ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00417

An Empirical Study of Collaborative Relationships in Open Scientific Data Sharing: The Case of Peking University Open Research Data Platform (Postprint)

Authors: Zhang Hui, Cheng Yuqi, Wang Chuanqing

Date: 2023-04-01T16:02:59+00:00

Abstract

[Purpose/Significance] To reveal the utilization status of three representative datasets from the Peking University Open Research Data Platform, providing a reference for research on scientific data open sharing. [Method/Process] Papers from CNKI that utilize the CFPS, CHARLS, and CLHLS datasets were selected as research objects, and network analysis methods and relevant tools were employed to analyze collaborative research from the dimensions of authors, institutions, and topics. [Results/Conclusion] After 2011, the annual collaboration degree of each research entity tends to stabilize, and as the scope of research entities expands, the collaboration rate gradually decreases. Three centrality indicators complement each other to identify important institutions. First-level institutions present a scenario with Peking University as the core, some institutions maintaining cooperation among themselves, and numerous institutions actively participating in collaboration; second-level institutions ranking high in the number of collaborating institutions and total collaboration times are relatively stable, maintaining solid cooperative relationships among themselves, and can be roughly divided into three categories of institutions. The core author group demonstrates that some author teams collaborate frequently with stable cooperative relationships, and there exist intermediary authors who can connect several collaborative groups. Collaborative research papers involve rich themes, with key research topics including physical and mental health of the elderly, household consumption and asset conditions, and social security for the aging population. To promote the development of scientific data open sharing, future improvements can be made in aspects such as building an authoritative integrated platform for scientific data open sharing, integrating academic databases, and organizing academic competitions.

ChinaRxiv 中X

Full Text

Preamble

Volume 65, Issue 23, December 2021

An Empirical Study on Collaborative Relationships in Scientific Data Open Sharing: A Case Study of Peking University Open Research Data Platform

Zhang Hui[1], Cheng Yuqi[1], Wang Chuanqing[2]

[1]School of Library, Information and Archives, Shanghai University, Shanghai 200444
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/Significance] This study aims to reveal the utilization status of three representative datasets from Peking University Open Research Data Platform and provide references for research on scientific data open sharing. [Method/Process] We selected CNKI papers using CFPS, CHARLS, and CLHLS datasets as research objects and employed network analysis methods with relevant tools to analyze collaborative research from the dimensions of authors, institutions, and themes. [Result/Conclusion] After 2011, the annual collaboration degree of each research entity stabilized, while the collaboration rate gradually decreased as the research entity scope expanded. The three centrality indicators complement each other to identify important institutions. At the first-level institution tier, Peking University serves as the core, with some institutions maintaining collaborations and many actively participating. The second-level institutions with high numbers of collaborative partners and total collaborations remain relatively stable, maintaining solid cooperative relationships and roughly falling into three categories. The core author groups show that some author teams collaborate frequently with stable relationships, and intermediary authors connect several collaborative groups. Collaborative research papers cover rich themes, focusing on elderly physical and mental health, family consumption and assets, and social security for aging populations. To promote scientific data open sharing, future improvements can be made by building authoritative integrated platforms, integrating academic databases, and organizing academic competitions.

Keywords: open scientific data; CFPS; CHARLS; CLHLS; collaboration analysis

Classification Number: G203
DOI: 10.13266/j.issn.0252-3116.2021.23.003

In recent years, open science research has attracted widespread attention from the global scientific community [1], with open scientific data being a crucial component [2]. As early as November 3, 2015, China's 13th Five-Year Plan first proposed "implementing a national big data strategy and promoting open sharing of data resources" [3]. Subsequently, the "Administrative Measures for

Scientific Data" issued by the State Council established the principle that "openness is the norm, non-openness is the exception" for scientific data sharing [4]. Open scientific data has received increasing attention in China, with current academic research extensively exploring data governance [5-6], influencing factors [7], sharing policies [8-9], mechanism models [10-11], stakeholder roles and responsibilities [12-13], and sharing platforms [9] in scientific data open sharing. Meanwhile, numerous scientific data open sharing platforms have been established domestically and internationally [9], including 20 national scientific data centers under construction in China [14-15] and some university scientific data open platforms [16-17].

Among these, Peking University Open Research Data Platform serves as an exemplary model. As of April 28, 2021, the platform had collected 73 data spaces and 314 datasets, featuring a collection of high-quality survey projects with significant domestic influence, such as China Family Panel Studies (CFPS), China Health and Retirement Longitudinal Study (CHARLS), and Chinese Longitudinal Healthy Longevity Survey (CLHLS), now renamed Chinese Longitudinal Healthy Longevity Survey and Happy Family (CLHLS-HF). These three longitudinal survey projects began in 2010, 2011, and 1998 respectively, and have become standardized, large-scale, and periodic survey datasets that have attracted widespread attention and utilization [18]. Additionally, the "National University Data-Driven Innovation Research Competition" organized based on Peking University Open Research Data Platform has vigorously promoted the reuse of scientific data and academic output. Scientific data reuse can effectively avoid duplicate data collection, save funds, and improve efficiency, though it also requires considerable effort to digest existing information [19]. Compared with traditional research collaborations, cooperation based on scientific data reuse can actively promote the development of scientific data open sharing. Therefore, this study selects Peking University Open Research Data Platform to explore collaborative relationships in scientific data open sharing.

As of May 2021, two papers had introduced the construction of Peking University Open Research Data Platform [20-21], but no research has been found on the platform's data utilization. Therefore, based on the CFPS, CHARLS, and CLHLS datasets from Peking University Open Research Data Platform and papers using these data in CNKI, this study employs network analysis methods and relevant tools to conduct a preliminary analysis of author and institutional collaborations, aiming to reveal the utilization status of Peking University open research data and further promote the development of scientific data open sharing.

3 Collaborative Research Analysis

3.1 Data Preprocessing

To analyze collaborative research using open scientific data from Peking University from the dimensions of institutional and author collaboration, further

processing was conducted based on 1,493 journal papers.

3.1.1 Institutional Collaboration Data Institutional collaboration analysis includes both first-level and second-level institutions (non-university institutions involve only first-level institutions, such as the Chinese Center for Disease Control and Prevention). For universities, first-level institutions refer to universities (e.g., Peking University), while second-level institutions refer to faculties or schools (e.g., Peking University Health Science Center, Shanghai University School of Economics). Institutional data were obtained from official university websites. For each paper, institutional collaboration frequency was ignored, with only collaborative relationships considered. Among the 1,493 papers, first-level institutional collaboration research accounted for 549 papers (36.77%), involving 530 first-level institutions; second-level institutional collaboration research accounted for 562 papers (37.64%), involving 785 second-level institutions.

3.1.2 Author Collaboration Data Among the 1,493 papers, author collaboration research accounted for 1,159 papers (77.63%), involving 2,642 authors. To more clearly reveal author collaboration using platform data, this study applied Price's Law to identify the core author group [25]. The most prolific author was Professor Zhao Yaohui from the National School of Development at Peking University, with 18 papers. The core author group included authors with publication numbers $N \geq 0.749\sqrt{18}$ 3 papers, comprising 199 authors, which formed the basis for subsequent author collaboration analysis.

3.1.3 Annual Collaboration Degree and Rate Changes in annual collaboration degree and rate for different research entities are shown in Table 1 , with research entities being first-level institutions, second-level institutions, and authors. Table 1 shows that after 2011, the annual collaboration degree of each research entity stabilized, with an average of 3 authors per article, involving 2 or 3 first-level or second-level institutions. Additionally, the annual collaboration rates of different research entities show clear distinctions and connections. Distinctions appear in vertical comparisons, with collaboration rates from left to right approximately 30%-55%, 40%-65%, and 70%-80%. Connections appear in horizontal comparisons, where the collaboration rate gradually decreases as the research entity scope expands.

3.2 Institutional Collaboration Analysis

This section analyzes the top 10 institutions by centrality from both first-level and second-level perspectives regarding collaborative research using platform data.

3.2.1 Institutional Centrality The top 10 institutions by collaboration network centrality and their indicator values are shown in Table 2 . Degree centrality reflects the total number of collaborating institutions for a given institution.

Among 530 first-level institutions, Peking University collaborated with 106 institutions, showing the highest degree centrality and a significant lead, followed by Zhejiang University and Duke University. Among 785 second-level institutions, the two institutions with highest degree centrality both came from Peking University: the Economics and Management Division and the Health Science Center, though with substantially different absolute values, collaborating with 103 and 39 second-level institutions respectively, followed by Zhejiang University School of Medicine and the Chinese Center for Disease Control and Prevention.

Closeness centrality reflects how close an institution is to other institutions in the network. The top 10 first-level institutions by this metric share 6 institutions with degree centrality, with Peking University and Zhejiang University being the two universities closest to other institutions in the network. Among second-level institutions, the top 5 by closeness centrality differ from degree centrality only in that West China Hospital of Sichuan University replaced the Chinese Center for Disease Control and Prevention, with 4 new institutions added compared to the previous two indicators.

Betweenness centrality reflects the extent to which an institution serves as a "bridge" in the network. Table 2 shows that for first-level institutions, this metric differs significantly from the first two indicators in the top 10 ranking, adding 3 institutions: Chinese Academy of Social Sciences, Nanjing Agricultural University, and Fudan University. Among second-level institutions, Peking University Economics and Management Division and Health Science Center, and Zhejiang University School of Medicine remain in the top 3, with 4 new institutions added compared to the previous indicators.

These three centrality measures can all reflect important institutions in the network and complement each other. Combined results identify 17 first-level institutions and 25 second-level institutions as top 10 important institutions, which are further demonstrated in the following analysis.

3.2.2 First-Level Institutional Collaboration Research Figure 1 [Figure 1: see original paper] shows the largest connected component of first-level institutional collaboration, where node size reflects the total number of collaborating institutions. First-level institutions collaborated on 549 papers, involving 530 institutions, with 447 included in Figure 1, indicating that 84.34% of first-level institutions are interconnected. The largest node is "Peking University," indicating it has the most collaborating institutions, having collaborated 224 times with 106 other institutions, including 11 institutions with $ $5 collaborations. These 106 institutions involve 31 foreign universities, 47 domestic universities, and 28 non-university institutions, demonstrating Peking University's broad collaboration scope across domestic, foreign, and non-university institutions, while maintaining strong collaborative relationships with some universities.

Zhejiang University ranks second in total collaborating institutions, having collaborated 61 times with 42 first-level institutions, including 5 collaborations

with Peking University and $ $2 collaborations with 6 other institutions. Duke University, a foreign institution, collaborated 69 times with 36 first-level institutions, including 8 collaborations with Peking University and $ $2 collaborations with 10 other institutions. Among 530 first-level institutions, 1,067 institutional collaborations were formed, with 909 pairs collaborating only once (85.19%), indicating low collaboration intensity for most institutions. Overall, the first-level institutional collaboration network presents a pattern with Peking University as the core, some institutions maintaining collaborations, and many institutions actively participating.

3.2.3 Second-Level Institutional Collaboration Research Figure 2 [Figure 2: see original paper] shows the largest connected component of second-level institutional collaboration. Second-level institutional collaboration research includes 562 papers involving 785 institutions and 1,239 collaborations, with the largest connected component retaining 475 nodes, indicating that 60.51% of institutions are interconnected. The largest node is Peking University Economics and Management Division, with the most collaborating institutions, having collaborated 170 times with 103 other institutions. The top 3 institutions by collaboration frequency are Duke University School of Medicine (7 times), Peking University Health Science Center (6 times), and Chinese Center for Disease Control and Prevention (6 times). Peking University Health Science Center ranks second in total collaborating institutions, having collaborated 59 times with 39 institutions. Zhejiang University School of Medicine collaborated 30 times with 26 institutions, ranking third in node size and fifth in total collaborations. Chinese Center for Disease Control and Prevention collaborated 41 times with 23 institutions, ranking fourth in node size and third in total collaborations. Duke University School of Medicine collaborated 35 times with 19 institutions, ranking fifth in node size and fourth in total collaborations.

In the second-level institutional collaboration network, institutions ranking high in both number of collaborators and total collaborations remain relatively stable and maintain solid cooperative relationships. Overall, among 1,239 second-level institutional collaborations, 1,100 pairs collaborated only once (88.78%), while only 9 pairs collaborated $ $5 times, indicating that most second-level institutional collaborations were not sustained, with only a few maintaining relatively stable relationships. Additionally, participating second-level institutions can be roughly classified into three categories: (1) medical faculties/schools (including public health schools), (2) economics and management faculties/schools, and (3) sociology faculties/schools. This distribution relates to the dataset types: CFPS is mainly used for economics or sociology research, while CHARLS and CLHLS are primarily used for medical research.

3.3 Author Collaboration Analysis

The "core author group" collaboration network formed by 199 authors is shown in Figure 3 [Figure 3: see original paper], where node size reflects an author's

total collaboration frequency. Figure 3 contains 9 relatively obvious collaborative networks (labeled -), with networks , , and shown in enlarged views for later analysis. Network contains 15 authors, the largest group. According to original data, 4 authors in this network rank in the top 5 of the entire network in total collaborations: Shi Xiaoming from Chinese Center for Disease Control and Prevention (11 papers, 90 collaborations with 40 authors), Zeng Yi from Peking University Center for Healthy Aging and Development (10 papers, 88 collaborations with 45 authors), Yin Zhaoxue from Chinese Center for Disease Control and Prevention (7 papers, 65 collaborations with 23 authors), and Lü Yuebin from Chinese Center for Disease Control and Prevention (7 papers, 65 collaborations with 31 authors). Shi Xiaoming and Zeng Yi are each other's most frequent collaborators. Yin Zhaoxue and Lü Yuebin's most frequent collaborators are Shi Xiaoming and Zeng Yi/Shi Xiaoming respectively. All other authors in the network have >20 total collaborations, indicating active author collaboration and stable relationships among the most active authors.

Networks and each have $ $10 collaborating authors. Network contains 14 authors, including Zhao Yaohui from Peking University National School of Development, who collaborated 86 times with 39 authors, and J. Strauss from University of Southern California, who collaborated 62 times with 28 authors, being each other's most frequent collaborator. Network contains 11 authors, including Jing Huiquan from Capital Medical University School of Health Management and Education (44 collaborations with 34 authors) and Ding Hua from Peking University China Center for Social Research (38 collaborations with 30 authors), who are the most frequent collaborators in this network and have the highest mutual collaboration frequency. The remaining 6 networks contain 6-8 authors each, with networks - being connected components. Calculating author betweenness centrality reveals that Zeng Yi, Shi Xiaoming, Jing Huiquan, and Ding Hua are top-ranked intermediary authors. In summary, the core author collaboration network shows that some author groups collaborate frequently with stable relationships, and intermediary authors connect several collaborative groups.

3.4 Author Collaboration Analysis of Key Research Themes

Keywords from 1,159 author collaboration papers were processed through English-Chinese alignment, synonym merging, and removal of meaningless terms (e.g., CFPS), yielding 2,062 keywords. To clearly reveal key research themes, we selected a keyword network with occurrence frequency $ $9 (Figure 4 [Figure 4: see original paper]). VOSviewer automatically clustered 71 keywords into 7 categories, which can be summarized into three key research themes:

3.4.1 Elderly Physical and Mental Health Keywords include chronic diseases, diabetes, hypertension, depressive symptoms, mental health, cognitive impairment, life satisfaction, and happiness. This theme features 3 active author

groups: Group 1, represented by Changwei Li, published 8 papers on this theme with 21 collaborators, focusing on physical health issues among middle-aged and elderly populations such as hypertension, arthritis, and diabetes. Group 2, represented by Zeng Yi, published 7 papers with 38 collaborators, focusing on cognitive function in the elderly. Group 3, represented by Gu Danan, published 6 papers with 10 collaborators, focusing on healthcare issues affecting elderly health.

3.4.2 Family Consumption and Assets Major keywords include income distribution, consumption structure, commercial insurance, and land transfer, involving multidimensional poverty issues such as targeted poverty alleviation, poverty reduction effects, catastrophic health expenditure, income disparities, and urban-rural differences. No obvious author collaboration groups were found in this theme. A few authors collaborated on >1 paper, such as Wang Xiaoquan (3 papers on "family commercial insurance" with 4 collaborators), Zhang Qilin (3 papers on "family poverty" with 4 collaborators), and Qian Long (3 papers on "farmland transfer" with 4 collaborators), but most collaborations were single-occurrence.

3.4.3 Social Security for Aging Population Keywords include new rural pension insurance, medical insurance, pension insurance, and socioeconomic status. In this theme, Li Jianxin and Xia Cuicui maintain a stable collaborative relationship with 3 co-authored papers. Most other authors had only single collaborations; among 18 papers on "new rural pension insurance," only Wang Xiaozeng and Wang Linping had 2 collaborations, while the remaining 39 authors had single collaborations. Statistics show that most authors published only 1 paper, with an average of only 11.52% publishing >1 paper across all themes.

In summary, author collaboration research covers rich themes, with research groups focusing on "family" and "elderly people." Family research emphasizes financial assets and consumption, while elderly research focuses on physical/mental health and life security. Additionally, the vast majority of authors in key research themes published only once, with only a few active authors maintaining collaborations with some partners, though single-occurrence collaborations are more common.

4 Conclusions and Discussion

Based on the three datasets from Peking University Open Research Data Platform, this empirical study reveals:

1. After 2011, the annual collaboration degree stabilized, with an average of 3 authors per article involving 2-3 first-level or second-level institutions. The collaboration rate gradually decreases as research entity scope expands. The three centrality indicators all reflect important institutions in the

network and complement each other. The top 10 important institutions include 17 first-level and 25 second-level institutions.

2. The first-level institutional collaboration network shows Peking University as the core, with some institutions maintaining collaborations and many actively participating. In the second-level institutional collaboration network, institutions ranking high in collaboration numbers and total collaborations remain stable and maintain solid relationships, roughly classifiable into three categories: medical faculties/schools, economics/management faculties/schools, and sociology faculties/schools. The core author collaboration network shows some author groups collaborating frequently with stable relationships, and intermediary authors connecting several collaborative groups.

3. Author collaboration research covers rich themes, focusing on elderly physical and mental health, family consumption and assets, and social security for aging populations. Research groups emphasize "family" and "elderly people." Additionally, the vast majority of authors in key themes published only once, with only a few active authors maintaining collaborations.

This empirical study based on three datasets from Peking University Open Research Data Platform demonstrates three issues:

1. At both institutional and author levels, the collaborative utilization scope of existing open research data is relatively limited, with participating institutions dominated by Peking University and its divisions, and active authors such as Zeng Yi and Zhao Yaohui all from Peking University.

2. The vast majority of collaborative relationships were not sustained, with >85% of institutional and author pairs collaborating only once.

3. Collaborative research themes depend on the type and content of open scientific data, meaning collaboration based on scientific data reuse is data-dependent, which somewhat limits this form of collaboration.

To promote scientific data open sharing, we propose three recommendations:

1. **Unify open scientific data sources and build an authoritative integrated platform for scientific data open sharing.** Current domestic platforms have different focuses: national scientific data centers emphasize scientific/technical data, while the platform used in this study focuses on social science data. An authoritative integrated platform would ensure scientific data comprehensiveness and unity, expand user scope, and alleviate the limitations in participating institutions, authors, and research themes identified in this study. Research outputs based on this platform could also provide more comprehensive empirical analysis materials.

2. **Integrate academic databases into scientific data open sharing platform construction.** Incorporate relevant research outputs, themes,

active authors, and other information into the platform to provide convenient services for users.

3. **Organize large-scale academic competitions based on the integrated platform** to actively promote scientific data reuse and efficient utilization.

This study has limitations as it only analyzed collaborative research using three datasets from Peking University Open Research Data Platform. Future research could utilize all datasets from this platform and other open scientific datasets on a larger scale.

References

[1] UNESCO. UNESCO mobilizes 122 countries to promote open science and reinforced cooperation in the face of COVID-19 [EB/OL]. [2021-11-19]. https://en.unesco.org/news/unesco-mobilizes-122-countries-promote-open-science-and-reinforced-cooperation-face-covid-19.

[2] Huang Ruhua, Zhao Yang, Huang Yuting. International open science research progress [J]. Library and Information Service, 2021, 65(1): 140-149.

[3] People's Daily Online. Full text of CPC Central Committee's proposals for 13th Five-Year Plan [EB/OL]. [2021-11-19]. http://politics.people.com.cn/n/2015/1103/c1001-27772701.html.

[4] General Office of the State Council. Notice on issuing Administrative Measures for Scientific Data [EB/OL]. [2021-11-19]. http://www.gov.cn/zhengce/content_{5279272}.htm.

[5] Sheng Xiaoping, Guo Daosheng. Research on data security governance in scientific data open sharing [J]. Library and Information Service, 2020, 64(22): 11-24.

[6] Sheng Xiaoping, Tian Jing, Xiang Guilin. Research on data quality governance in scientific data open sharing [J]. Library and Information Service, 2020, 64(22): 25-36.

[7] Sheng Xiaoping, Yuan Yuan. Review of influencing factors in scientific data open sharing at home and abroad [J]. Information Studies: Theory & Application, 2021, 44(8): 173-179.

[8] Wen Fangfang. Research on foreign scientific data open sharing policies [J]. Library Science Research, 2017(9): 91-100.

[9] Sheng Xiaoping, Wu Tong. Review of scientific data open sharing research at home and abroad [J]. Library and Information Service, 2019, 63(17): 6-14.

[10] Li Chengzan, Zhang Lili, Hou Yanfei, et al. Scientific big data open sharing: models and mechanisms [J]. Information Studies: Theory & Application, 2017, 40(11): 45-51.

[11] KITCHIN R, COLLINS S, FROST D. Funding models for open access digital data repositories [J]. Online information review, 2015, 39(5): 664-681.

[12] WESSELS B, FINN RL, LINDE P, et al. Issues in the development of open access to research data [J]. Prometheus, 2014, 32(1): 49-66.

[13] Sheng Xiaoping, Wu Hong. Analysis of different stakeholders' motivations in scientific data open sharing activities [J]. Library and Information Service, 2019, 63(17): 40-50.

[14] Zhang Xianen. National scientific data sharing project [J]. Scientific Chinese, 2004(9): 11-13.

[15] National Science and Technology Infrastructure Center. National scientific data centers [EB/OL]. [2021-11-19]. https://www.escience.org.cn/data-center/.

[16] China Survey and Data Center, Renmin University of China. China General Social Survey [EB/OL]. [2021-11-19]. http://cgss.ruc.edu.cn/index.htm.

[17] China Survey and Data Center, Renmin University of China. Overview of China Longitudinal Aging Social Survey [EB/OL]. [2021-11-19]. http://class.ruc.edu.cn/xmjs/xmgk.htm.

[18] Peking University Open Research Data Platform. Introduction to Peking University Open Research Data Platform [EB/OL]. [2021-11-19]. https://opendata.pku.edu.cn/about.xhtml.

[19] Li Jialu. Research on influencing factors and promotion strategies of researchers' data reuse behavior [D]. Changchun: Northeast Normal University, 2019.

[20] Luo Pengcheng, Zhu Ling, Cui Haiyuan, et al. Construction of Peking University Open Research Data Platform based on Dataverse [J]. Library and Information Service, 2016, 60(3): 52-58.

[21] Zhu Ling, Nie Hua, Cui Haiyuan, et al. Construction of Peking University Open Research Data Platform: exploration and practice [J]. Library and Information Service, 2016, 60(4): 44-51.

[22] Zhao Rongying, Yu Bo. Analysis of collaboration patterns and impact factors of Altmetrics Top 100 papers [J]. Information Science, 2020, 38(4): 3-8.

[23] Wei Ruibin. Evolution analysis of research institution collaboration networks in scientometrics [J]. Intelligence Journal, 2012, 31(12): 40-45.

[24] Zhang Xue, Zhang Zhiqiang, Chen Xiujuan. Author collaboration characteristics based on journal papers and their impact on research output: taking highly productive authors in international medical informatics as an example [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(1): 29-37.

[25] Huang Lixia, Ji Sutong. Research on author collaboration relationships in domestic reading promotion field based on SNA [J]. Library and Information Service, 64(7): 119-126.

ChinaRxiv 中X

Author Contributions

Cheng Yuqi: Assisted with data processing
Zhang Hui: Responsible for topic selection, writing, and revision
Wang Chuanqing: Responsible for topic selection and revision

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

Source: ChinaXiv — Machine translation. Verify with original.