Information Research - Vol. 30 No. iConf (2025)

How do data authors perform in data-intensive research activities? Evidence from author contribution statement in data papers

Yang Heng , Yu Yonglin, and Liu Fenghong

DOI: https://doi.org/10.47989/ir30iConf47524

Abstract

Introduction. Despite the increasing prevalence of data-intensive scientific research, the division of labor in these activities and the performance of data authors remain underexplored. By employing the Contributor Roles Taxonomy (CRediT), this study examines the division of scientific labor in data papers from *Data in Brief.*

Method and analysis. Utilizing methods of mathematical statistics and data visualization, we analysed the connections between the 14 CRediT roles within data papers. We also explored the relationship between the distribution of labor and the size and discipline of the authorial team, as well as the associations between key authors and their respective CRediT roles.

Results. The results show that 1) data papers rarely make full use of the 14 CRediT roles to describe author contributions. 2) Team size and discipline have a significant impact on the labor division of data-intensive scientific research activities. 3) The need for data collection and analysis is the main reason for the expansion of team size, which is particularly evident in the natural sciences. 4) Corresponding authors and first authors continue to take on core roles. 5) Meanwhile, undertaking data analysis and processing-related tasks, such as '*Software*', helps authors advance in the author order of data papers.

Conclusion. This study provides insights into the division of labor in data-intensive scientific research and shows that CRediT has limitations in fully capturing the research workflow of data papers. We propose developing a taxonomy specific to data papers, such as DP - CRediT.

Introduction

Scientific data, as a crucial research output, has now widely gained recognition for its academic and economic value (Greenberg, Wu, et al., 2023; Pasquetto et al., 2017; Wilson et al., 2014). Under the impetus of open science, the forms of data publication are becoming increasingly diverse and the content is continually enriched (Landi et al., 2020; Wittenburg, 2021). As the infrastructure for data openness and sharing is progressively established and refined (Benhamed et al., 2023; Greenberg, McClellan, et al., 2023), the issue of how to incentivize researchers' willingness to open their data has become a new challenge for scientific data sharing in an open environment (Faniel & Jacobsen, 2010; Tenopir et al., 2011; Treadway et al., 2016).

In recent years, it has been recognized that appropriate and meaningful incentives are essential to capitalize on the promise of data sharing (Lo & DeMets, 2016) and that crediting data generators is key in this effort (Kalager et al., 2016). As stated by the International Council for Science (ICSU), '*Scientists should be recognized and given credit for the scientific contribution of the data sets that they produce as well as for the analysis of those data*' (ICSU). Consequently, Barbara E. et al. proposed the designation of '*data authors*' as an incentive for data sharing, with explicit identification in publications (Bierer et al., 2017). Notably, more than a decade ago, Jillian C.'s exploratory study found that for many participants, the term '*author*' in the context of data was not fitting, raising the question: according to scientific researchers, is data something that can be authored (Wallis & Borgman, 2011)? We believe that this shift in perception is related to the development of data publication and the rise of data papers over the past decade. In the dissemination of data papers, in order to be cited as a data author, a person must have made substantial contributions to the original acquisition, quality control, and curation of the data, be accountable for all aspects of the accuracy and integrity of the data provided, and ensure that the available data set follows FAIR Guiding Principles (Bierer et al., 2017). However, accurately defining data responsibilities to clarify the identity of data authors is a current challenge.

In data-intensive scientific research activities, establishing the identity of data authors requires that their contributions to the data are quantifiable and evaluable. With the proliferation of author contribution statements, various contributor role ontologies and taxonomies (CROTs) have emerged (Hosseini, Colomb, et al., 2023), offering standardized lists of roles or terms to designate individuals' contributions to research. Among these, CRediT (NISO), as a standardized method for describing author contributions, has been vigorously promoted by numerous journal publishers and widely adopted since its introduction in 2014. It is also gradually being applied to data papers to describe the contributions of authors in the production and publication process of scientific data.

In this research, we employ CRediT to examine the division of scientific labor in a sample of data papers from *Data in Brief*, exploring how research contributions are allocated in data-intensive scientific activities. More specifically, we first explore the intercorrelations among the 14 CRediT roles to assess the utilization of CRediT in data papers. We also consider the relationship between the division of scientific labor and the size of the author team, as well as the discipline of the research. Finally, we investigate the correlation between key authors in data papers, such as corresponding authors and first authors, and the CRediT roles they undertake.

Related work

Invisible labor in data-intensive science

As the fourth paradigm (Nielsen, 2009; Tolle et al., 2011), data-intensive science has been widely discussed in the academic community. Ramachandran et al. regard data-intensive science as a scientific discovery process that is driven by knowledge extracted from large volumes of data rather than the traditional hypothesis-driven discovery process, and they introduce the concept of '*data prospecting*' to address the challenges of data-intensive science (Ramachandran et al., 2013). Data prospecting requires more interdisciplinary collaboration, Cheruvelil and Soranno argue that data-intensive science will be most successful when used in combination with open science and team science (Cheruvelil & Soranno, 2018).

The division of labor in data-intensive science diverges significantly from traditional research models (Pietsch, 2015), primarily due to the sheer scale and complexity of data management and analysis. Unlike research activities that may be more individualistic or limited in scope, data-intensive science often requires collaborative efforts across multiple disciplines and relies heavily on technological infrastructures (Lenhardt et al., 2016; Schultes et al., 2022). This collaborative aspect introduces a new dimension to the scientific process, where the labor is not just intellectual, but also deeply intertwined with the technical and logistical support systems that enable data collection, processing, and interpretation at massive scales.

Scroggins et al. apply the concept of invisible labor to data-intensive science (Scroggins & Pasquetto, 2020). Drawing on a fifteen-year corpus of research into multiple domains of data-intensive science, they used a series of ethnographic vignettes to offer a snapshot of the varieties and valences of labor in data-intensive science. They finally pointed out that a full and nuanced understanding of data-intensive science can only be obtained by starting with the in-situ work and labor of scientific practice in all its manifold forms. Their work underscores a critical perspective: that the comprehensive grasp of data-intensive science is not merely about acknowledging the presence of invisible labor, but also about deciphering its unique characteristics when compared to other scientific endeavours.

Moreover, the importance of invisible labor in data-intensive science cannot be overstated. It is the often-uncredited work of data cleaning, metadata creation, and algorithm development that forms the backbone of robust scientific inquiry (Resnik et al., 2017; Shamoo, 2013). These tasks, while critical, are frequently overshadowed by the final published research outputs, which tend to be the metrics by which scientific success is commonly judged (Dance, 2012). The labor behind data maintenance, ensuring the integrity and accessibility of datasets, is equally vital, yet it remains largely invisible to those outside the immediate circle of data-intensive scientific practice.

In contemporary scientific research, data has become the central element propelling innovation and discovery. With the rapid evolution of big data technologies, data-intensive science has emerged as a prominent field, bringing revolutionary changes to various academic disciplines. This research methodology relies on the use of extensive datasets to drive experimentation, simulations, and analyses, simpler models with a lot of data supposedly trump more elaborate models with less data (Halevy et al., 2009), thereby enabling the acquisition of profound insights and knowledge. However, there is a paucity of research on data-intensive scientific activities. Approaching the understanding of data-intensive scientific activities from the perspective of division of labor and collaboration can better reveal the patterns of the fourth paradigm.

Evolution and adoption of CRediT

As the international community places increasing emphasis on the construction of research integrity, there is a growing call in the journal industry for the use of CRediT (Contribution Roles Taxonomy) to facilitate more transparent and granular descriptions of author contributions (Das & Das, 2020; Rahman & Verhagen, 2023; Udey, 2018). However, the forms and content of scientific research activities are evolving rapidly, the current taxonomy may need to evolve as science and the types of contributions that may become less or more important change (Allen et al., 2019). Therefore, for CORTs (Contributor Role Taxonomies), maintaining an up-to-date list of roles is essential to meet the evolving needs of users and is one of the factors that promote the adoption of CROTs (Vasilevsky et al., 2021). As emphasized by its developers, CRediT was initially designed as a contributor role taxonomy for life and physical sciences, and thus may not be suitable for all disciplines(Allen et al., 2019), more specific roles should be added to the CRediT, and a more granular lexicon of contribution elements should be established. (Steele et al., 2021)。

Holcombe compared the author roles reflected by CRediT with the authorship criteria provided by the ICMJE (International Committee of Medical Journal Editors), elucidating the significance of the CRediT role taxonomy (Holcombe, 2019). Some scholars have discussed the boundaries of CRediT usage and proposed modifications. For instance, Alpi and Akers noted that CRediT does not currently describe all the roles played by librarians (Alpi & Akers, 2021). Larivière et al. employed the CRediT, combined with data from PLOS journals, to study the division of scientific labor from aspects such as author order, gender, and contribution combinations, highlighting the need for increased attention to labor division across different disciplines and research teams (Lariviere et al., 2021). Ding et al. proposed a new method for co-author credit allocation based on CRediT, demonstrating through empirical analysis that this approach can effectively prevent credit inflation and reasonably reflect author contributions, particularly mitigating the impact of the number of co-authors on the first author's credit (Ding et al., 2021).

Numerous scholars have taken CRediT as a foundation and, in conjunction with domain research or specific scenarios, have empirically conducted certain additions, deletions, and modifications to enhance its applicability. Some studies have pointed out the deficiencies of CRediT in the field of Randomized Controlled Trials (RCTs), with some attributing these to improper use by authors, such as not distinguishing between manuscript editing and drafting the initial manuscript (Steele et al., 2021). Others have identified inherent design flaws in CRediT, leading to the proposal of CRediT-RCT for researchers in this field (Zhang et al., 2019). Matarese and Shashok argued that CRediT overlooks some non-author contributions, including technical support, translation, and manuscript editing (Matarese & Shashok, 2019). In response, they proposed specific recommendations to improve three roles within CRediT (Investigation; Writing - Original Draft; Writing - Review & Editing) and to add two new roles (Technical support; Translating or editing the manuscript, as non-author). Alliez et al. emphasized the importance of software development, proposing nine more refined categories to better represent the tasks involved (design, debugging, maintenance, coding, architecture, documentation, testing, support, and management) (Alliez et al., 2020). Fitzgerald et al. summarized the shortcomings of the CRediT in the social sciences and proposed improvements, suggesting 12 librarian author role classifications (literature synthesis, conceptualization, methodology, instruments, software, investigation, data curation, data analysis, interpretation, visualization, writing, editing) (Fitzgerald et al., 2020).

In summary, existing research provides empirical analyses of CRediT across various disciplines and application scenarios, along with recommendations for improvement. The data paper, as an emerging form of academic communication, represents a unique type of scholarly output generated from data-intensive research activities, and the description and attribution of author contributions within this context also warrant attention.

Dataset and methods

Data source

We selected data papers published in the Data in Brief journal as our sample data source. The reason for choosing this particular data source is that Data in Brief is a purely data-focused interdisciplinary journal where all papers published are descriptions of datasets and serve as records of data-intensive scientific research activities. Additionally, Data in Brief utilizes CRediT to describe authors' contributions, which provides us with a consistent data source for analysis.

We utilized the web scraping tool Web Scraper (Scraper) to crawl and collect information related to data papers of *Data in Brief*. The collected data included the disciplinary fields of the articles, titles, DOIs, author names, author order, corresponding authors, and CRediT Author Statements, from 10 volumes (V41-V50) of *Data in Brief* published between 2022 and 2023, totalling 1,724 articles. After excluding corrigendum, articles without author contribution statements, and those not fully utilizing the CRediT format, we ultimately included 1,513 data papers for analysis, which comprised 7,697 CRediT contribution entries.

Research route

Scholars have conducted extensive research on author task division using CRediT (Lariviere et al., 2021; Zhang et al., 2019). Building upon their research approaches, we have formulated the research route for this paper, as detailed below:

（1）We initially present some basic information about the dataset, then employ Kendall's coefficient to examine the correlation between CRediT roles and conduct a statistical analysis of the symmetry in the roles undertaken by authors.

（2）Then, we investigate the general division of labor characteristics of data-intensive scientific research activities from the perspective of the CRediT roles involved in data papers. Firstly, the usage of CRediT roles, includes the number of CRediT roles used in data papers and the frequency of use for each role. Second, we examined the variation in CRediT roles with the number of authors per article, ranging from single-authored works to collaborative papers with up to 16 authors. Third, we designed the CRediT Concentration Index (CCI) and utilized the Gini coefficient to investigate the preference for CRediT roles across different disciplinary fields. The formula for calculating CCI is as follows:

$$CCI = \frac{\mathrm{U}_{\mathrm{max\_role}}}{\sum \mathrm{U}_{\mathrm{role}}}$$

$\mathrm{U}_{\mathrm{max\_role}}$ represents the number of times the most frequently used role appears in a particular academic discipline, and $\mathrm{U}_{\mathrm{role}}$ represents the usage count of each role within that discipline. It should be noted that to prevent excessive bias in the data, the calculation of CCI excludes the relevant data for the two writing roles, '*Writing – original draft*' and '*Writing – review and editing*'.

（3）Finally, we examine the detailed division of labor characteristics of data-intensive scientific research activities by considering the CRediT roles undertaken by key authors. Descriptive statistics were conducted based on whether an author was the corresponding author and whether they were the first author (without considering joint first authorship), and the chi-square test was used to examine if the differences between groups were significant. The number of roles undertaken by each author was treated as skewed data, and the Wilcoxon rank-sum test was employed to assess the significance of the differences between groups. All 14 CRediT roles were coded as binary variables. A logistic regression model was employed to examine the influence of individual roles on the status of the corresponding author and the first author. Due to the right-skewed distribution of the data, a generalized linear model with a Gamma distribution and a log link function was constructed to assess the impact of individual roles on the author order status.

Results

Overview

Demography of the dataset

As shown in Table 1, the dataset of data papers utilized in this study spans 24 disciplines, including Agricultural Sciences and Arts and Humanities, as designated by Data in Brief. The number of data papers varies across each discipline, with the highest count in Biological Sciences, totalling 205 papers, and the lowest in Mathematics, with only 3 papers. There is also a significant variation in the average number of authors per paper, ranging from 1.5 to 7.3 authors per paper. However, the average number of CRediT roles per paper and per author across different disciplines is relatively stable, with most papers in the dataset involving an average of 6 to 10 CRediT roles per paper, and each author undertaking 3 to 5 CRediT roles.

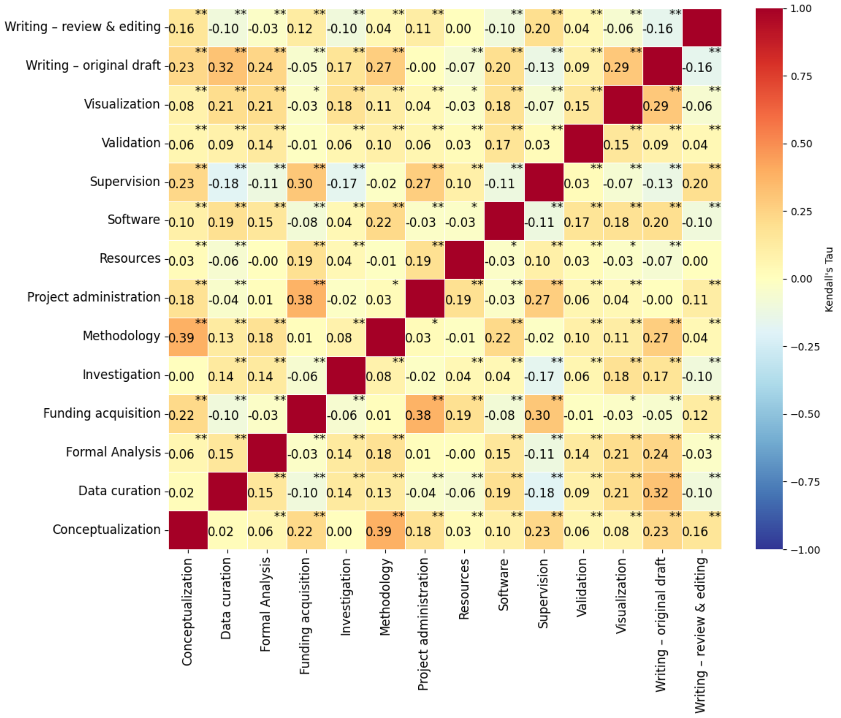
Disciplines	No. papers	Mean No. authors per paper	Mean No. roles per paper	Mean No. roles per author
Agricultural Sciences	164	5.0	8.5	3.3
Arts and Humanities	17	3.5	7.9	4.0
Biological Sciences	205	5.6	9.1	3.4
Business, Management, and Decision Sciences	64	3.3	8.5	4.1
Chemistry	58	5.2	8.9	3.1
Computer Science	166	4.5	9.3	3.7
Data Science	105	4.5	9.1	3.8
Earth and Planetary Sciences	90	6.0	9.2	3.5
Economics, Econometrics and Finance	34	2.9	8.7	4.9
Energy	43	5.6	9.2	3.4
Engineering	108	4.5	9.0	4.0
Environmental Science	100	5.5	8.8	3.3
Food Science	10	6.4	10.7	3.7
Health and Medical Sciences	103	7.3	9.2	3.4
Materials Science	56	4.6	9.2	3.5
Mathematics	3	2.3	11.7	7.3
Microbiology	25	5.8	8.9	3.1
Neuroscience	20	6.4	9.8	3.2
Omics	34	5.1	8.6	3.3
Pharmaceutical Sciences	6	7.2	7.5	2.4
Physical sciences	4	1.5	6.3	4.5
Plant Science	10	5.8	8.9	2.9
Psychology	30	4.3	8.1	3.8
Social Sciences	57	4.4	8.0	3.5

Table 1. Characteristics of CRediT roles from data papers in different disciplines

Correlation between CRediT roles

Figure 1 illustrates the correlations between various CRediT roles. Notably, the strongest correlation is observed between '*methodology*' and '*conceptualization*' (correlation coefficient = 0.39), followed by the correlation between '*funding acquisition*' and '*project administration*' (0.38), and the correlation between '*data curation*' and '*writing – original draft*' (0.32).

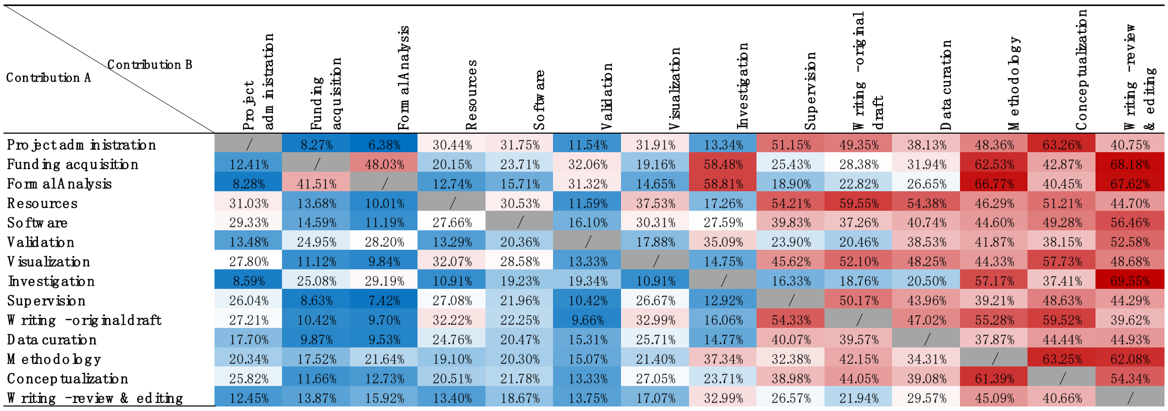
Figure 2 reflects the asymmetry in the associations between CRediT roles. More specifically, it shows the percentage of authors who have performed contribution A who have also performed contribution B (Lariviere et al., 2021). For instance, the figure shows that while 68.18% of authors who contributed to funding acquisition also reviewed and edited the manuscript, only 13.87% of authors who reviewed and edited the manuscript were involved in funding acquisition, representing the most asymmetric relationship. Following this is the asymmetry between theoretical and practical work, such as the observation that a significant portion of authors who undertook hands-on roles like '*formal analysis*' '*software*' '*validation*' and '*visualization*' also took on theoretical roles such as '*conceptualization*' and '*methodology*'. However, among the group of authors engaged in theoretical work, a relatively small fraction participated in the aforementioned practical tasks.



**. At the 0.01 level (two-tailed), the correlation is significant.

*. At the 0.05 level (two-tailed), the correlation is significant.

Figure 1. Heatmap showing the pairwise correlation of author roles defined in the CRediT



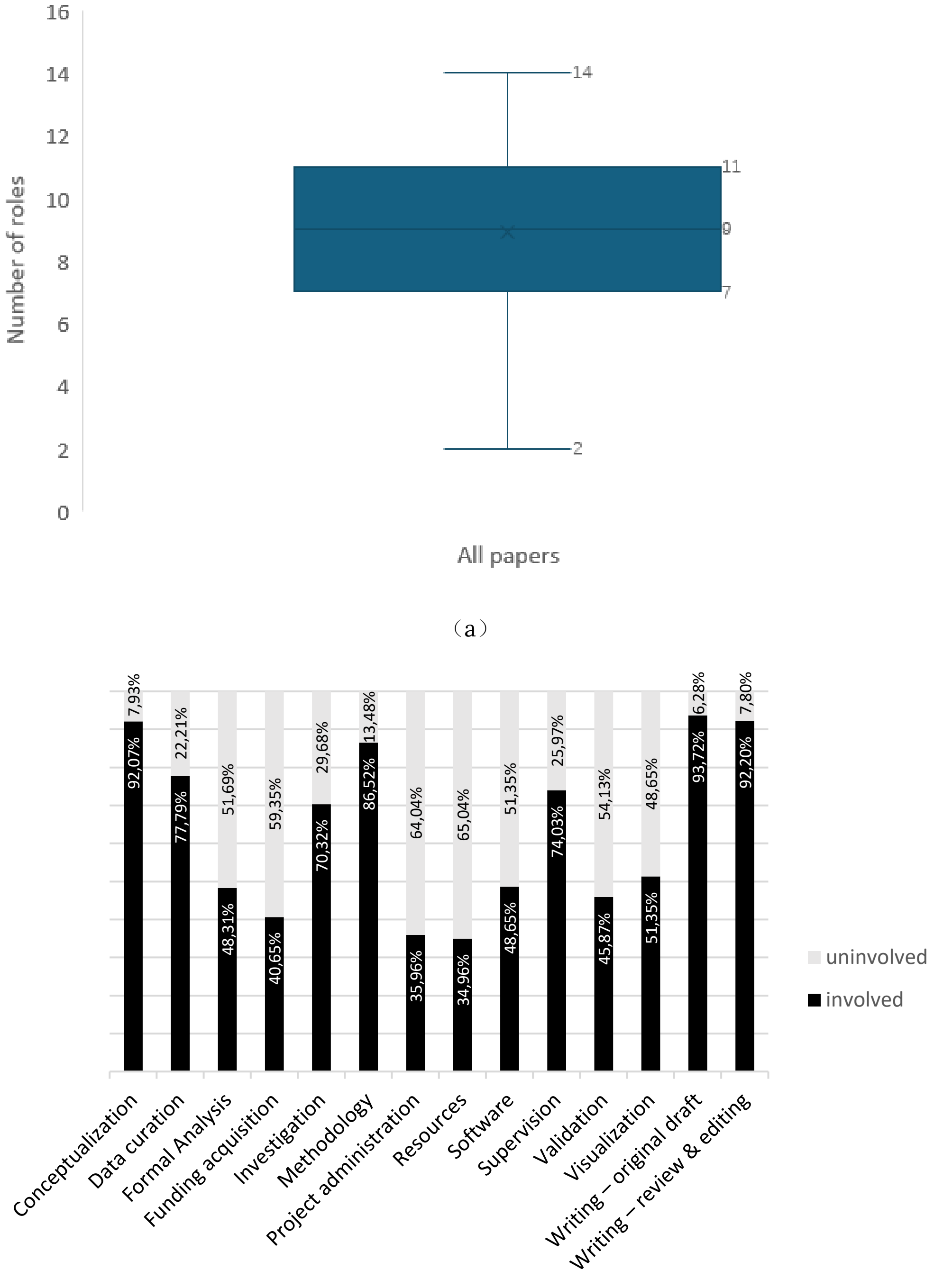
Contribution A \ Contribution B	Project administration	Funding acquisition	Formal Analysis	Resources	Software	Validation	Visualization	Investigation	Supervision	Writing -original draft	Data curation	Methodology	Conceptualization	Writing -review & editing
Project administration	/	8.27%	6.38%	30.44%	31.75%	11.54%	31.91%	13.34%	51.15%	49.35%	38.13%	48.36%	63.26%	40.75%
Funding acquisition	12.41%	/	48.03%	20.15%	23.71%	32.06%	19.16%	58.48%	25.43%	28.38%	31.94%	62.53%	42.87%	68.18%
Formal Analysis	8.28%	41.51%	/	12.74%	15.71%	31.32%	14.65%	58.81%	18.90%	22.82%	26.65%	66.77%	40.45%	67.62%
Resources	31.03%	13.68%	10.01%	/	30.53%	11.59%	37.53%	17.26%	54.21%	59.55%	54.38%	46.29%	51.21%	44.70%
Software	29.33%	14.59%	11.19%	27.66%	/	16.10%	30.31%	27.59%	39.83%	37.26%	40.74%	44.60%	49.28%	56.46%
Validation	13.48%	24.95%	28.20%	13.29%	20.36%	/	17.88%	35.09%	23.90%	20.46%	38.53%	41.87%	38.15%	52.58%
Visualization	27.80%	11.12%	9.84%	32.07%	28.58%	13.33%	/	14.75%	45.62%	52.10%	48.25%	44.33%	57.73%	48.68%
Investigation	8.59%	25.08%	29.19%	10.91%	19.23%	19.34%	10.91%	/	16.33%	18.76%	20.50%	57.17%	37.41%	69.55%
Supervision	26.04%	8.63%	7.42%	27.08%	21.96%	10.42%	26.67%	12.92%	/	50.17%	43.96%	39.21%	48.63%	44.29%
Writing -original draft	27.21%	10.42%	9.70%	32.22%	22.25%	9.66%	32.99%	16.06%	54.33%	/	47.02%	55.28%	59.52%	39.62%
Data curation	17.70%	9.87%	9.53%	24.76%	20.47%	15.31%	25.71%	14.77%	40.07%	39.57%	/	37.87%	44.44%	44.93%
Methodology	20.34%	17.52%	21.64%	19.10%	20.30%	15.07%	21.40%	37.34%	32.38%	42.15%	34.31%	/	63.25%	62.08%
Conceptualization	25.82%	11.66%	12.73%	20.51%	21.78%	13.33%	27.05%	23.71%	38.98%	44.05%	39.08%	61.39%	/	54.34%
Writing -review & editing	12.45%	13.87%	15.92%	13.40%	18.67%	13.75%	17.07%	32.99%	26.57%	21.94%	29.57%	45.09%	40.66%	/

Figure 2. Percentage of authors who have performed contribution A who also have performed contribution B

CRediT roles involved in data papers

Usage of CRediT roles

Figure 3(a) presents the distribution of the number of CRediT roles involved across all articles. It can be observed that the majority of articles utilize approximately 9 CRediT roles to describe the research work. Figure 3(b) reflects the percentage of papers employing a particular CRediT role out of all data papers. It is evident that the usage of various roles in data papers is not uniform, the most frequently used roles are '*writing – original draft*', '*writing – review & editing*', and '*conceptualization*', all with a prevalence of over 90%. Following these are '*methodology*', which is associated with 86.52% of the papers, and roles such as '*data curation*', '*supervision*', and ‘*investigation*’, which are included in 77.79%, 74.03%, and 70.32% of the data papers, respectively. It is also noticeable that ‘*project administration*’ and ‘*resources*’ are less commonly utilized in data papers.



（a）

（b）

Figure 3. Statistical Overview of CRediT Role Usage

The division of labor varies with the number of authors

The division of labor in scientific research varies with the number of authors involved. Figure 4 presents the percentage of authors who have performed a given task, for papers between 1 and 16 authors (N=7296 papers, 94.8% of the data set). We assume that the sole author of a paper undertakes all 14 CRediT roles, and as the number of authors increases, the distribution of different tasks among authors begins to shift. The changes can be broadly categorized into three types: one type of role is consistently carried out by a smaller proportion of the team (dashed line), such as '*project administration*', '*funding acquisition*', and '*resources*', which are consistently undertaken by about one-fifth of the total team members; a second type of role is consistently undertaken by a larger proportion of team members (thick line), such as '*writing – review and editing*' and '*investigation*', which are consistently handled by about one-third to one-half of the authors; and a third type of role sees a significant decrease in the proportion of authors undertaking it as the number of authors increases (solid line), such as '*conceptualization*', '*methodology*', and '*writing – original draft*', which are initially undertaken by half or more of the authors when the number of authors is low, but drop to being handled by only about one-fifth of the authors when the team size exceeds 10.

It is noteworthy that as the size of the author team increases, particularly when it reaches more than 10 members, the proportion of authors assuming CRediT roles begins to fluctuate. However, roles associated with data collection or analysis tend to emerge with disproportionately high levels of involvement, such as '*validation*' and '*formal analysis*'. Notably, the role of '*investigation*' stands out; when the number of authors exceeds 11, the proportion of authors undertaking the '*investigation*' role surpasses that of roles like '*conceptualization*', '*methodology*', and '*data curation*'. Alongside '*writing – review and editing*', '*investigation*' becomes one of the roles with the highest proportion of authors involved.

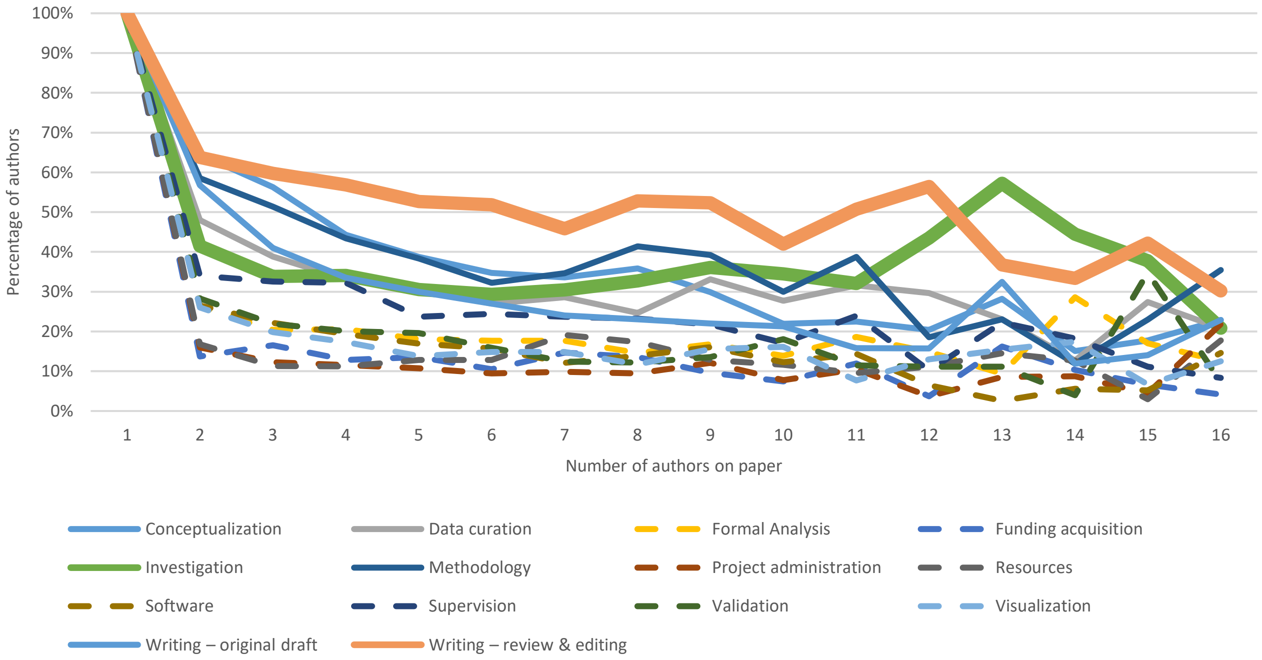


Figure 4. Statistical overview of CRediT role variation with the number of authors

CRediT role concentration across different disciplines

Disciplinary differences exist in the concentration of CRediT roles, and Figure 5 reflects the concentration of CRediT roles involved in articles from the top 13 most represented fields in the sample dataset (N=6767, 87.9%).

Overall, the concentration of CRediT roles in data papers across various academic disciplines fluctuates between 10% and 20%. The highest concentration is observed in the social sciences field, at 19.18%, while the lowest is in the computer science field, at 13.44%. Relatively speaking, social science fields, such as social sciences, business, management, and decision sciences, exhibit significantly higher CRediT role concentration compared to some natural science fields, like biological sciences and energy. The Gini coefficient, a measure of inequality, further highlights the disparities in CRediT role distribution across disciplines. Higher values of the Gini coefficient indicate greater inequality in role distribution, which aligns with the observed higher concentration in social science fields (0.31). This metric underscores the need for a more balanced distribution of CRediT roles across all academic disciplines.

Furthermore, the specific CRediT roles that authors concentrate on vary by discipline. Upon closer examination, with the exception of the '*writing – review and editing*' role, authors in social science fields predominantly focus on the '*conceptualization*' role, whereas authors in natural science fields are more inclined toward the '*investigation*' and '*methodology*' roles (Figure 6).

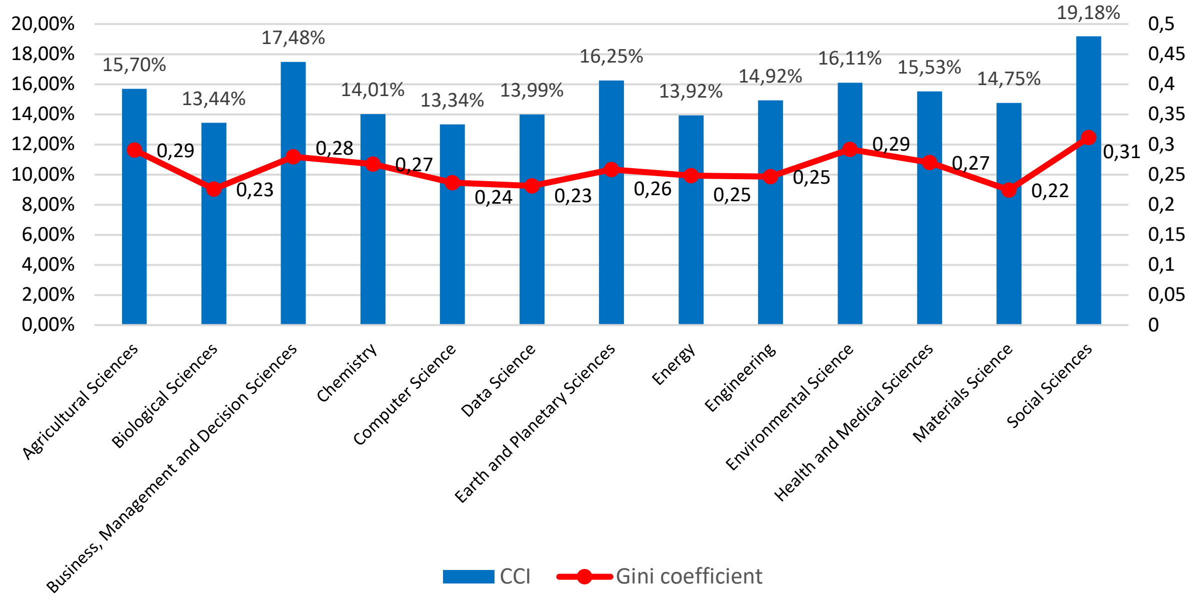
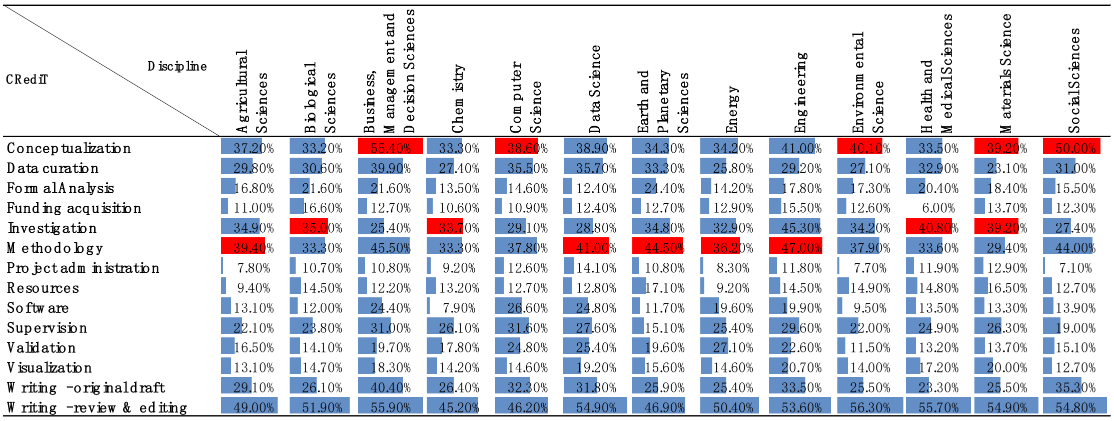


Figure 5. CRediT Concentration Index (CCI) and Gini coefficient across different disciplines



Discipline CRediT	Agricultural Sciences	Biological Sciences	Business, Management and Decision Sciences	Chemistry	Computer Science	Data Science	Earth and Planetary Sciences	Energy	Engineering	Environmental Science	Health and Medical Sciences	Materials Science	Social Sciences
Conceptualization	37.20%	33.20%	55.40%	33.30%	38.60%	38.90%	34.30%	34.20%	41.00%	40.10%	33.50%	39.20%	50.00%
Data curation	29.80%	30.60%	39.90%	27.40%	35.50%	35.70%	33.30%	25.80%	29.20%	27.10%	32.90%	23.10%	31.00%
Formal Analysis	16.80%	21.60%	21.60%	13.50%	14.60%	12.40%	24.40%	14.20%	17.80%	17.30%	20.40%	18.40%	15.50%
Funding acquisition	11.00%	16.60%	12.70%	10.60%	10.90%	12.40%	12.70%	12.90%	15.50%	12.60%	6.00%	13.70%	12.30%
Investigation	34.90%	35.00%	25.40%	33.70%	29.10%	28.80%	34.80%	32.90%	45.30%	34.20%	40.80%	39.20%	27.40%
Methodology	39.40%	33.30%	45.50%	33.30%	37.80%	41.00%	44.50%	36.20%	47.00%	37.90%	33.60%	29.40%	44.00%
Project administration	7.80%	10.70%	10.80%	9.20%	12.60%	14.10%	10.80%	8.30%	11.80%	7.70%	11.90%	12.90%	7.10%
Resources	9.40%	14.50%	12.20%	13.20%	12.70%	12.80%	17.10%	9.20%	14.50%	14.90%	14.80%	16.50%	12.70%
Software	13.10%	12.00%	24.40%	7.90%	26.60%	24.80%	11.70%	19.60%	19.90%	9.50%	13.50%	13.30%	13.90%
Supervision	22.10%	23.80%	31.00%	26.10%	31.60%	27.60%	15.10%	25.40%	29.60%	22.00%	24.90%	26.30%	19.00%
Validation	16.50%	14.10%	19.70%	17.80%	24.80%	25.40%	19.60%	27.10%	22.60%	11.50%	13.20%	13.70%	15.10%
Visualization	13.10%	14.70%	18.30%	14.20%	14.60%	19.20%	15.60%	14.60%	20.70%	14.00%	17.20%	20.00%	12.70%
Writing - original draft	29.10%	26.10%	40.40%	26.40%	32.30%	31.80%	25.90%	25.40%	33.50%	25.50%	23.30%	25.50%	35.30%
Writing - review & editing	49.00%	51.90%	55.90%	45.20%	46.20%	54.90%	46.90%	50.40%	53.60%	56.30%	55.70%	54.90%	54.80%

Figure 6. Percentage of authors per CRediT Role across different disciplines

CRediT roles undertaken by key authors

Factors associated with the corresponding author

A statistical analysis of 7967 author contributions across 1513 articles revealed that (Table 2), in general, apart from writing roles, authors were most involved in '*methodology*' (38.9%) and '*conceptualization*' (37.8%), while a smaller proportion of authors participated in "Project administration" (10.6%). The median number of roles per author was 3 (IQR: 2-5).

For corresponding authors, the proportion of those assuming the '*conceptualization*' role is significantly higher compared to other roles, accounting for 65.0%, and they are more likely to take on any of the 14 CRediT roles relative to non-corresponding authors. It is noteworthy that the majority of corresponding authors are also first authors, with 899 out of 1958 (45.9%). Corresponding authors take on a greater number of roles than non-corresponding authors [5 (3-6) vs. 3 (2-4); P<0.001], and they are ranked higher in author order [2 (1-4) vs. 4 (2-6); P < 0.001]. Additionally, there is no significant difference in involvement in the '*investigation*' and '*resources*' roles based on whether an author is a corresponding author or not (P>0.1).

Variables	Overall（n=7697）	Non-corresponding author (n=5739)	Corresponding author (n=1958)	P
Conceptualization, n (%)	2906 (37.8)	1633 (28.5)	1273 (65.0)	<0.001
Data_curation, n (%)	2400 (31.2)	1623 (28.3)	777 (39.7)	<0.001
Formal_Analysis, n (%)	1403 (18.2)	917 (16.0)	486 (24.8)	<0.001
Funding_acquisition, n (%)	942 (12.2)	531 (9.3)	411 (21.0)	<0.001
Investigation, n (%)	2633 (34.2)	1937 (33.8)	696 (35.5)	0.156
Methodology, n (%)	2994 (38.9)	1932 (33.7)	1062 (54.2)	<0.001
Project_administration, n (%)	814 (10.6)	479 (8.3)	335 (17.1)	<0.001
Resources, n (%)	1046 (13.6)	762 (13.3)	284 (14.5)	0.184
Software, n (%)	1222 (15.9)	799 (13.9)	423 (21.6)	<0.001
Supervision, n (%)	1898 (24.7)	1199 (20.9)	699 (35.7)	<0.001
Validation, n (%)	1323 (17.2)	909 (15.8)	414 (21.1)	<0.001
Visualization, n (%)	1199 (15.6)	720 (12.5)	479 (24.5)	<0.001
Writing-original draft, n (%)	2216 (28.8)	1174 (20.5)	1042 (53.2)	<0.001
Writing-review & editing, n (%)	4001 (52.0)	2886 (50.3)	1115 (56.9)	<0.001
First_author, n (%)	1513 (19.7)	614 (10.7)	899 (45.9)	<0.001
roles_number, median (IQR)	3 (2, 5)	3 (2, 4)	5 (3, 6)	<0.001
Author_order, median (IQR)	3 (2, 5)	4 (2, 6)	2 (1, 4)	<0.001

IQR, interquartile range.

Table 2. Comparison of CRediT roles between corresponding author and other co-authors

The logistic regression model indicates that authors who perform '*conceptualization*' are twice as likely to be corresponding authors (OR: 2.33; 95% CI: 2.04-2.67; P<0.001). Furthermore, authors engaged in '*funding acquisition*' (OR: 1.91; 95% CI: 1.60-2.29; P<0.001), '*supervision*' (OR: 2.26; 95% CI: 1.95-2.61; P<0.001), '*writing - original draft*' (OR: 2.37; 95% CI: 2.03-2.77; P<0.001), and those serving as first authors (OR: 4.65; 95% CI: 3.94-5.50; P<0.001) are more likely to be corresponding authors (Table 3).

	OR (95% CI)	P
Conceptualization	2.33 (2.04, 2.67)	<0.001
Data_curation	1.12 (0.97, 1.29)	0.115
Formal_Analysis	1.00 (0.85, 1.17)	1
Funding_acquisition	1.91 (1.60, 2.29)	<0.001
Investigation	0.75 (0.66, 0.86)	<0.001
Methodology	0.95 (0.83, 1.09)	0.461
Project_administration	1.19 (0.98, 1.45)	0.078
Resources	1.05 (0.87, 1.26)	0.599
Software	1.04 (0.87, 1.23)	0.68
Supervision	2.26 (1.95, 2.61)	<0.001
Validation	1.02 (0.87, 1.20)	0.819
Visualization	1.23 (1.03, 1.45)	0.018
Writing-original draft, n (%)	2.37 (2.03, 2.77)	<0.001
Writing-review & editing, n (%)	1.37 (1.21, 1.56)	<0.001
First_author	4.65 (3.94, 5.50)	<0.001

OR odds ratio; CI, confidence interval.

Table 3. Logistic regression model investigating the factors associated with the corresponding author role

Factors associated with the first author

Due to a significant overlap between the first authors and corresponding authors, the roles undertaken by the first authors are similar to those of the corresponding authors. However, there are subtle differences. There are no significant differences between first authors and non-first authors in the roles of '*funding acquisition*' and '*project administration*' (Table 4).

Variables	Overall（n=7697）	Non-first author (n=6184)	First author (n=1513)	P
Conceptualization, n (%)	2906 (37.8)	1910 (30.9)	996 (65.8)	<0.001
Data_curation, n (%)	2400 (31.2)	1541 (24.9)	859 (56.8)	<0.001
Formal_Analysis, n (%)	1403 (18.2)	839 (13.6)	564 (37.3)	<0.001
Funding_acquisition, n (%)	942 (12.2)	795 (12.9)	147 (9.7)	0.001
Investigation, n (%)	2633 (34.2)	1863 (30.1)	770 (50.9)	<0.001
Methodology, n (%)	2994 (38.9)	1949 (31.5)	1045 (69.1)	<0.001
Project_administration, n (%)	814 (10.6)	660 (10.7)	154 (10.2)	0.607
Resources, n (%)	1046 (13.6)	913 (14.8)	133 (8.8)	<0.001
Software, n (%)	1222 (15.9)	728 (11.8)	494 (32.7)	<0.001
Supervision, n (%)	1898 (24.7)	1686 (27.3)	212 (14.0)	<0.001
Validation, n (%)	1323 (17.2)	968 (15.7)	355 (23.5)	<0.001
Visualization, n (%)	1199 (15.6)	658 (10.6)	541 (35.8)	<0.001
Writing-original draft, n (%)	2216 (28.8)	980 (15.8)	1236 (81.7)	<0.001
Writing-review & editing, n (%)	4001 (52.0)	3379 (54.6)	622 (41.1)	<0.001
Author_corresponding, n (%)	1958 (25.4)	1059 (17.1)	899 (59.4)	<0.001
roles_number, median (IQR)	3 (2, 5)	3 (2, 4)	5 (4, 7)	<0.001

IQR, interquartile range.

Table 4. Comparison of CRediT roles between the first author and other co-authors

The logistic regression model indicates a significant correlation between the role of '*writing - original draft*' and being designated as the first author (OR: 11.23; 95% CI: 9.56-13.23; P<0.001). However, first authors are seldom involved in '*writing - review and editing*' (OR: 0.71; 95% CI: 0.61-0.84; P<0.001), '*supervision*' (OR: 0.53; 95% CI: 0.42-0.66; P<0.001), or '*resources*' (OR: 0.57; 95% CI: 0.43-0.74; P<0.001) (Table 5).

	OR (95% CI)	P
Conceptualization	3.50 (2.94, 4.17)	<0.001
Data_curation	1.60 (1.36, 1.87)	<0.001
Formal_Analysis	1.72 (1.44, 2.05)	<0.001
Funding_acquisition	0.86 (0.65, 1.14)	0.304
Investigation	1.49 (1.27, 1.74)	<0.001
Methodology	1.88 (1.59, 2.22)	<0.001
Project_administration	0.89 (0.66, 1.19)	0.423
Resources	0.57 (0.43, 0.74)	<0.001
Software	1.60 (1.33, 1.93)	<0.001
Supervision	0.53 (0.42, 0.66)	<0.001
Validation	0.87 (0.71, 1.06)	0.179
Visualization	1.79 (1.48, 2.15)	<0.001
Writing-original draft, n (%)	11.23 (9.56, 13.23)	<0.001
Writing-review & editing, n (%)	0.71 (0.61, 0.84)	<0.001

OR odds ratio; CI, confidence interval.

Table 5. Logistic regression model investigating the factors associated with the first author designation

Factors associated with the author order

Figure 7 revealed a significant correlation between the position of each author in an article and the number of roles they undertake ($R^2=0.073$, $P<0.001$). The results of the generalized linear model analysis (Table 6) indicate a significant correlation between '*conceptualization*', '*data curation*', '*software*', and '*visualization*' with the order of authorship. For instance, authors who contributed to the '*software*' are likely to be ranked higher compared to those who did not participate in this role, with a coefficient of 0.28 (95% CI: 0.22-0.34).

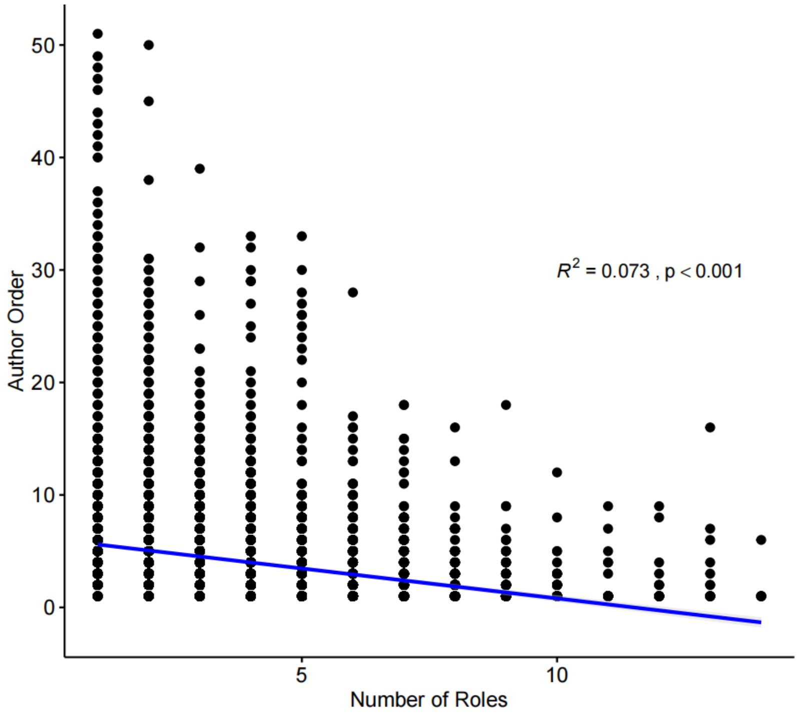


Figure 7. Scatter plot showing the correlation between the author order and the number of roles per author. There was a significant correlation (R^2 =0.073; P<0.001) between the two variables.

	Coefficient (95% CI)	P
Conceptualization	-0.23 (-0.28, -0.18)	<0.001
Data_curation	-0.15 (-0.19, -0.10)	<0.001
Formal_Analysis	-0.10 (-0.16, -0.04)	<0.05
Funding_acquisition	0.07 (-0.00, 0.14)	0.059
Investigation	-0.07 (-0.11, -0.03)	<0.05
Methodology	-0.08 (-0.12, -0.03)	<0.05
Project_administration	0.04 (-0.03, 0.12)	0.26
Resources	0.09 (0.03, 0.15)	<0.05
Software	**-0.28 (-0.34, -0.22)**	**<0.001**
Supervision	0.04 (-0.01, 0.09)	0.142
Validation	-0.04 (-0.09, 0.02)	0.193
Visualization	-0.16 (-0.22, -0.10)	<0.001
Writing-original draft	-0.59 (-0.64, -0.54)	<0.001
Writing-review & editing	-0.10 (-0.15, -0.06)	<0.001

Table 6. Generalized linear model investigating the factors associated with the author order

Discussion

This study employs CRediT to explore the division of labor in data-intensive scientific research activities, represented by data papers. According to our result, the pattern of labor division in data articles is different from traditional research papers in many ways. For instance, in data papers, the core contribution of '*conceptualization*' is most strongly correlated with the contribution of '*methodology*', while it is '*funding acquisition*' with research articles in the RCT field (Zhang et al., 2019). However, the roles that play a significant part in data paper collaborations show no significant difference from those in research articles: almost all articles involve roles such as '*conceptualization*', '*writing – original draft*', and '*writing – review and editing*' (Lariviere et al., 2021). Furthermore, some roles exhibit high autocorrelation and lack distinct differentiation, suggesting that they could be considered for consolidation within the context of contributor roles in data papers.

Data-intensive science is a collaborative endeavour, and our research indicates that as the number of co-authors increases, there is a corresponding rise in the number of individuals taking on roles such as investigation, validation, and formal analysis. In contrast, the number of individuals assuming core roles like conceptualization, methodology, and Writing – original draft, or managerial roles such as project administration, funding acquisition, and resources, remains relatively stable. This suggests that the primary driver of team size expansion is the need for practical tasks such as data handling and investigation, while the number of scientific leaders remains scarce (Robinson-Garcia et al., 2020). As Larivière et al. have posited, '*the bureaucratization of science can be considered as an inevitable consequence of the ubiquity of collaborative science*' (Larivière et al., 2015).

The publication of data papers reflects the work patterns and research cultures across various disciplines. Different academic fields exhibit varying degrees of concentration in the use of CRediT roles, with data papers in the social sciences generally showing higher concentrations of CRediT role usage compared to those in the natural sciences. We suggest that the differences in research methodologies, types of data, and the nature of research questions are the primary causes of this phenomenon. For instance, data papers in the social sciences more prominently represent the '*paper*' dimension, with authors tending to focus on the '*conceptualization*' role, while data papers in the natural sciences emphasize the '*data*' dimension more, with authors concentrating more on '*investigation*' and '*methodology*'.

The attribution of academic credit is one of the key concerns for researchers, with a common understanding that corresponding authors and first authors make significant contributions and play major leadership roles (Bhandari et al., 2014; Perneger et al., 2017; Teixeira da Silva, 2021; Yang et al., 2017). The statistical analysis indicates that in data papers, corresponding authors often undertake more leadership and coordination tasks and are more likely to be involved in key roles such as '*conceptualization*', '*funding acquisition*', and '*supervision*'. These roles pertain not only to the initial design and theoretical construction of the research but also encompass the supervision of the research process and financial support. The first authors, on the other hand, exhibit higher engagement in roles like '*writing - original draft*' and '*data curation*', which are directly related to the implementation of the research and the accuracy of the data. Comparatively, corresponding authors are more likely than first authors to take on any given role, implying that they act as versatile players within the team (Lu et al., 2020) and are more likely to be the corresponding authors of a data paper. Furthermore, our research also reveals that the more roles an author takes on, and their involvement in key data processing roles such as '*data curation*', '*software*', and '*visualization*', the more it aids in the author's ranking in terms of by-line order.

Our research indicates that some technical contributions related to data processing are clearly very important for the attribution of credit to authors of data papers. However, in the CRediT, the connotations of roles such as '*data curation*' encompass both work requiring profound professional background knowledge, such as data annotation, and technical tasks like data cleaning and maintenance. The roles of technical contributions (Smith, 2023) and human intervention are not adequately differentiated and described. This could lead to a situation where some authors' significant contributions are not recognized as they should be, while others enjoy excessive credit due to their role designation. This situation could be more pronounced in data papers, ultimately affecting the fairness of academic credit attribution for data papers.

CRediT, as a standardized method for describing author contributions, aids in clarifying the specific contributions of each author and determining the data responsibilities of different authors in data papers. However, the content standards and organizational forms of data papers differ from those of traditional research articles (Callaghan et al., 2012). Data papers focus more on the collection, processing, and presentation of data, some of which do not have direct corresponding categories within the 14 CRediT roles. The majority of data papers use only about 9 roles to describe the work conducted, indicating that CRediT has limitations in fully capturing the research workflow. Some roles, such as '*validation*' and '*software*', which ensure data quality and reuse value, are missing in most data papers. This leads to an underestimation or neglect of the actual significance and unique contributions of these roles in research work, affecting a comprehensive understanding of the entire research process. This uneven usage also implies that CRediT fails to accurately and consistently provide all important types of contributions when describing the research work in data papers (Alliez et al., 2020; Fitzgerald et al., 2020; Matarese & Shashok, 2019).

Conclusion

Over the last few decades, data-intensive scientific research has become increasingly prevalent, with some of its labor division characteristics reflected through the publication of data papers. Our study arrives at the following main conclusions:

(1) Data papers rarely make full use of the 14 CRediT roles to describe author contributions, with '*project administration*' and '*resources*' being unmentioned in over half of the data paper samples.

(2) Team size and discipline have a significant impact on the labor division of data-intensive scientific research activities. The need for data collection and analysis is the main reason for the expansion of team size, which is particularly evident in the natural sciences, where authors' roles are more concentrated on '*investigation*' and '*methodology*'.

(3) Corresponding authors and first authors continue to take on core roles, such as '*methodology*' and '*conceptualization*', but at the same time, undertaking data analysis and processing-related tasks, such as '*software*', helps authors advance in the author order of data papers.

Data papers provide an excellent window into studying data-intensive scientific research activities, and CRediT offers a useful framework for characterizing data-centric scientific workflows, but it requires refinement to reflect the characteristics of data papers and the diversity of research work more comprehensively. Developing a taxonomy of contributor roles specific to data papers—DP-CRediT (Data paper-CRediT)—could be a good option. For instance, it might be beneficial to add specific roles such as '*data collection*' and '*metadata management*' to accurately reflect the contributions of these key steps in data papers. Moreover, the importance of technical work such as software development and data processing is increasingly recognized. Alliez et al. argue that since software development in research involves '*significant innovation*', there is a need for appropriate human intervention and qualitative information to ensure accurate reporting of contributions (Alliez et al., 2020). Building on the existing roles of '*software*' and '*formal analysis*', it could be considered to further refine technical contributions by introducing roles such as '*algorithm development*' and '*data engineering*' to describe technical work more comprehensively. Adding roles like '*critical analysis*' or '*data interpretation*' could emphasize the role of human intervention in data analysis and interpretation, ensuring a balance between technical contributions and human judgment. While refining roles, it is also necessary to maintain consistency among existing CRediT roles. Some CRediT roles describe project-level tasks while others describe paper-level tasks(Hosseini, Gordijn, et al., 2023), which could lead to an imbalance in the allocation of academic credit.

Acknowledgements

This work was funded by the project "Publishing Model linked Research Paper and Scientific Data and its application in FAIR-compliant Manner"（23BXW097）supported by National Social Science Foundation of China

About the authors

YANG Heng is a Ph.D. candidate at the National Science Library, Chinese Academy of Sciences. His research interests include scientific data management, data publishing and dissemination, FAIR principles, and related areas. He can be contacted at yangheng@mail.las.ac.cn

YU Yonglin is a master's candidate at the National Science Library, Chinese Academy of Sciences. Her research interests include semantic publishing, scientific information editing and dissemination, and related fields. She can be contacted at yuyonglin@mail.las.ac.cn

LIU Fenghong is a research librarian at the National Science Library, Chinese Academy of Sciences. Her research interests include data publishing, scientific data management, FAIR principles, semantic publishing, and related areas. She can be contacted at liufh@mail.las.ac.cn

References

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship [Article]. Learned Publishing, 32(1), 71-74. https://doi.org/10.1002/leap.1210

Alliez, P., Di Cosmo, R., Guedj, B., Girault, A., Hacid, M.-S., Legrand, A., & Rougier, N. (2020). Attributing and Referencing (Research) Software: Best Practices and Outlook From Inria [Article]. Computing in Science & Engineering, 22(1), 39-51. https://doi.org/10.1109/mcse.2019.2949413

Alpi, K. M., & Akers, K. G. (2021). CRediT for authors of articles published in the Journal of the Medical Library Association [Editorial Material]. Journal of the Medical Library Association, 109(3), 362-364. https://doi.org/10.5195/jmla.2021.1294

Benhamed, O. M., Burger, K., Kaliyaperumal, R., da Silva Santos, L. O. B., Suchánek, M., Slifka, J., & Wilkinson, M. D. (2023). The FAIR Data Point: Interfaces and Tooling. Data Intelligence, 5(1), 184-201. https://doi.org/10.1162/dint_a_00161

Bierer, B. E., Crosas, M., & Pierce, H. H. (2017). Data Authorship as an Incentive to Data Sharing. New England Journal of Medicine, 376(17), 1684-1687. https://doi.org/doi:10.1056/NEJMsb1616595

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P. J., Bowie, R. C., Leadbetter, A. M., Lowry, R. K., Moncoiffe, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. G. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. Int. J. Digit. Curation, 7, 107-113.

Cheruvelil, K. S., & Soranno, P. A. (2018). Data-Intensive Ecological Research Is Catalyzed by Open Science and Team Science. BioScience, 68(10), 813-822. https://doi.org/10.1093/biosci/biy097

Dance, A. (2012). Authorship: Who's on first? Nature, 489(7417), 591-593. https://doi.org/10.1038/nj7417-591a

Das, N., & Das, S. (2020). 'Author Contribution Details' and not 'Authorship Sequence' as a merit to determine credit: A need to relook at the current Indian practice [Review]. National Medical Journal of India, 33(1), 24-30, Article Pmid 33565483. https://doi.org/10.4103/0970-258x.308238

Ding, J., Liu, C., Zheng, Q., & Cai, W. (2021). A new method of co-author credit allocation based on contributor roles taxonomy: proof of concept and evaluation using papers published in PLOS ONE [Article]. Scientometrics, 126(9), 7561-7581. https://doi.org/10.1007/s11192-021-04075-x

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Computer Supported Cooperative Work (CSCW), 19(3), 355-375. https://doi.org/10.1007/s10606-010-9117-8

Fitzgerald, S., Budd, J., Beile, P., & Kaspar, W. (2020). Modeling Transparency in Roles: Moving from Authorship to Contributorship. 2020(7). https://doi.org/10.5860/crl.81.7.1056

Greenberg, J., McClellan, S., Rauch, C., Zhao, X., Kelly, M., An, Y., Kunze, J., Orenstein, R., Porter, C., Meschke, V., & Toberer, E. (2023). Building Community Consensus for Scientific Metadata with YAMZ. Data Intelligence, 5(1), 242-260. https://doi.org/10.1162/dint_a_00211

Greenberg, J., Wu, M., Liu, W., & Liu, F. (2023). Metadata as Data Intelligence. Data Intelligence, 5(1), 1-5. https://doi.org/10.1162/dint_e_00212

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8-12. https://doi.org/10.1109/MIS.2009.36

Holcombe, A. O. (2019). Contributorship, Not Authorship: Use CRediT to Indicate Who Did What [Article]. Publications, 7(3), Article 48. https://doi.org/10.3390/publications7030048

Hosseini, M., Colomb, J., Holcombe, A. O., Kern, B., Vasilevsky, N. A., & Holmes, K. L. (2023). Evolution and adoption of contributor role ontologies and taxonomies [Article]. Learned Publishing, 36(2), 275-284. https://doi.org/10.1002/leap.1496

Hosseini, M., Gordijn, B., Wafford, Q. E., & Holmes, K. L. L. (2023). A systematic scoping review of the ethics of Contributor Role Ontologies and Taxonomies [Review; Early Access]. Accountability in Research-Ethics Integrity and Policy. https://doi.org/10.1080/08989621.2022.2161049

ICSU. Priority Area Assessment on Scientific Data and Information. ICSU. https://council.science/publications/priority-area-assessment-on-scientific-data-and-information/

Kalager, M., Adami, H.-O., & Bretthauer, M. (2016). Recognizing Data Generation. New England Journal of Medicine, 374(19), 1898-1898. https://doi.org/doi:10.1056/NEJMc1603789

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., & Roos, M. (2020). The "A" of FAIR – As Open as Possible, as Closed as Necessary. Data Intelligence, 2(1-2), 47-55. https://doi.org/10.1162/dint_a_00027

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323-1332. https://doi.org/10.1002/asi.23266

Lariviere, V., Pontille, D., & Sugimoto, C. R. (2021). Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT) [Article]. Quantitative Science Studies, 2(1), 111-128. https://doi.org/10.1162/qss_a_00097

Lenhardt, W. C., Conway, M., Scott, E., Blanton, B., Krishnamurthy, A., Hadzikadic, M., Vouk, M., Wilson, A., & Ieee. (2016, 2016

Sep 13-15). Cross-Institutional Research Cyberinfrastructure for Data Intensive Science.IEEE High Performance Extreme Computing Conference [2016 ieee high performance extreme computing conference (hpec)]. IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA.

Lo, B., & DeMets, D. L. (2016). Incentives for Clinical Trialists to Share Data. New England Journal of Medicine, 375(12), 1112-1115. https://doi.org/doi:10.1056/NEJMp1608351

Lu, C., Zhang, Y., Ahn, Y.-Y., Ding, Y., Zhang, C., & Ma, D. (2020). Co-contributorship Network and Division of Labor in Individual Scientific Collaborations [Article]. Journal of the Association for Information Science and Technology, 71(10), 1162-1178. https://doi.org/10.1002/asi.24321

Matarese, V., & Shashok, K. (2019). Transparent Attribution of Contributions to Research: Aligning Guidelines to Real-Life Practices [Article]. Publications, 7(2), Article 24. https://doi.org/10.3390/publications7020024

Nielsen, M. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Nature, 462(7274), 722-723. https://doi.org/10.1038/462722a

NISO. Contributor Roles Taxonomy. NISO. https://credit.niso.org/

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. Data Science Journal. https://doi.org/10.5334/dsj-2017-008

Pietsch, W. (2015). Aspects of Theory-Ladenness in Data-Intensive Science. Philosophy of Science, 82(5), 905-916. https://doi.org/10.1086/683328

Rahman, M. T., & Verhagen, J. V. (2023). Implementing Quantitative Declarations of Authorship Contribution: A Call to Action [Article]. Journal of Scientometric Research, 12(2), 431-435. https://doi.org/10.5530/jscires.12.2.039

Ramachandran, R., Rushing, J., Lin, A., Conover, H., Li, X., Graves, S., Nair, U. S., Kuo, K. S., & Smith, D. K. (2013). Data Prospecting–A Step Towards Data Intensive Science. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(3), 1233-1241. https://doi.org/10.1109/JSTARS.2013.2248133

Resnik, D. B., Elliott, K. C., Soranno, P. A., & Smith, E. M. (2017). Data-Intensive Science and Research Integrity. Accountability in Research-Ethics Integrity and Policy, 24(6), 344-358. https://doi.org/10.1080/08989621.2017.1327813

Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V., & Nane, G. F. (2020). Task specialization across research careers. Elife, 9, e60586. https://doi.org/ARTN e60586

10.7554/eLife.60586

Schultes, E., Roos, M., Bonino da Silva Santos, L. O., Guizzardi, G., Bouwman, J., Hankemeier, T., Baak, A., & Mons, B. (2022). FAIR Digital Twins for Data-Intensive Research. Front Big Data, 5, 883341. https://doi.org/10.3389/fdata.2022.883341

Scraper, W. About us. Web Scraper. https://www.webscraper.io/about-us

Scroggins, M. J., & Pasquetto, I. V. (2020). Labor Out of Place: On the Varieties and Valences of (In)visible Labor in Data-Intensive Science. Engaging Science Technology and Society, 6, 111-132. https://doi.org/10.17351/ests2020.341

Shamoo, A. E. (2013). Data Audit as a Way to Prevent/Contain Misconduct. Accountability in Research-Policies and Quality Assurance, 20(5-6), 369-379. https://doi.org/10.1080/08989621.2013.822259

Smith, E. (2023). "Technical" Contributors and Authorship Distribution in Health Science [Article]. Science and Engineering Ethics, 29(4), Article 22. https://doi.org/10.1007/s11948-023-00445-1

Steele, L., Lee, H. L., Earp, E., Hong, A., & Thomson, J. (2021). Who writes dermatology randomized controlled trials? The need to specify the role of medical writers [Article]. Clinical and Experimental Dermatology, 46(6), 1086-1088. https://doi.org/10.1111/ced.14711

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLOS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tolle, K. M., Tansley, D. S. W., & Hey, A. J. G. (2011). The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. Proceedings of the IEEE, 99(8), 1334-1337. https://doi.org/10.1109/JPROC.2011.2155130

Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O'Donnell, D., Science, D., & Hook, D. (2016). The State of Open Data Report. Figshare. https://figshare.com/articles/report/The_State_of_Open_Data_Report/4036398?file=6558051

Udey, M. C. (2018). Giving Credit where Credit Is Due (and Assigning Individual Responsibilities) [Editorial Material]. Journal of Investigative Dermatology, 138(7), 1451-1452. https://doi.org/10.1016/j.jid.2018.05.010

Vasilevsky, N. A., Hosseini, M., Teplitzky, S., Ilik, V., Mohammadi, E., Schneider, J., Kern, B., Colomb, J., Edmunds, S. C., Gutzman, K., Himmelstein, D. S., White, M., Smith, B., O'Keefe, L., Haendel, M., & Holmes, K. L. (2021). Is authorship sufficient for today's collaborative research? A call for contributor roles [Article]. Accountability in Research-Ethics Integrity and Policy, 28(1), 23-43. https://doi.org/10.1080/08989621.2020.1779591

Wallis, J. C., & Borgman, C. L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. Proceedings of the American Society for Information Science and Technology, 48(1), 1-10. https://doi.org/https://doi.org/10.1002/meet.2011.14504801188

Wilson, A., Downs, R. R., Lenhardt, W. C., Meyer, C., Michener, W., Ramapriyan, H., & Robinson, E. (2014). Realizing the Value of a National Asset: Scientific Data. Eos, Transactions American Geophysical Union, 95(50), 477-478. https://doi.org/https://doi.org/10.1002/2014EO500006

Wittenburg, P. (2021). Open Science and Data Science. Data Intelligence, 3(1), 95-105. https://doi.org/10.1162/dint_a_00082

Zhang, Z., Wang, S. D., Li, G. S., Kong, G., Gu, H., & Alfon, F. (2019). The contributor roles for randomized controlled trials and the proposal for a novel CRediT-RCT [Article]. Annals of Translational Medicine, 7(24), 812, Article 812. https://doi.org/10.21037/atm.2019.12.96