataset usage is concentrated in the latter sections of papers with relatively few formal citations; and the corresponding standards and norms require further strengthening.

Full Text

Research on the Use Characteristics of Scientific Datasets Combining Quantitative and Content Analysis

Yang Ning[1],[2], **Zhang Zhiqiang**[1],[2] [1] Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041 [2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract

[Purpose/Significance] This study investigates the usage characteristics of scientific datasets from both quantitative and content analysis perspectives,

quantitatively evaluating the impact of scientific datasets on disciplinary development to provide references for scientific data management services and policy research. [**Method/Process**] Integrating text mining and bibliometric methods, we analyzed the full-text literature from PubMed Central, comprehensively examining dataset usage across seven dimensions including temporal distribution and usage intensity. Based on this analysis, we evaluated the actual impact of scientific datasets on disciplinary development. [**Results/Conclusions**] The findings demonstrate that the influence of scientific datasets on biomedical research is growing daily. Data publishing and high-level journals promote the openness and sharing of scientific datasets. Dataset usage concentrates in the latter half of papers, with formal citations being relatively rare, indicating that corresponding standards and specifications require further strengthening.

Keywords: quantitative analysis; content analysis; scientific dataset; usage characteristics

1. Introduction

Scientific datasets are data materials or products generated during research activities or through reprocessing, with certain specifications and complete descriptions, primarily including experimental data, observational data, and statistical data [1]. With the rise of the open science movement, the sharing and reuse of scientific datasets have become increasingly common, gradually emerging as important research objects and output types throughout the entire research process. Studying the usage characteristics and impact of scientific datasets can, on one hand, help understand current data usage patterns and grasp researchers' needs and utilization behaviors regarding data; on the other hand, it can concretize and quantify the actual contribution value of scientific datasets to research activities and inform rational planning of scientific resource allocation and enrich research evaluation indicators.

Current research on scientific dataset usage characteristics generally employs either quantitative analysis or content analysis methods. Quantitative analysis is a quantitative approach based on mathematics and statistics that examines external and macroscopic features of knowledge entities [2]. From a quantitative perspective, studies typically use metrics such as citation frequency, download counts, and mention frequency to evaluate usage characteristics and impact. C. W. Belter et al. [3] studied dataset citation behavior in oceanography using citation counts to assess dataset impact. Jiao Hong et al. [4] employed bibliometric methods to conduct multi-dimensional analyses of reuse characteristics of biomedical datasets. Content analysis delves into full-text content of academic papers, examining usage behavior characteristics of various knowledge entities through manual interpretation or natural language processing [5]. From a content analysis perspective, studies typically examine usage patterns, locations, and intensity. Wang Xue et al. [6] analyzed literature from 10 disciplines

in CNKI, using content analysis to compare data reuse behaviors across disciplines from perspectives such as mention patterns, usage locations, and source types. Li Longfei et al. [7] adopted content analysis methods to study scientific dataset usage patterns and quantitatively measure their value from an altmetrics perspective. Content analysis operates at a more micro level, enabling investigation of dataset usage characteristics and impact from a granular article structure perspective.

However, due to significant disciplinary differences in scientific dataset usage characteristics and challenges in identifying and extracting dataset information from literature, most existing studies rely on manual annotation or small-scale datasets, with relatively broad analysis levels and indicators. This study examines large-scale academic paper collections in the biomedical field, combining quantitative and full-text content analysis methods to comprehensively investigate and analyze the usage characteristics of scientific datasets in academic papers, and further analyze their actual impact on disciplinary development from different perspectives. The significance of this research lies in utilizing full disciplinary paper collections, employing rule-based extraction and natural language processing techniques to explore dataset usage characteristics from both macro and micro perspectives, providing new perspectives for scientific data management and services while offering novel approaches for subsequent research.

2. Research Methods

2.1 Basic Approach

The full-text data for this study comes from PubMed Central (PMC), an open-access subset provided by the U.S. National Center for Biotechnology Information (NCBI) [8]. In addition to PMC, NCBI provides over 60 biomedical databases and related research tools, assigning unique identifiers—accession numbers—to scientific datasets in various formats. This study employs pattern matching based on custom rules to identify and extract dataset mentions from full texts as evidence of dataset usage in papers. We analyze dataset usage characteristics through both quantitative and content analysis dimensions, and summarize the actual influence of scientific datasets on biomedical research and disciplinary development. The overall research framework is shown in Figure 1 [Figure 1: see original paper].

2.2 Data Acquisition

We bulk-downloaded PMC file packages via FTP service before May 25, 2021. After merging index files, we obtained basic literature information and local file locations. PMC full-text data is stored in XML format using the National Library of Medicine (NLM) Document Type Definition (DTD) standard [9]. Ultimately, we acquired 3,219,908 full-text articles.

Scientific dataset identification employs pattern matching through regular expressions in full-text content. Since accession number rules vary across NCBI databases and many consist solely of numbers that cannot be automatically extracted via pattern matching, this study selected five commonly used databases with well-defined formats and detailed accession number specifications: GEO [10], RefSeq [11], SRA [12], CDD [13], and Assembly [14]. GEO is currently the largest and most comprehensive gene expression database, collecting microarray and high-throughput sequencing data submitted and shared by researchers worldwide. RefSeq is a reference sequence database containing genomic and transcript sample information, providing sequence data and related materials for various organisms. The SRA database primarily stores next-generation sequencing raw data and associated quality control reports. CDD is a protein conserved domain database collecting extensive conserved domain sequence and protein sequence information. The Assembly database provides assembled genomic structures, related metadata, and assembly reports. Table 1 shows the regular expressions constructed based on each database's accession number rules.

Additionally, some literature contains batch usage patterns such as "GSE4357-GSE4380" or "SRX001799 to SRX001808," requiring separate batch extraction rules with a maximum extraction threshold of 500; beyond this limit, such patterns are ignored. After identification and extraction, we found 162,200 articles using datasets from the five databases, with a total of 435,920 datasets used 2,606,552 times. Articles exhibiting dataset usage behavior account for 5.04% of all articles. Table 2 shows the distribution of dataset usage across the five databases, with RefSeq containing 238,023 used datasets (approximately 55% of the total), indicating that datasets from this database receive substantial attention and usage in the biomedical field.

2.3 Quantitative Analysis Indicators

Quantitative analysis employs direct indicators from datasets and dataset-using literature to examine usage characteristics, including temporal distribution, document type, disciplinary distribution, and high-frequency datasets. Quantitative analysis uses the CountOne method [15], counting multiple uses of a dataset in a paper as only one occurrence. Specific analysis contents include: Temporal distribution: analyzing annual trends in literature volume and dataset usage counts to identify temporal patterns; Document type: beyond research articles and reviews, dataset usage appears in reports, briefings, comments, etc., with statistical analysis revealing characteristic patterns across document types; Disciplinary distribution: exploring differential data usage needs across disciplines from the perspective of publishing journals' disciplinary affiliations; High-frequency datasets: analyzing characteristics of frequently used datasets by ranking papers using specific datasets to identify research hotspots and researcher preferences.

2.4 Content Analysis Indicators

Content analysis employs detailed information about dataset mentions and usage in literature as indirect indicators, including usage intensity, usage sections, and usage locations. Content analysis adopts the CountX method [16], incorporating all usage records of a dataset in a paper into the analysis. Table 3 provides detailed explanations of each indicator: Usage intensity: using average usage frequency per paper to evaluate dataset impact within articles; Usage sections: dividing data usage into five detailed parts by section type to compare dataset usage across different paper sections; Usage location: categorizing eight data usage and presentation locations to compare dataset usage characteristics in papers.

3. Results Analysis

3.1 Quantitative Analysis Results

3.1.1 Temporal Distribution From 1998 to 2021, 162,200 biomedical articles used 435,920 datasets. Figure 2 [Figure 2: see original paper] shows the annual distribution of articles and dataset usage counts. After 2006, with the transformation of research paradigms and the rise of data-driven disciplines such as bioinformatics and medical informatics, both the number of articles using datasets and dataset usage quantities began to grow dramatically. The number of articles increased from 724 in 2006 to 27,279 in 2020, with an average annual growth rate of 35.5%. Dataset usage counts grew from 24,783 in 2006 to 400,320 in 2020, with an average annual growth rate of 31.5%. Scientific data sharing and reuse are profoundly influencing the development of biomedical research fields, particularly opening new development avenues for biomedicine in the past decade.

3.1.2 Document Type We identified 29 document types with dataset usage behavior. Ranked by quantity, these include: research articles, briefings, reviews, case reports, others, data papers, communications, corrections, product reviews, abstracts, methods papers, editorials, systematic reviews, reports, article commentaries, meeting reports, protocols, calendars, appendices, announcements, retractions, chapter articles, concern statements, replies, book reviews, research letters, descriptions, news. Research articles account for approximately 92% of the total literature. Figure 3 [Figure 3: see original paper] shows the distribution across document types.

Excluding research articles, Figure 4 [Figure 4: see original paper] shows the annual publication volume distribution for the seven document types with relatively high dataset usage. Among these, product reviews first used datasets in 2004 (two articles using RefSeq and CDD datasets for gene database construction and protein specificity alignment software development testing [17-18]). Subsequently, datasets began appearing in briefings, reviews, and case reports, with review literature showing steady year-over-year growth in dataset usage,

indicating that datasets have become research materials integrated into disciplinary development histories. Additionally, data papers emerging since 2014 have grown rapidly. As a new academic publication format primarily describing data structure, processing methods, and reusability, data papers are actively promoting scientific data development and utilization [19].

3.1.3 Disciplinary Distribution Literature using datasets was published in 3,127 journals. The journal with the most publications is *PLOS ONE*, with 20,931 articles using datasets. To ensure broad coverage and enhance interpretability, we excluded journals with fewer than 100 publications, yielding 229 journals with a total of 131,359 articles (approximately 81% of the total literature). Using the 2019 Journal Citation Reports from the National Science Library, Chinese Academy of Sciences [20], we examined and evaluated the research fields and impact of these top 229 journals. We found 181 SCI-indexed journals, with 120 (66%) classified as Q1 or Q2. Disciplinary distribution and journal quartiles are shown in Figure 5 [Figure 5: see original paper].

From a disciplinary perspective, biology journals account for 56% of the total, with biochemistry, molecular biology, genetics, and cell biology showing the most frequent dataset usage. In medical fields, research & experimental medicine, oncology, and psychiatry journals are most numerous, representing the medical disciplines with highest dataset usage. The results also include comprehensive disciplines, food science, and agricultural sciences, demonstrating the interdisciplinary and cross-disciplinary nature of scientific dataset usage.

3.1.4 High-Frequency Datasets Statistical analysis of dataset usage frequency reveals that 346,115 datasets (79% of the total) were used only once. Plotting dataset usage frequency against dataset count yields the relationship shown in Figure 6 [Figure 6: see original paper] in both original and double logarithmic coordinates. Univariate linear regression produces: log(dataset count) = 4.59 - 1.91 log(dataset usage frequency), with $R^2 = 0.88$, showing a clear linear relationship. The results indicate that numerous datasets receive minimal usage while a small number receive extensive usage.

Table 4 details the top 20 high-frequency datasets. Five datasets originate from GEO, with the remaining 15 from RefSeq. The most frequently used dataset, "GPL570," is a commercial dataset from Affymetrix, a renowned U.S. biochip company; the other four GEO datasets also come from this company's chip products. Among the top 20 datasets, five focus on tumor research, three on actin function, three on the human genome, three on glyceraldehyde-3-phosphate dehydrogenase, with others related to interleukins, *Mycobacterium tuberculosis*, *Escherichia coli*, and SARS-CoV-2 research. High-frequency dataset usage intuitively reflects disciplinary research hotspots.

3.2 Content Analysis Results

3.2.1 Usage Intensity Traditional frequency metrics only indicate whether a dataset appears in a paper. However, if dataset A is used repeatedly in a paper while dataset B appears only once, dataset A's impact on that article should be greater. Therefore, this study employs usage intensity to analyze dataset usage characteristics and influence. The dataset "NR_{033736}" from RefSeq exhibits the highest usage intensity, being used 768 times in a single paper [21]. Based on overall usage patterns, we divided scientific dataset usage intensity into 11 intervals, with results shown in Figure 7 [Figure 7: see original paper].

Figure 7 reveals that biomedical dataset usage intensity primarily falls between 1-6, with the highest concentration in the 2-3 interval, followed by 1, 5-6, and 1-2. This differs markedly from paper citations, as scientific datasets exhibit more high-intensity usage phenomena, indicating that a single dataset may be used repeatedly throughout the research process.

3.2.2 Usage Sections Different paper sections carry varying importance, and datasets used in different sections consequently have different significance and impact. Following the IMRaDC structure of empirical research papers [22], we divided sections into five parts: abstract, introduction, data & methods, results & discussion, and conclusion. Tables and figures listed in appendices were assigned to corresponding sections via "id" markers. For non-research articles such as data papers and product reviews that cannot be mapped to these five sections, we manually assigned them to functionally similar sections; those that could not be assigned were excluded from statistical analysis (a small proportion that does not significantly affect results). Section distribution results are shown in Figure 8 [Figure 8: see original paper].

Figure 8 shows that 49% of dataset usage occurs in the "data & methods" section, followed by "results & discussion." The "abstract" provides overviews without extensive data elaboration; the "introduction" includes brief background on methods and datasets; the "data & methods" and "results & discussion" sections focus on experimental data analysis and interpretation, making them the two most dataset-intensive sections, accounting for approximately 95% of total dataset usage. The "conclusion" section contains minimal specific dataset descriptions. Overall, dataset usage section distribution shows extreme imbalance, reflecting the emphasis on empirical analysis and results interpretation in biomedical literature and demonstrating the critical importance and influence of scientific datasets in this field.

3.2.3 Usage Locations Similar to usage sections, datasets used in different locations carry different importance and impact. We categorized usage locations into eight types: main text, tables, figures, references, acknowledgments, appendices, footnotes, and annotations. Main text includes datasets appearing in titles, abstracts, and body text. Location distribution results are shown in Figure 9 [Figure 9: see original paper].

Figure 9 indicates that biomedical datasets most frequently appear in tables, followed by main text descriptions and figure captions. In biomedical literature, tables and figures carry equal importance to main text, warranting attention to table and figure data identification and utilization in related research. Notably, datasets in references account for only 0.04% of total usage, indicating that formally cited datasets remain rare. This suggests that scientific dataset formal citation issues warrant greater attention, as standardized citation is crucial for enhancing data value and promoting researchers' enthusiasm for data sharing and reuse.

4. Discussion and Conclusions

Based on the above results, we draw the following conclusions:

(1) The influence of scientific datasets on biomedical research is increasing daily. Statistics based on paper counts and usage intensity represent the breadth and depth of dataset usage, respectively. Broader usage scope indicates greater actual impact, and the dramatic growth in papers using scientific datasets over the past decade demonstrates their increasing influence. Usage intensity reveals unique dataset usage characteristics, showing significantly higher intensity than citation intensity for papers [23] and books [24], indicating that datasets are less frequently mentioned as background references and more often directly used in conjunction with research results.

(2) Data publishing and high-level journals promote scientific dataset openness and sharing. Analysis of document types and disciplinary distribution shows that datasets are gradually becoming independent research materials playing key roles in scientific communication. Current data publishing models—including data repositories, data journals, and joint data-paper publishing—particularly the emergence of data journals, have made data papers the fastest-growing vehicle for scientific data publication, establishing datasets as evaluable, measurable research outputs. Investigation of Q1 journals reveals that all 51 journals provide detailed dataset submission requirements in author guidelines, demonstrating that high-level journal initiatives in open data accelerate data sharing and reuse, advancing research progress.

(3) Dataset usage concentrates in the latter half of papers, with few formal citations. Usage sections and locations show that datasets most frequently appear in tables, followed by main text mentions. Researchers should 重视 table and figure data mining. The most common sections are "data & methods" and "results & discussion," contrasting with other fields where papers and books are typically cited in "introduction" sections [25-26]. Biomedical papers frequently cite literature and use datasets in "results & discussion," indicating this is the most important section, with approximately 95% of dataset usage occurring in the latter half of papers. The

small proportion of formally cited datasets in references suggests that datasets are primarily listed through informal mentions, indicating both the large number of datasets involved in biomedical research and the need for further development of data citation standards. Formal citation is essential for enhancing data value and motivating researchers to share and reuse data.

This study provides comprehensive analysis from macro and micro perspectives, offering more complete and reliable results than previous research and providing references for scientific data management and services. Future efforts should: (1) advance establishment of scientific data citation standards and improve unique identifier and version management by scientific databases; (2) ensure scientific database construction is professional, timely, and open, with specialized databases maintained by professional teams and peer reviewers, and guarantee free and open access through optimized multi-channel funding; (3) strengthen scientific data talent cultivation in universities and libraries, including data management researchers, data analysts, and data curators to meet rapidly developing data management and service needs.

However, this study has limitations: (1) due to identification method constraints, we only examined datasets from five NCBI databases with relatively standardized accession numbers, limiting research scope; (2) we analyzed only from the paper perspective without examining dataset metadata and content, equating dataset mentions with usage without further investigating usage intentions, leaving room for deeper analysis. Future work will improve dataset identification scope and accuracy, analyzing dataset usage characteristics and impact from more granular perspectives.

References

[1] Qu Baoqiang, Wang Kai. Current status and research progress of scientific data citation[J]. Information Theory and Practice, 2016, 39(5): 118-138. [2] Zhu Shaoqiang, Qiu Junping. Bibliometrics and content analysis: mining implicit information in document collections[J]. Library and Information Service, 2005(6): 19-23. [3] BELTER C W, BROWMAN H I. Measuring the value of research data: a citation analysis of oceanographic datasets[J]. PloS one, 2014, 9(3): e92590. [4] Jiao Hong, Yang Bo, Zhou Qi. Research on reuse characteristics of scientific datasets in biomedical field[J]. Information Theory and Practice, 2021, 44(9): 90-96. [5] Wang Yuefen, Lu Fei, Wu Xiaolei. Comparative and comprehensive study of bibliometrics and content analysis[J]. Library and Information Service, 2005, 49(9): 72-75. [6] Wang Xue, Ma Shengli, She Zengli, et al. Research on citation behavior and impact of scientific data[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(11): 1132-1139. [7] Li Longfei, Yu Houqiang, Yin Zihan, et al. Quantitative measurement of scientific dataset value from altmetrics perspective[J]. Information Theory and Practice, 2020, 43(9): 47-52, 71. [8] Shen Xibin, Lü Xiaodong, Hao Xiuyuan, et al. Introduction to PubMed Central and its evalu-

ation and inclusion of journals[J]. Chinese Journal of Scientific and Technical Periodicals, 2006, 17(5): 866-868. [9] Shen Xibin, Gu Jia, Bao Jingling, et al. Interpretation of reference marking in NLM DTD 3.0 journal storage and exchange tag set[J]. Chinese Journal of Scientific and Technical Periodicals, 2013, 24(2): 233-237. [10] NCBI. Gene expression omnibus[EB/OL].[2021-07-12]. https://www.ncbi.nlm.nih.gov/geo/. [11] NCBI. Reference sequence database[EB/OL].[2021-07-12]. https://www.ncbi.nlm.nih.gov/refseq/. [12] NCBI. Sequence read archive[EB/OL].[2021-07-12]. https://trace.ncbi.nlm.nih.gov/Traces/sra/. [13] NCBI. Conserved domains database[EB/OL].[2021-07-12]. https://www.ncbi.nlm.nih.gov/cdd/. [14] NCBI. Assembly[EB/OL].[2021-07-12]. https://www.ncbi.nlm.nih.gov/assembly/. [15] WAN X, LIU F. WL-index: leveraging citation mention number to quantify an individual's scientific impact[J]. Journal of the American Society for Information Science & Technology, 2014, 65(12): 2509-2517. [16] DING Y, LIU X, GUO C, et al. The distribution of references across texts: some implications for citation analysis[J]. Journal of informetrics, 2013, 7(3): 583-592. [17] WANG B B, BRENDEL V. The asrg database: identification and survey of arabidopsis thaliana genes involved in pre-mRNA splicing[J]. Genome biology, 2004, 5(12): 1-23. [18] MEEREIS F, KAUFMANN M. Pcog: phylogenetic cog ranking as an online tool to judge the specificity of cogs with respect to freely definable groups of organisms[J]. BMC bioinformatics, 2004, 5(1): 150-150. [19] Qu Baoqiang, Wang Kai. Emergence and development of data papers[J]. Library and Information, 2015(5): 1-8. [20] National Science Library, Chinese Academy of Sciences. Journal Citation Reports of Chinese Academy of Sciences[EB/OL].[2021-07-12]. http://www.fenqubiao.com/. [21] LI J, JIN K, LI M, et al. A host cell long noncoding RNA nr_{033736} regulates type I interferon-mediated gene transcription and modulates intestinal epithelial anti-cryptosporidium defense[J]. PLoS Pathogens, 2021, 17(1): e1009241. [22] LIN L, EVANS S. Structural patterns in empirical research articles: a cross-disciplinary study[J]. English for specific purposes, 2012, 31(3): 150-160. [23] Hu Zhigang. Full-text citation analysis methods and applications[M]. Beijing: Science Press, 2017. [24] Zhang Chengzhi, Li Zhuo, Zhao Mengyuan, et al. Construction of standardized datasets for citation content analysis of Chinese books[J]. Library Forum, 2016(8): 48-53. [25] Zhang Mengying, Lu Chao, Zheng Rujia, et al. Standardized dataset construction for citation content analysis[J]. Journal of Library Science in China, 2019, 45(3): 96-109. [26] CHIPS. Differing disciplinary citation concentration patterns of book and journal literature?[J]. Journal of informetrics, 2016, 10(3): 814-829.

Author Contributions

Yang Ning: Data collection, experimental validation, and initial draft writing.
Zhang Zhiqiang: Supervised paper revision and refined research points.

Research on the Use Characteristics of Scientific Datasets Combined with Quantitative Analysis and Content Analysis

Yang Ning[1],[2], **Zhang Zhiqiang**[1],[2]

[1] Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

[**Purpose/significance**] This paper analyzes the use characteristics of scientific datasets from the perspectives of quantitative analysis and content analysis, quantitatively evaluates the impact of scientific datasets on disciplinary development, and provides references for scientific data management services and policy research. [**Method/process**] Methods of text mining and bibliometrics were used to analyze the full-text literature in PubMed Central. This study comprehensively investigated the use of scientific datasets from seven aspects such as time distribution and use intensity, and on this basis, evaluated the actual impact of scientific datasets on disciplinary development. [**Result/conclusion**] The research results show that the influence of scientific datasets on scientific research in the biomedical field is increasing with each passing day. Data publishing and high-level journals promote the opening and sharing of scientific datasets. The use of scientific datasets is concentrated in the second half of the paper and there are few formal references. The corresponding standards and specifications need to be further strengthened.

Keywords: quantitative analysis; content analysis; scientific dataset; use characteristics

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