An Exploratory Analysis of Knowledge Diffusion Characteristics in Scientific Datasets: A Case Study of Gene Expression Datasets (Postprint)

Authors: Yang Ning, Zhiqiang Zhang

Date: 2023-04-01T15:51:27+00:00

Abstract

[Purpose/Significance] By investigating the characteristics and patterns of knowledge diffusion of scientific datasets, this study explores their actual role in disciplinary development, providing references for the scientific and technological evaluation of datasets and the formulation of management policies.

[Method/Process] Taking datasets from the GEO database and full-text data from PubMed Central that reuse datasets as analysis objects, this study employs content analysis combined with knowledge diffusion indicators such as diffusion breadth, diffusion intensity, and diffusion speed to investigate the knowledge diffusion characteristics of scientific datasets.

[Results/Conclusion] The results indicate that the breadth and intensity of knowledge diffusion from scientific datasets are increasing, that data reuse can accelerate the speed of knowledge diffusion, and that China's status in the global scientific data domain is continuously improving.

Full Text

Preamble

Exploring the Characteristics of Knowledge Diffusion in Scientific Datasets: A Case Study of Gene Expression Datasets

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] By investigating the characteristics and patterns of knowledge diffusion in scientific datasets, this study explores their practical role in disciplinary development, providing references for scientific dataset evaluation and management policy formulation. [Method/Process] Using datasets from the GEO database and full-text data of dataset-reusing publications from PubMed Central as analytical objects, this paper employs content analysis combined with knowledge diffusion indicators such as diffusion breadth, diffusion intensity, and diffusion speed to examine the knowledge diffusion characteristics of scientific datasets. [Result/Conclusion] The results demonstrate that both the breadth and intensity of knowledge diffusion in scientific datasets are increasing daily, that data reuse can accelerate the speed of knowledge diffusion, and that China's position in the global scientific data field continues to strengthen.

Keywords: scientific dataset; knowledge diffusion; characteristic analysis; measurement indicators
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2022.12.008

Introduction

Scientific data refers to data materials generated or reprocessed during research activities, primarily including experimental data, observational data, and statistical data. While any unit of data can be called scientific data, collections or products of related scientific data created, collected, and organized for specific research purposes constitute scientific datasets. With the widespread adoption of data-driven research paradigms across various disciplines, scientific datasets have gradually become important research objects and outputs throughout the scientific process. These data materials obtained through experiments or observation not only accelerate research progress but also propagate, inherit, and innovate their inherent knowledge value during dataset sharing and reuse, thereby achieving knowledge diffusion.

Knowledge diffusion refers to the cross-temporal flow of knowledge through certain carriers, promoting the generation of new knowledge and scientific innovation through knowledge absorption and recombination. Analyzing the knowledge diffusion of scientific datasets holds significant practical importance for expanding the scope of knowledge diffusion research, deeply understanding the academic value of scientific datasets, promoting standardized data citation, and facilitating data sharing and reuse.

In 1924, W.S. Learned of the Carnegie Foundation first studied knowledge diffusion in his book *The American Public Library and the Diffusion of Knowledge*. Current domestic and international research on knowledge diffusion can be categorized into three main types: (1) Knowledge diffusion unit research, which

investigates diffusion characteristics and patterns based on various units such as papers, patents, authors, journals, and disciplines. For instance, Huang Lucheng et al. proposed a framework for studying technical knowledge diffusion characteristics based on complete patent citation information, exploring diffusion from the perspectives of knowledge utilization and propagation. Zhao Rongying et al. constructed author knowledge diffusion networks to identify processes and patterns while evaluating author contributions. Yue Zenghui et al. examined disciplinary knowledge diffusion characteristics using social network disciplines as research objects. Wang Jingjing et al. analyzed cross-disciplinary knowledge diffusion trends in international digital humanities research, finding that core disciplines like library and information science were declining in centrality while arts, humanities, and engineering gained importance. (2) Knowledge diffusion indicator research, which measures diffusion through bibliometric or network indicators. Y.X. Liu et al. proposed indicators for disciplinary knowledge diffusion breadth, intensity, and speed based on ESI classifications. Yu Liping et al. developed the CJH index to reflect knowledge diffusion depth in academic journals, referencing the h-index calculation method. H. Nakamura et al. proposed using the time difference between citing and cited publications as a knowledge diffusion delay indicator. Song Ge utilized diffusion theory, social network analysis, and citation analysis to propose measurement indicators for innovation diffusion breadth, speed, intensity, and delay from a knowledge network structure perspective. (3) Knowledge diffusion model research, which examines diffusion and evolution processes through various models. I.Z. Kiss et al. proposed an individual-based directed weighted knowledge diffusion model based on epidemic models to describe how research topics diffuse across disciplines. X. Gao et al. developed a citation-based temporal network knowledge diffusion model integrating social network analysis, visualization, and citation analysis to reveal diffusion processes from a network structure perspective.

Existing research primarily focuses on papers, patents, and authors as carriers, constructing networks through citation or co-occurrence relationships to study diffusion characteristics. Although recent studies have examined knowledge diffusion using books, software, and funding as knowledge units, few have analyzed scientific datasets as research outputs. This gap stems from two main reasons: First, the lack of unified data citation standards means scientific datasets often appear in papers as informal mentions rather than formal citations, making their usage difficult to trace and quantify. Research has found that datasets indexed in the Data Citation Index (DCI) exhibit high rates of zero citations, with limited diffusion breadth and depth, creating significant challenges for citation-based studies. Second, current research on dataset citation still relies primarily on sampling surveys and manual content analysis, with limited document quantities and scope, failing to delve into dataset metadata or identify macro-level patterns applicable to knowledge diffusion research.

This study addresses these limitations by using gene expression datasets from the biomedical domain as research objects. It obtains author, institutional, and dataset publication date information from metadata, identifies literature

reusing datasets through content analysis, establishes citation relationships between datasets and publications, and employs knowledge diffusion indicators to analyze dataset diffusion characteristics in scholarly communication. This research contributes by: (1) introducing scientific data into knowledge diffusion research to enrich theoretical and methodological understanding of dataset diffusion processes, and (2) expanding research evaluation content and application areas to provide new perspectives for scientific data management and services.

---

2. Research Methods

2.1 Basic Approach

The scientific dataset data in this study originates from the Gene Expression Omnibus (GEO) database, a global high-throughput molecular abundance database created and maintained by the National Center for Biotechnology Information (NCBI). GEO is currently the world's largest and most comprehensive gene expression database, collecting microarray and high-throughput sequencing data submitted and shared by researchers worldwide. GEO classifies and stores user-submitted data, assigning each dataset a unique and permanent accession number. It requires that research papers using shared data make the data publicly available to facilitate subsequent research.

This study first obtains GEO dataset metadata for multi-perspective bibliometric and trend analysis. Knowledge diffusion research focuses on data reuse relationships—also called data reutilization or secondary data analysis—which began in the 1990s and refers to the process of reanalyzing original or recombined datasets to reproduce results or for new research purposes. We retrieve full-text biomedical literature from PubMed Central (PMC), a free full-text biomedical journal database provided by NCBI. Using pattern matching with regular expressions, we extract accession numbers from full texts to identify dataset usage information. Publications with dates after dataset publication dates are defined as dataset-reusing literature, establishing citation relationships between datasets and papers for knowledge diffusion analysis. The overall research framework is shown in Figure 1 [Figure 1: see original paper].

2.2 Data Acquisition

GEO's raw data is organized into three entity databases: platform (GPL), sample (GSM), and series (GSE), which are independent yet interconnected. Platforms contain description and annotation information for chips or sequencing platforms, typically including samples from multiple submitters. Samples record gene expression measurement data for individual samples as the basic unit of raw experimental results. Series comprise multiple samples forming biologically meaningful datasets. Additionally, GEO classifies raw data from "experiment" and "gene" perspectives into Datasets and Profiles databases. This study retrieved complete dataset information for all three raw data types.

We also bulk-downloaded PMC file packages before May 25, 2021, via FTP service, merged index files to obtain literature metadata and local file locations, and parsed PMC full-text data using Python, ultimately acquiring 3,219,908 full-text articles. Dataset usage identification employed pattern matching with regular expressions. Table 1 shows basic information for scientific datasets, including quantities, main metadata fields, and extraction rules for the three raw data types.

Analysis of extraction results revealed bulk usage patterns such as "GSE4357-GSE4380" or "GSE4357 to GSE4380," requiring separate bulk extraction rules with a maximum threshold of 500 datasets; beyond this threshold, extraction was not performed. Publications with dates after dataset publication dates were defined as reusing literature. Extraction identified 39,189 papers reusing GEO datasets, totaling 57,841 datasets with 294,517 reuse instances, representing 1.22% of all literature.

2.3 Scientific Dataset Knowledge Diffusion Indicators

Drawing on previous indicator definitions for other knowledge diffusion units and considering scientific dataset characteristics, this study proposes four indicators: data knowledge diffusion breadth, intensity, speed, and delay.

(1) Data Knowledge Diffusion Breadth (DKDB): This indicator analyzes diffusion coverage—the more papers reusing a dataset, the greater its diffusion breadth and the more knowledge recipients. I. Rowlands first proposed knowledge diffusion breadth metrics in 2002, subsequently refined by T.F. Frandsen, Qiu Junping, and others. We propose DKDB as calculated in Formula (1):

$$DKDB = N_i / Y_{pub} \quad \text{Formula (1)}$$

where N_i represents the number of papers reusing datasets published in the statistical year, and Y_{pub} represents dataset age. As knowledge diffusion is a dynamic cumulative process, we also examine cumulative diffusion breadth (DKDB*), calculated in Formula (2):

$$DKDB^* = \sum_{i=1}^{n} N_i \quad \text{Formula (2)}$$

where $1 \leq i \leq n$, N_i represents papers reusing datasets published in year i, and n is the total number of statistical years. These indicators reflect annual and cumulative diffusion breadth trends.

(2) Data Knowledge Diffusion Intensity (DKDI): This indicator analyzes reuse frequency—the more frequently a dataset is reused, the greater its diffusion intensity and influence on knowledge recipients. Similar to breadth measurement, we examine both DKDI and cumulative intensity (DKDI*), calculated in Formulas (3) and (4):

$$DKDI = N_j / Y_{pub} \quad \text{Formula (3)}$$

where N_j represents total reuse frequency of datasets published in the statistical year. Cumulative intensity (DKDI*) is calculated as:

$$DKDI^* = \sum_{j=1}^{n} N_j \quad \text{Formula (4)}$$

where $1 \leq j \leq n$, N_j represents total reuse frequency for datasets published in year j. These indicators reflect annual and cumulative diffusion intensity trends.

(3) Data Knowledge Diffusion Speed (DKDS): This indicator analyzes propagation distance per unit time. In 2005, R. Rousseau proposed the "average diffusion speed" metric, defined as the ratio of journals citing a paper to paper age. We adapt this for datasets, defining DKDS as the ratio of journals publishing papers reusing a dataset to dataset age, calculated in Formula (5):

$$DKDS_y = \frac{1}{m} \sum_{i=1}^{m} (P_i / Y_{pub}) \quad \text{Formula (5)}$$

where $DKDS_y$ is the average diffusion speed for datasets published in a given year, P_i represents the number of journals publishing papers reusing datasets from that year, Y_{pub} represents dataset age, and m is the number of datasets published that year.

(4) Data Knowledge Diffusion Delay (DKDD): This indicator references the "citation lag" metric, representing the time difference between a dataset's first reuse and its publication date, revealing diffusion efficiency from a temporal perspective. It is calculated in Formula (6):

$$DKDD = T_1 - T_0 \quad \text{Formula (6)}$$

where T_1 represents the publication year of the first paper reusing the dataset, and T_0 represents the dataset's publication year.

---

3. Results and Analysis

3.1 Basic Information of Scientific Datasets

The reused GEO datasets were published between 2000-2021, totaling 57,841 datasets with 294,517 reuse instances (average 5.092 reuses per dataset). Papers exhibiting GEO dataset reuse were published between 2004-2021, totaling 39,189 papers (average 1.476 datasets per paper) in 1,337 journals. The earliest

reuse record dates to 2004 in M.V. Osier et al.'s study, which used four GEO datasets (GPL205, GPL218, GPL229, GPL356) to test a microarray data analysis protocol. Annual distributions are shown in Table 2 and Figure 2 [Figure 2: see original paper].

Due to publication lag in biomedical research, heavily reused datasets concentrate in 2008-2017, aligning with the rise of data-driven disciplines. Ranking datasets by reuse frequency revealed that 43,217 datasets (74.72%) were reused only once. The most reused dataset was GPL570 from Affymetrix, with 1,634 reusing papers. Plotting reuse frequency (X-axis) against dataset count (Y-axis) yields Figure 3 [Figure 3: see original paper], showing a clear power-law distribution ($R^2 = 0.99$): most datasets receive minimal reuse while a minority receives extensive reuse.

From a geographical perspective, 53,419 datasets included contributor nationality information. The United States ranked first with 27,187 datasets, while China ranked second with 3,976 datasets. Early datasets from 2000-2001 came exclusively from the U.S. After 2002, more countries began publishing reusable datasets. China's first published and reused dataset appeared in 2005 (lentinus edodes gene expression data from Chinese University of Hong Kong). With China's increasing data sharing policies and researcher awareness, China's data contribution share has grown from 1.35% in 2005 to over one-third in recent years. Figure 4 [Figure 4: see original paper] shows annual publication share changes for the top 10 countries.

3.2 Data Knowledge Diffusion Breadth

Diffusion breadth examines coverage from a scope perspective—annual distribution and trends of papers reusing datasets. Calculations for GEO datasets are shown in Table 3 . DKDB increased from 3.182 in 2000 to a peak of 495.333 in 2019, with an annual average of 233.534. Using $\log_{1000}(DKDB)$ as the annual indicator, Figure 5 [Figure 5: see original paper] compares annual and cumulative diffusion breadth. Annual breadth showed fluctuating growth until 2019, then declined. Cumulative breadth exhibited an S-curve, indicating datasets gradually gained influence and sustained attention, driving knowledge integration and innovation in biomedicine. The recent decline reflects the time lag between dataset publication and reuse.

3.3 Data Knowledge Diffusion Intensity

Diffusion intensity examines impact from a frequency perspective—annual distribution and trends of reuse frequency. Calculations are shown in Table 4 . DKDI increased from 23.273 in 2000 to a peak of 3,350.400 in 2017 before declining, with an annual average of 1,607.756. Using $\log_{1000}(DKDI)$ as the annual indicator, Figure 6 [Figure 6: see original paper] compares annual and cumulative intensity. Annual intensity grew fluctuatingly until 2017, then declined due to dataset lag. The correlation coefficient between DKDI and DKDB

is 0.998, showing extremely high correlation. Both indicators exhibit nearly identical annual trends, reflecting increasing researcher dependence on datasets as biomedical research advances in sequence alignment and gene identification.

3.4 Data Knowledge Diffusion Speed

Diffusion speed reflects researcher attention and utilization efficiency—shorter times indicate faster diffusion. Faster propagation reduces knowledge aging losses and innovation costs, accelerating scientific development. DKDS ranged from a minimum of 0.006 (2005) to a maximum of 0.515 (2021), with an annual average of 0.040 (Figure 7 [Figure 7: see original paper]). Initial high attention was followed by slowed diffusion after 2003 as gene sequencing costs decreased and researchers preferred generating their own data. After 2014, as database content improved, researchers gradually returned to dataset reuse, reducing research costs and accelerating progress.

3.5 Data Knowledge Diffusion Delay

Diffusion delay reveals efficiency by measuring the time from dataset publication to first reuse. The maximum delay was 20 years, minimum 0 years, with an average of approximately 3.8 years. Comparing this to the average 4.3-year cycle from data processing to original publication shows that reusing data improves research efficiency by 13.16% on average. Figure 8 [Figure 8: see original paper] shows the distribution: 26.08% of datasets were reused within one year (the highest proportion), while original data papers peaked at a 2-year cycle (17.99%). For delays $ $4 years, reused-data papers accounted for 57% versus 50.58% for original-data papers, confirming that data reuse shortens publication cycles and accelerates diffusion.

---

4. Conclusion

This study analyzed GEO datasets and PMC full-text data to examine gene expression dataset sharing and reuse in biomedicine, proposing dataset-specific knowledge diffusion indicators. The findings reveal datasets' actual value in research, confirm that reuse improves efficiency, and quantitatively demonstrate datasets' importance in biomedical research:

1. Heavily reused GEO datasets concentrate in 2008-2017, aligning with the paradigm shift from knowledge-driven to data-driven science. Due to reuse lag, reused dataset counts show a growth-then-decline trend, demonstrating clear regularity consistent with the ~4-year diffusion delay.
2. From 2001-2021, annual diffusion breadth and intensity showed fluctuating growth before recent declines. Cumulative breadth and intensity exhibited S-curves, reflecting the time lag in dataset publication and reuse. Overall, datasets' short-term and long-term knowledge value in biomedicine is

increasing, with researchers using more datasets, reflecting biomedicine's data-driven nature.

3. GEO dataset diffusion speed experienced three phases: initial rapid uptake, stable period, and accelerated growth. Early datasets gained quick attention. After 2003, decreasing sequencing costs led researchers to generate their own data, slowing diffusion. After 2014, improved databases and accumulated datasets renewed researcher interest, especially with exponential growth post-2014.

4. Although China started late in scientific data sharing, initially concentrated in Hong Kong through international collaborations, China's position has strengthened with policies like the *Scientific Data Management Measures* and *CAS Scientific Data Management and Open Sharing Measures*. These efforts provide foundation and assurance for China to become a data powerhouse.

This study expands knowledge diffusion theory and methods by incorporating scientific datasets as a new knowledge entity, providing references for data management and services. Future directions include: (1) enriching disciplines and samples beyond biomedical GEO datasets; (2) analyzing relationships between diffusion characteristics and factors like contributors, institutions, impact factors, and regions; and (3) exploring diffusion features from network structure and collaboration perspectives.

References

[1] CHEN C M, HICKS D. Tracing knowledge diffusion[J]. Scientometrics, 2004, 59(2): 199-211.
[2] LEARNED W S. The American public library and the diffusion of knowledge[J]. Journal of the American Medical Association, 1924, 83(20): 1611-1611.
[3] Huang Lucheng, Liu Yumin, Wu Feifei, et al. A framework for studying technical knowledge diffusion characteristics based on complete patent citation information[J]. Science and Technology Management, 2017, 38(4): 149-161.
[4] Zhao Rongying, Wei Xuqiu. Exploring author knowledge diffusion patterns from a citation perspective[J]. Information Studies: Theory & Application, 2016, 39(8): 12-17.
[5] Yue Zenghui, Xu Haiyun. Research on knowledge diffusion characteristics in disciplinary citation networks[J]. Library and Information Service, 2015, 59(15): 119-126.
[6] Wang Jingjing, Ye Ying. Analysis of cross-disciplinary knowledge diffusion in international digital humanities research[J]. Journal of Library Science in China, 2015, 41(1): 62-75.
[7] LIU Y X, ROUSSEAU R. Knowledge diffusion through publications and citations: a case study using ESI-fields as unit of diffusion[J]. Journal of the American Society for Information Science and Technology, 2010, 61(2):

340-351.
[8] Yu Liping, Wan Xiaoyun, Xiang Yiming, et al. A new indicator for evaluating academic journal knowledge diffusion depth—the CJH index[J]. Journal of Intelligence, 2019, 38(8): 145-150.
[9] NAKAMURA H, SUZUKI S, HIRONORI T, et al. Citation lag analysis in supply chain research[J]. Scientometrics, 2011, 87(2): 221-232.
[10] Song Ge. Research on the diffusion process of academic innovation[J]. Journal of Library Science in China, 2015, 41(1): 62-75.
[11] KISS I Z, BROOM M, CRAZE P, et al. Can epidemic models describe the diffusion of topics across disciplines?[J]. Journal of Informetrics, 2010, 4(1): 74-82.
[12] GAO X, GUAN J C. Network model of knowledge diffusion[J]. Scientometrics, 2012, 90(3): 749-762.
[13] Wei Xuqiu, Guo Fengjiao, Yu Miao. Analysis of book knowledge diffusion characteristics from a micro perspective[J]. Information Science, 2021, 39(3): 37-43.
[14] Yu Xiaotong, Pan Xuelian, Hua Weina. Analysis of software citation and diffusion in knowledge graph research[J]. Information and Documentation Services, 2019, 40(2): 19-29.
[15] Zhang Lingling, Zhang Yu'e, Du Li. Knowledge diffusion research in library and information science from the perspective of National Social Science Fund projects[J]. Journal of Academic Libraries, 2021, 39(2): 45-51, 61.
[16] PARK H, YOU S, 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.
[17] Meng Xiangbao, Qian Peng. Research on characteristics of humanities and social sciences data from a data lifecycle perspective[J]. Library and Information Knowledge, 2017(1): 76-88.
[18] Ding Wenyao, Li Jian, Han Yi. Research on scientific data citation characteristics of papers in China's library and information science field[J]. Library and Information Service, 2019, 63(22): 118-128.
[19] Liu Yanan, Liu Jiangrong, Xiao Ming, et al. Research on data citation behaviors in funded project papers[J]. Library Tribune, 2019, 39(7): 75-83.
[20] ZHAO M N, YAN E J, LI K. Dataset mentions and citations: a content analysis of full-text publications[J]. Journal of the Association for Information Science and Technology, 2018, 69(1): 32-46.
[21] Yan Xiaoni, Tian Guoxiang, Guo Xiaojuan, et al. GEO database architecture, application and data extraction methods[J]. Chinese Journal of Evidence-Based Cardiovascular Medicine, 2019, 11(2): 134-137.
[22] Wang Xue, Yang Bo. Research on disciplinary differences in scientific data reuse[J]. Journal of Intelligence, 2021, 40(7): 122-126+156.
[23] Ruan Ji, Wang Yue, Liu Qian, et al. Implementation of PubMed Central citation data display on Chinese STM journal platforms[J]. Science-Technology & Publication, 2020(3): 125-128.
[24] ROWLANDS I. Journal diffusion factors: a new approach to measuring

research influence[J]. Aslib Proceedings, 2002, 54(2): 77-84.
[25] FRANDSEN T F, ROUSSEAU R, ROWLANDS I. Diffusion factors[J]. Journal of Documentation, 2006, 62(1): 58-72.
[26] Qiu Junping, Qu Hui, Luo Li. Bibliometric research on disciplinary knowledge diffusion based on journal citation relationships—taking Chinese "library, information and archives science" as an example[J]. Information Science, 2012, 30(4): 481-485, 491.
[27] Li Jiang. Review of citation-based knowledge diffusion research[J]. Information and Documentation Services, 2013(4): 36-40.
[28] Tang Yibing, Huang Zuqing, Zhang Baoyou. Research on knowledge diffusion and integration based on citation networks—taking supply chain research as an example[J]. Journal of Intelligence, 2012, 31(1): 119-122.
[29] OSIER M V, ZHAO H Y, CHEUNG K H. Handling multiple testing while interpreting microarrays with the gene ontology database[J]. BMC Bioinformatics, 2004, 5.
[30] GEO. Development stages of lentinus edodes[EB/OL]. [2021-11-12]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2167.

Author Contributions

Yang Ning: Data collection, experimental validation, and initial draft writing. Zhang Zhiqiang: Supervision, guidance on manuscript revision, and research conceptualization.

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