DSI DATA SCIENCE AND INFORMETRICS

RESEARCH ARTICLE

Critical remarks on current practices of data article publishing: Issues, challenges, and recommendations

Quan-Hoang Vuong [a], Viet-Phuong La [a,b], Minh-Hoang Nguyen [a,*]

a. Centre for Interdisciplinary Social Research, Phenikaa University, Yen Nghia Ward, Ha Dong District, Hanoi, Vietnam

b. A.I. for Social Data Lab (AISDL), Vuong & Associates, Hanoi, Vietnam

ABSTRACT

The contribution of the data paper publishing paradigm to the knowledge generation and validation processes is becoming substantial and pivotal. In this paper, through the information-processing perspective of Mindsponge Theory, we discuss how the data article publishing system serves as a filtering mechanism for quality control of the increasingly chaotic datasphere. The overemphasis on machine-actionality and technical standards presents some shortcomings and limitations of the data article publishing system, such as the lack of consideration of humanistic values, radical race for big data, and inadequate use of expertise in data evaluation. Without addressing the shortcomings and limitations, the reusability of data will be hindered, and scientific investment to facilitate data sharing will be wasted. Thus, we suggest that the current data paper publishing paradigm needs to be updated with a new philosophy of data.

KEYWORDS

Data paper; FAIR Principles; Editing and reviewing processes; Quality control; Open science; Mindsponge Theory

> "First come the ideas, then comes an action plan. Never mind the planning required, he excels at this — if a plan is incomplete or not assuring enough, he would correct it. Perfection naturally calls for dedication and diligence."
>
> — In "The Perfect Plan"; *The Kingfisher Story Collection* (2022b)

* Corresponding Author: hoang.nguyenminh@phenikaa-uni.edu.vn

1. Importance of data sharing and data article

Data and datasets are pivotal constituents of scientific research, but they were once mostly considered invisible properties underlying research articles or patents (MacMillan, 2014). However, data and datasets have been increasingly acknowledged and disseminated as official scientific outputs for the last two decades. Some disciplines even consider the generated dataset more important for the research agenda than the associated published literature (Akers & Doty, 2013; Castelli et al., 2013; Reilly et al., 2011). Due to the demand for a huge amount of data, researchers have innovated the production and storage of large international datasets, like the Sloan Digital Sky Survey in Astronomy, the Large Hadron Collider in Physics, the Human Genome Project in life sciences, and NOAA's Climate Data Center in climate science.

Since the beginning of the 21st century, the rise of open science has promoted a culture of openness, which emphasizes openly sharing not only scientific knowledge but also all the properties associated with the process of conducting science (Bartling & Friesike, 2014; National Academies of Sciences, 2018; Space Studies Board & National Academies of Sciences, 2018; Vuong, 2020). Subsequently, data sharing, enabled by technological advances, has become a more common practice for open collaboration, which makes open science possible (Ramachandran et al., 2021). In essence, the importance of data sharing seen through scientific activities reflects its inherent value in humanity's knowledge management as an evolving information collective. Its benefits help enhance the entire system's effectiveness and efficiency in generating new knowledge and filtering irrelevant information.

Data sharing provides resources for scientific conduct, which expedites innovation and discovery. Various disciplines have emphasized data sharing as a crucial driver of innovation (Borgman, 2012; Davis et al., 2021; Lawler et al., 2015; Sanchez & Sivaram, 2017). One of the most remarkable innovations enabled by data-sharing practices is the rapid development of Covid-19 vaccines. Without the early sharing of the first genome sequence of the SARS-CoV-2 on GISAID and Nextstrain, the speedy creation of vaccines might not have been accomplished (shortened from a decade to less than a year) and saved millions of lives (Shu & McCauley, 2017; Vuong, Le, et al., 2022; Zastrow, 2020). In addition to fostering innovation, data sharing can also help alleviate the reproducibility crisis in multiple disciplines (Baker, 2016; Camerer et al., 2018; Hutson, 2018; Open Science Collaboration, 2015; Van Noorden, 2023) by allowing researchers to reproduce and validate research results, identify methodological errors, and conduct open review and dialogue (Borgman, 2012; Vuong, 2017). Although data production and storage costs have significantly declined thanks to the development of digital technologies, they still account for a huge proportion of expenses for scientific conduct. Making data widely accessible can greatly lower the cost of doing science and reproduction (Vuong, 2018), which is even more crucial when national budgets for scientific activities are on the verge of decline (Editorial, 2017; Mallapaty, 2019; Tollefson, 2023).

Actualizing the "data sharing and reuse" culture in academia is challenging. Many obstacles still exist, including methodological, legal, and technical barriers, as well as the lack of incentives for researchers to share their data (Asher et al., 2013; Bourne, 2010; Bourne et al., 2012; Candela et al., 2015; Douglass et al., 2014). To embrace data publication, which is a prerequisite for data sharing and reuse, several attempts were conducted (Candela et al., 2015):

1) publishing data as an integral part of the article, and
2) publishing data residing in the supplementary files attached to the article

However, each publishing model had its own drawbacks, leading to the demand for a new data-publishing paradigm (Candela et al., 2015). When the data are an integral part of the article, they will be difficult to separate from the rest of the materials, hindering its dissemination. Meanwhile, when the data are stored as supplementary files, they require the curation and preservation of such files, and they cannot be shared independently, limiting the findability and accessibility of the data.

As a result, a data-publishing paradigm based on the concept of "data papers" or "data paper" started to emerge as a favorable alternative (Kunze et al., 2011), especially in the biodiversity and Earth sciences (Chavan & Penev, 2011; Pfeiffenberger & Carlson, 2011). It is defined as "a scholarly publication of a searchable

metadata document describing a particular online accessible dataset, or a group of datasets, published according to the standard academic practices" (Chavan & Penev, 2011). Data papers are analogous to conventional research articles, which are products of the academic publishing system. Thus, they are published by journals as primary objects of concern and can be processed by any of the tools and services available for research articles, such as indexing and citation analysis (Candela et al., 2015). However, there still exist differences between data and research articles in terms of their purpose, focus, content, data presentation, and data deposit (see **Table 1**).

Table 1 Differences between research and data articles

	Research article	Data article
Purpose	It presents new knowledge, validation of knowledge, empirical evidence, theories, or insights into a particular field of study.	It facilitates the sharing and reusing of data by other researchers, potentially enabling new analyses and discoveries and validating existing knowledge.
Focus	It presents the findings and analysis of original research.	It describes and provides access to the dataset.
Content (structure)	It typically includes sections such as Introduction, Literature Review, Methodology, Results, Discussion, Conclusion, and Limitations.	It typically includes sections such as Background and Summary, Data Description, Experimental Design and Methods, Technical Validation, Usage Notes (Limitations), and Code Availability.
Data presentation	It may include data presentation, primarily for interpreting and discussing the findings.	Visualizations, tables, and other data representations are required to help readers understand the dataset.
Data deposition	Data deposition is optional, but journals increasingly require reporting of the Data Availability Statement.	Data deposition is mandatory. Prior to peer review, datasets must be deposited in established, community-recognized data repositories (e.g., Zenodo, Figshare, Dryad Digital Repository, and Harvard Dataverse).

International Journal of Robotics Research seemed to be the first journal to solicit and publish a new genre of journal paper (Newman & Corke, 2009). The journal's publication of data papers has two main objectives that are different from previous data-sharing practices. The first objective is to "facilitate and encourage the release of high-quality, peer-reviewed datasets to the robotics community," while the second objective aims to credit the authors for releasing their valuable datasets, as regular peer-reviewed research papers do. Other authors in multiple disciplines, like biodiversity science, earth science, and neuroscience, also promoted quality control through the peer-review process and credit attribution to authors as the main goals of publishing a data paper (Chavan & Penev, 2011; Kennedy et al., 2011; Pfeiffenberger & Carlson, 2011).

With the endorsement of publishers, incentivized authors, and third-party organizations specializing in data archiving, the data-article-publishing paradigm is expanding quickly in terms of data journal and article numbers, substantially increasing the roles of data publication in the current knowledge generation and validation systems. By 2015, more than 100 data journals had been launched (Candela et al., 2015). The largest

data journal is *Data in Brief*, a mega journal that has published over 9,000 data articles since its launch in 2014.

Nevertheless, we found that the current data publishing system presents major weaknesses and limitations based on our editing, reviewing, and authoring experiences in multiple data journals, like *Scientific Data* (Nature), *Data Intelligence* (MIT), *Data* (MDPI), and *Data in Brief* (Elsevier). Without acknowledging and addressing these weaknesses and limitations early, they can significantly affect the quality of data articles, hinder the reusability and operationality of data, and undermine the values of science utilizing data articles. One instance is the typical analysis of Thelwall (2020) on the usefulness of *Data in Brief*. According to the study, even though the journal states to adopt the FAIR guidelines, which provide four fundamental principles – Findability, Accessibility, Interoperability, and Reusability – for both humans and machines to overcome the obstacles of data discovery and reuse (Wilkinson et al., 2016), its published data are rarely reused.

Through this paper, we aim to indicate some underlying weaknesses and limitations of the current data-article-publishing system and provide recommendations to alleviate the problems and improve the system. For elaboration, we will adopt the information-processing perspective of Mindsponge Theory in explaining the data-article-publishing system throughout the paper (Nguyen, Le et al., 2023; Vuong, 2023). This approach is adopted from the metaphysical notion that our world (or physical reality) is composed of information. Specifically, instead of viewing information from the orthodox view:

> Mathematics → Physics → Information, the conceptual hierarchy of information is expressed as Information → Laws of physics → Matter (Davies & Gregersen, 2014).

Due to this characteristic of the information-processing exploratory scheme, it offers great flexibility in explaining many complex and dynamic phenomena, such as theoretical physics, evolutionary biology, and brain sciences (Davies & Gregersen, 2014; Dyson, 1999; Li et al., 2022). Data is a hard-to-be-defined concept that can exist in multiple forms, and the publishing system is a dynamic process, so it will be clearer and more consistent when examining the data-article-publishing system through the information-processing perspective. In fact, the information-processing perspective of the mindsponge mechanism was utilized to examine ideological homogeneity in the knowledge generation process of the entrepreneurial finance field (Vuong et al., 2021).

The current paper is structured into four main sections. In the first Section, we introduce the background of data-sharing practices and the data paper publishing system and highlight the current paper's objective. Before presenting the weaknesses and limitations, we redefine data and data article publishing systems through the information-processing lens in the second Section. Then, the observed weaknesses and limitations are presented in the third Section. The final Section provides recommendations to address the shortcomings and limitations.

2. Data article publishing as a quality filtering mechanism

Since the early days of human civilizations, observing the natural world has been an important way for humankind to accumulate knowledge and generate innovations and discoveries. Despite the differences in its appearance and what it is called, this pattern of humankind has remained the same for thousands of years, from the written records of Ancient Egypt and Mesopotamia in around 3000 to 1200 BCE to the natural philosophy of classical antiquity, which gave birth to modern science (Harrison, 2020; Lindberg, 2010). In the modern age, science has become the most widely accepted way of accumulating knowledge and generating innovation. From the information-processing perspective of the Mindsponge Theory, the scientific community as a whole (e.g., scientific institutions and scientists) can be deemed an information collection-cum-processor that absorbs and processes information from the natural world to generate knowledge and innovations that can benefit humankind.

Within such a process, scientists (as parts of the knowledge generation system) absorb information from the

surrounding environment through their sensory systems to form understandings, explain, and propose theories. However, the sensory systems of humans have limits, hindering them from observing a majority of the natural phenomena. Many tools have been invented and developed to overcome such limits, such as capturing information about natural phenomena that cannot be perceived through humans' sensory systems and storing them as data. For example, Relativistic Heavy Ion Collider is required to acquire data on quarks, elementary particles, and a fundamental constituent of matter; Magnetic Resonance Imaging (MRI) is required to acquire data on neuronal tracts and blood flow in the nervous system; camera traps are required to capture data of wild animals; surveys are required to obtain the data of people's perceptions, thinking, beliefs, and behaviors. In some scenarios, data are also used for validating understandings, explanations, and theories, as they are not always precise and subject to change and subjective biases.

Although the term "data" is very common, it is hard to define as it can be in many forms. One of its most widely cited definitions is the one provided by the National Academies of Sciences, Engineering, and Medicine (National Research Council, 2000): "Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." In a more recent internal document, the Academies also referred to "data" as "data and databases that generally require the assistance of computational machinery and software to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines," in addition to digital manifestations of literature (Uhlir & Cohen, 2011). While the former definition does not specify the digital nature of data, the latter emphasizes data are digital manifestations of observed phenomena that require machines to be processed and analyzed.

To make later explanations clearer and more consistent, we redefine data following the information-processing perspective based on the definition of the Academies. Through the viewpoint of information processing, data, in essence, are the frames (i.e., facts, numbers, letters, and symbols) that are used to manifest information about observed phenomena (e.g., an object, idea, condition, situation, or other factors) and can be processed and analyzed by humans and machines to generate insights that increase human's understanding of reality. Therefore, data are crucial for testing and updating the accuracy of interpretations, explanations, or even theories – fitting the subjective collective mind to the objective world (Nguyen, Le et al., 2023; Vuong, 2023). The results of this process are later translated into knowledge that informs the decision- and policy-making of individuals and organizations.

Being the information that reflects the surrounding environment, including those that human sensory systems cannot observe, data are crucial for knowledge accumulation and innovation generation, and data availability is indispensable to drive science forward and enrich humanity's collective pool of knowledge. Data sharing is an effective way to increase information availability, enabling information diffusion and exchange. Thanks to the development of digital technologies (e.g., cloud and Internet) and the promotion of FAIR principles, researchers can now access better, more cost-effective computational power and more substantial and affordable storage (Ramachandran et al., 2021).

However, while facilitating information diffusion, the Open Science (including Open Data) movement also presents a big challenge: "garbage information." In fact, any researchers can easily deposit their datasets to open repositories (e.g., Figshare, Mendeley Data, Dryad, Harvard Dataverse, Open Science Framework, and Zenodo), significantly raising the entropy of the scientific infosphere. This data chaos, in turn, fuels conflict, increases workload and decreases the quality and reusability of the available data resources.

Problems induced by the chaotic datasphere contribute to the rising demand for a data-article-publishing system. The main difference between normally deposited datasets and data articles is the data publishing system's involvement, similar to the conventional academic publishing of research articles. Academic publishing is essentially the generation, transmission, and diffusion of knowledge within the infosphere of humans (Facer, 2020; Teixeira da Silva & Vuong, 2023; Vuong, 2023). Within such a process, the peer review system acts as a filtering mechanism, with editors and reviewers being the quality evaluators who are responsible for assuring the papers being processed meet the publishing quality and standards (Vuong, 2023; Vuong et al., 2021). In particular, the editors are tasked with initially evaluating manuscripts to determine their

suitability for the journal. They then review the manuscripts within the journal's scope, either independently or by selecting reviewers, before making the final decision on publication (Hames, 2001; Vuong, 2022a). Normally, editors tend to select reviewers within their social networks, known within the field and to the journal, and referrals. When invited, reviewers provide their evaluation of the manuscript's potential, quality, and rigor to support the editors' decision-making process. If the assessed manuscript is a data article, the editors and reviewers will evaluate the technical standards of the datasets and data deposition. Thus, data articles are generally the outcomes of such a quality control process, which involves journal-related stakeholders (e.g., editors and reviewers), authors, and third-party organizations specializing in data archiving.

In general, the data article publishing system can be viewed as one of the filtering mechanisms of the scientific system (besides the research article publishing system) that helps squeeze out the poor-quality and irrelevant data. Data that passes the quality control of the publishing system (i.e., peer review) can be deemed reliable and valuable information for the subsequent knowledge generation within the scientific system. However, the existing weaknesses and limitations in the filtering mechanism have hampered the reusability and operationality of the published datasets and degraded the reliability of future knowledge generation based on the published datasets.

3. Weaknesses and risks of the current data article publishing system

The data published through the data-article-publishing system are believed to become resources for subsequent knowledge generation processes and validation of research findings. To actualize these goals, the published data must be operational and reusable. Operationality refers to the capability of the dataset to be operationalized by humans and machines, while reusability refers to the capability of the dataset to be reused to generate new knowledge and validate the accuracy of interpretations, explanations, or even theories.

Numerous data management guidelines have been proposed to enhance the quality of data operationality and reusability. FAIR principles are widely endorsed by data management and stewardship guidelines (Wilkinson et al., 2016). Besides the acceptance and implementation by many governments and international organizations, most data journals also adopt the FAIR guidelines to direct their operations. For example, *Scientific Data*, a leading data journal launched by Nature Publishing Group, states explicitly that their six foundational principles are designed to align with and support the FAIR guidelines(https://www.nature.com/sdata/principles).

While the spirit of FAIR needs to be advocated and actively supported, in reality, the application of FAIR principles as guidelines for directing the quality control of data articles still presents weaknesses and limitations. Due to such weaknesses and limitations, although the operationality of published data has been upheld through the emphasis on the notion of machine actionability, some problems still exist, and the reusability of the data article remains low, wasting a significant proportion of resources for data storage and dissemination.

3.1. Insufficient evaluation system

The current data article publishing system is built referring to the research article publishing system, so editors and reviewers serve as the benchmarks to decide which data article should be accepted for publication and which should be rejected. However, some problems exist, hindering the effective evaluation of the editors and reviewers.

Firstly, editors and reviewers lack the software/tools specialized for operationality assessment. To examine the operationality of the dataset, they have to download the dataset to computers and use their most familiar tools or software to examine the quality of the data. Although this approach is convenient, it incurs a severe problem: editors are putting their trust in reviewers. If the reviewers cannot use the software/tools or use inappropriate software/tools to evaluate the dataset, their evaluation will be incorrect, contributing to misdirecting the editors' decisions. Therefore, the trust of editors towards reviewers can be deemed as untrustworthy trust.

Second, although machine actionability is essential for the operationality of the dataset, overemphasizing it will lead editors, reviewers, and authors to neglect the other factors underlying the dataset, such as rationale,

design, logical foundation, philosophy, ethics, etc. Typically, FAIR guidelines mainly focus on presenting (meta)data and machine actionability but do not include these factors. Most well-known data journals, such as *Scientific Data* and *Data in Brief*, also concentrate on the rigor of creating, managing, and storing the data (as indicated in the journals' objectives). The negligence of expertise-related content might be influenced by the early perceptions regarding the data paper evaluation process when it was first proposed. For example, in the essay calling for the integration of data papers into the publishing system, Chavan and Penev (2011) presented the technical qualities as the main criteria for the evaluation process:

> "Peer review of the potential data paper manuscript is expected to evaluate completeness and quality of the metadata. This may include the validity of methods used and standards conformance during the collection, management and curation of data. To meet the reviewers' expectations for accuracy and usefulness, the metadata needs to be as complete and descriptive as possible."

The concept of data paper was initially proposed by researchers in natural sciences, so the concentration on the technical aspects of evaluating the data papers' quality can be acknowledged. Nevertheless, even though the dataset can be operationalized well, without sufficient expertise-related content, it is still very challenging for other authors to recognize the dataset's values and how to use it to address the knowledge gap in the discipline. This significantly hinders the data's reusability for knowledge generation and validation.

3.2. The radical race for big data and lack of humanities

Data journals' skew toward the importance of technical standards for operationality can do more harm than good for the reusability of data and the values of data themselves. Specifically, it puts editors, reviewers, and authors in a position that favors the race for a bigger sample size and undermines the data's humanistic values.

The digital age has upgraded the way we conduct science by offering researchers access to big data and its disruptive potential for science and society. One of its significant benefits is the explosive growth of data production/acquisition/navigation capabilities. The expectation toward big data is so big that it leads to some radical stances, like the provocative statement of Anderson (2008): "With enough data, the numbers speak for themselves. […] Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all." Later, some researchers even called for accepting data as science (Hanson et al., 2011): "We must all accept that science is data and that data are science […]." Theoretically, big data are not as sexy as many researchers expected, as explained by Succi and Coveney (2019). However, we will not present how the excessively huge number of data is not promising for complex systems like Succi and Coveney (2019) did. Instead, we would like to highlight the misconception caused by big data: "small samples" are something "bad" and worthless, and big data can replace theories. This perception trivializes the value of thinking processes that lead to the creation of data (e.g., rationale, design, logical foundation, philosophy, and ethics), leading to the race for bigger number of data.

A data race is not a healthy tendency, for overdependence on large samples can diminish serendipity capability – a natural skill underlying human innovations (Vuong, 2022c). Think about the legend of Newton watching an apple fall down and conceptualizing his theory of gravitation. How many falling apples were enough data for Newton to formulate such a crucial law in physics? The answer can be one, or it can also be zero. With the representation of the falling object in his mind, curiosity and thinking were everything he needed. Indeed, if science was all about data, the well-known mass-energy equivalence formula arising from the theory of relativity of theoretical physicist Albert Einstein had not been born. Similarly, how many mold-infested Petri dishes of Staphylococcus bacteria were needed until Dr. Alexander Fleming discovered the crucial antibiotic penicillin? If that single moldy Petri dish had been ignored due to being an insignificantly small number, modern medical history would have been very different.

The overemphasis on machine actionability requires researchers to collect sufficient data points that meet the technical standards so that the machine can work properly and generate reliable inferences. However, each data point is not simply a piece of information that researchers analyze on the computer; in some cases, it is a manifestation of a human's life. Weighing too much on the technical qualities of a dataset (e.g., sample size,

completeness, randomness, and collection methods) while neglecting the reality represented by those data – real people and their life situations – can take away the humanity of scientific endeavors even to the verge of being immoral.

For example, when collecting data about how destitution due to medical treatment led to tragedies for patients and their family members, the author initially planned to gather 1,000 data points (Vuong, 2015). However, reality soon struck. By the time the first 40 responses were gathered, the author noticed cases when the respondents could not finish the survey after 3-4 weeks. Even though the patients were very cooperative, they were constantly busy with paperwork, borrowing money, taking care of the patients, etc. Some respondents even burst into tears while doing the survey and could not continue. For many respondents, meeting 5-6 times was required to complete the survey, and it took about one week between each time.

In the end, experiences from this data collection have required us to think seriously about the humanistic values of data. Behind one data point could be a person suffering on the verge of death. Do we wait until collecting all 1000 responses to have a dataset that is considered "pretty, reliable, and valuable"? The study based on the cut-off dataset, which had 330 data points (Vuong, 2015), was more valuable than those based on the updated one with 1042 data points (Ho et al., 2019). This is because insights generated from the initial dataset must be published as soon as possible for the government, the healthcare system, and the public to take action. Every passing moment means more people have to face financial destitution and the risk of "near-suicide" (Vuong, Le et al., 2023; Vuong, 2015).

3.3. The absence of expertise-related assessment criteria

Reusability is one of the fundamental qualities of data articles. To be reused, the dataset needs to have the potential of being analyzed and used to contribute to the existing pool of knowledge. However, while there are many guidelines for evaluating the technical standards of data articles, expertise-related assessment criteria remain unclear. In most cases, editors and reviewers consider presenting results generated from the data to evaluate their reusability. However, many authors abuse this method. Specifically, they use a substantive part of the data articles for result presentation and neglect the data's details and factors that lead to the creation of the data (e.g., rationale, design, logical foundation, philosophy, and ethics). Meanwhile, those details and factors are important conditions for data reusability. If the result presentation continues to be misused, it can eventually diminish the value of data articles, hinder their reusability, and turn data journals into low-quality research journals that publish studies with superficial rationales, conceptualization, logic, and explanation.

The lack of clearly defined expertise-related assessment standards pertaining to data reusability might also result in data journals being platforms where low-quality data are made open in exchange for recognition. Data articles currently give the authors credit and recognition for the publication of the datasets, which is believed to incentivize authors to make their datasets open. As the number of new data journals is on the rise and indexed in scientific databases, like Scopus and Web of Science, publishing data articles has increasingly been recognized as a way to meet the institutions' KPIs and obtain career promotion. For this reason, data journals become an ideal place for "dumping" datasets that meet technical standards but have limited value for knowledge generation. Sometimes, data papers are published because of the huge sample size. Still, they cannot be used to generate any meaningful results due to the poor design, conceptualization, and logical foundation. In other cases, some researchers only consider publishing their datasets after exploiting all the possible findings that can be generated from them.

Furthermore, without well-defined expertise-related criteria, the data article publishing system might face the risks of being capitalized for metric manipulation. For example, the dataset acquires 500 data points after the first data collection, sufficient for publication. Then, the authors take another year to collect another 1000 data points. Although the second dataset is more valuable than the first one because it is invested with more resources and offers higher validity, it does not have any change in the logical foundation. In this case, should we continue to publish that 1000-observation dataset?

4. Solutions and recommendations for editors, reviewers, and authors

Integrating data papers into the publishing system and the appearance of guidelines for improving the operationality and reusability of data (i.e., FAIR principles) are significant advancements for science. However, some weaknesses and limitations are emerging as the number of data journals and data papers is on the rise, not only hindering the reusability of data papers and wasting scientific resources but also corroding the foundation of science in the long term.

To address the current weaknesses and limitations of the data publishing system, data collecting, reviewing, editing, and publishing activities should not focus too much on the operationality (as reflected through the machine actionability of data) and overstates the values of data (e.g., big data) over the human's thinking processes and humanistic values. We suggest that these activities must be based on two principles empowered by the new philosophy of data.

Firstly, data should be considered as the frames used to manifest information about observed phenomena (including people, animals, events, phenomena, etc.) and can be processed and analyzed by humans and machines to generate insights that increase human understanding of reality, but not lifeless numbers. For such frames to be utilized effectively and appropriately, they must obtain a clear rationale, design, and logical foundation and be based on justified philosophy or ethics. Secondly, data publishing is for the sake of knowledge generation and validation, not data presentation. The reusability of data for generating and validating knowledge needs to be considered equally important as data operationality.

By realizing these key principles, the processes of collecting, reviewing, editing, and publishing data will share a unified vision in alignment with the essence of science: a generation and filtering process for qualified scientific resources (see Figure 1).

Editors, reviewers, and authors need to recognize the value of data based on the properties of the reflected phenomena, especially the dimensions of ethics and humanity. The core question to be asked is whether the data can further our understanding of such phenomena. While technical standards are essential, they should not be overemphasized nor be the sole reference for data evaluation. On this note, the assistance of artificial intelligence (AI) can be beneficial, but over-dependence on using AI for evaluating technical aspects will quickly erode the human value of data. The data validation process also requires including experts in respective scientific fields who understand the data's actual value.

Evaluating data's value and logical foundation should be based on dynamic methods with the information processing approach. One such approach is the Bayesian Mindsponge Framework (BMF) analytics (Nguyen, La et al., 2022; Vuong, Nguyen et al., 2022). The method offers users an analytical framework based on the information-processing view of the Mindsponge Theory to establish and imagine the information process of a system (e.g., a human mind, an ecosystem, a publishing system, etc.) retrospectively using the data at hand. We have successfully employed the method to capitalize on data constructed and published by other researchers in data journals for studying various topics (Nguyen et al., 2024; Nguyen, Jin et al., 2022; Nguyen, Nguyen et al., 2023; Vuong et al., 2024; Vuong, La et al., 2023). Thus, BMF analytics and similar methods are expected to aid data papers' production and evaluation processes.

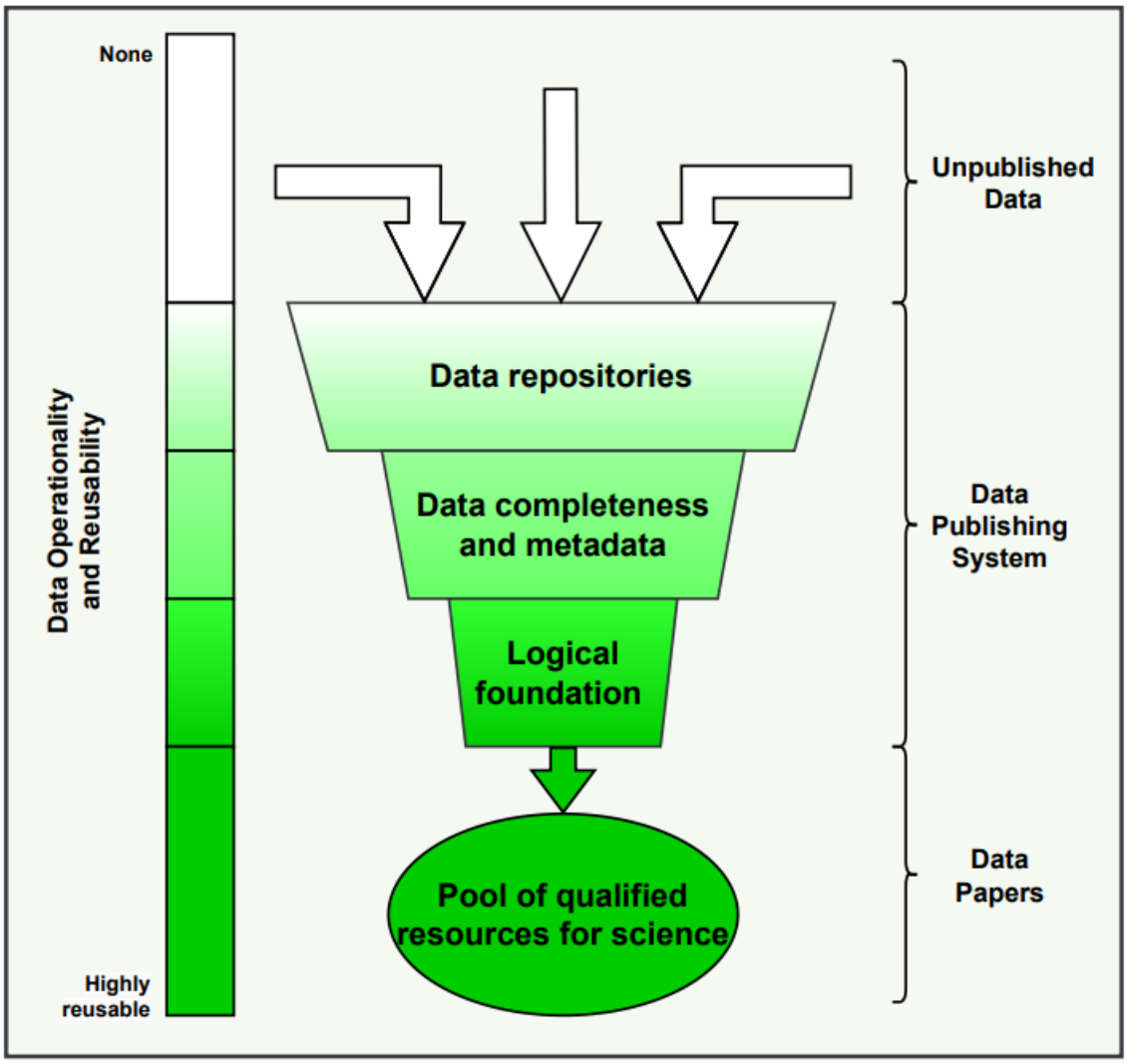

Figure 1 Data paper publishing paradigm with the new philosophy of data

Pragmatically, this new philosophy of data should be applied by the authors, editors, and reviewers throughout the production, evaluation, and publishing processes of data papers.

- *Authors:* The authors should design and collect the data based on a clear conceptual framework, theoretical reasoning, logical foundation, and justified philosophical or ethical standpoints. The data papers must also demonstrate these details to demonstrate the data's values, facilitate the peer-review process, and improve the data reusability.
- *Editors:* Editors should consider the underlying conceptual framework, theoretical reasoning, logic, and justified philosophical or ethical standpoints of the data paper in their decision-making process. The inclusion of such details in the data paper should be set as a requirement for the authors. Editors should consider the disciplinary expertise during the reviewer selection process. Moreover, expertise-related assessment criteria of the data paper should be shown in the instructions to authors and provided to the reviewers as guidelines.
- *Reviewers:* Reviewers should evaluate the underlying conceptual framework, theoretical reasoning, logic, and justified philosophical or ethical standpoints of the data paper, besides the completeness and

metadata quality of the data. Restructuring data and running analysis in various ways are currently not being practiced adequately. Taking a subset in a dataset for testing or randomly examining some parameters can enhance perceptions of data value. This not only informs the authors of the data values but also helps reviewers and editors learn more about the data. Currently, only one side judges the other, which is not the most constructive approach to generating a good-quality data paper, to say the least.

Currently, citations are often used as a key indicator to assess the values and impacts of data articles, just like research articles. However, this citation-based assessment needs to be considered differently and more selectively, as citations can be unequally important. The significance of data articles lies in the reusability of data and its logical underpinning to generate and validate knowledge. Therefore, the citation from a research paper developed from the published data article is much more meaningful than citations for other reasons. This difference also implies that using Web of Science and Scopus citation systems to evaluate impacts and values is unreasonable. Grigori Perelman did not need to publish his work in journals indexed in Web of Science and Scopus to be awarded a Fields Medal (although he refused to accept the prize). In Vietnam, Ngo Bao Chau, the 2010 Field Medalist, has never appeared in the list of most influential Vietnamese scientists generated from the Web of Science and Scopus data. In fact, we might party if our work is cited by a Nobel medalist once.

References

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation, 8*(2), 5–26. https://doi.org/10.2218/ijdc.v8i2.263

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *WIRED*. https://www.wired.com/2008/06/pb-theory/

Asher, A., Deards, K., Esteva, M., Halbert, M., Jahnke, L., Jordan, C., . . . Stark, S. (2013). *Research data management: Principles, practices, and prospects*. Council on Library and Information Resources.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News, 533*(7604), 452-454. https://doi.org/10.1038/533452a

Bartling, S., & Friesike, S. (2014). Towards another scientific revolution. In S. Bartling & S. Friesike (Eds.), *Opening science: The evolving guide on how the internet is changing research, collaboration scholarly publishing* (pp. 3-15). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_1

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. https://doi.org/10.1002/asi.22634

Bourne, P. E. (2010). What do I want from the publisher of the future? *PLoS Computational Biology, 6*(5), e1000787. https://doi.org/10.1371/journal.pcbi.1000787

Bourne, P. E., Clark, T. W., Dale, R., de Waard, A., Hovy, E. H., & Shotton, D. (2012). Improving the future of research communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). *Dagstuhl Manifestos, 1*(1), 11331. https://doi.org/10.4230/DagMan.1.1.41

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Pfeiffer, T. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637-644. https://doi.org/10.1038/s41562-018-0399-z

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science Technology, 66*(9), 1747-1762. https://doi.org/10.1002/asi.23358

Castelli, D., Manghi, P., & Thanos, C. (2013). A vision towards scientific communication infrastructures: On bridging the realms of research digital libraries and scientific data centers. *International Journal on Digital Libraries, 13*(155–169), 155-169. https://doi.org/10.1007/s00799-013-0106-7

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics, 12*, S2. https://doi.org/10.1186/1471-2105-12-S15-S2

Davies, P., & Gregersen, N. H. (2014). *Information and the nature of reality: From physics to metaphysics*. Cambridge University Press. https://doi.org/10.1017/CBO9781107589056

Davis, A. M., Engkvist, O., Fairclough, R. J., Feierberg, I., Freeman, A., & Iyer, P. (2021). Public-private

partnerships: Compound and data sharing in drug discovery and development. *SLAS DISCOVERY: Advancing the Science of Drug Discovery, 26*(5), 604-619. https://doi.org/10.1177/2472555220982268

Douglass, K., Allard, S., Tenopir, C., Wu, L., & Frame, M. (2014). Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. *Journal of the Association for Information Science and Technology, 65*(2), 251-262. https://doi.org/10.1002/asi.22988

Dyson, F. (1999). *Origins of life*. Cambridge University Press.

Editorial. (2017). Budget cuts fuel frustration among Japan's academics. *Nature, 548,* 259. https://doi.org/10.1038/nature.2017.22444

Facer, K. (2020). Convening publics? Co-produced research in the entrepreneurial university. *Philosophy and Theory in Higher Education, 2*(1), 19-43. https://doi.org/10.3726/ptihe.2020.01.0

Hames, I. (2001). Editorial boards: Realizing their potential. *Learned Publishing, 14*(4), 247-256.

Hanson, B., Sugden, A., & Alberts, B. (2011). Making data maximally available. *Science, 331*(6018), 649-649. https://doi.org/10.1126/science.1203354

Harrison, P. (2020). *The territories of science and religion*. University of Chicago Press.

Ho, M.-T., La, V.-P., Nguyen, M.-H., Vuong, T.-T., Nghiem, K.-C. P., Tran, T., . . . Vuong, Q.-H. (2019). Health care, medical insurance, and economic destitution: A dataset of 1042 stories. *Data, 4*(2), 57. https://doi.org/10.3390/data4020057

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science, 359*(6377), 725-726. https://doi.org/10.1126/science.359.6377.725

Kennedy, D. N., Ascoli, G. A., & De Schutter, E. (2011). Next steps in data publishing. *Neuroinformatics, 9*, 317-320. https://doi.org/10.1007/s12021-011-9131-0

Kunze, J. A., Cruse, P., Hu, R., Abrams, S., Hastings, K., Mitchell, C., & Schiff, L. R. (2011). Practices, trends, and recommendations in technical appendix usage for selected data-intensive disciplines. *CDL Staff Publications*. https://escholarship.org/uc/item/9jw4964t

Lawler, M., Siu, L. L., Rehm, H. L., Chanock, S. J., Alterovitz, G., Burn, J., . . . North, K. N. (2015). All the world's a stage: Facilitating discovery science and improved cancer care through the global alliance for genomics and health. *Cancer Discovery, 5*(11), 1133-1136. https://doi.org/10.1158/2159-8290.CD-15-0821

Li, T., Zheng, Y., Wang, Z., Zhu, D. C., Ren, J., Liu, T., & Friston, K. (2022). Brain information processing capacity modeling. *Scientific Reports, 12*(1), 1-16.

Lindberg, D. C. (2010). *The beginnings of Western science: The European scientific tradition in philosophical, religious, and institutional context, prehistory to AD 1450*. University of Chicago Press.

MacMillan, D. (2014). Data sharing and discovery: What librarians need to know. *The Journal of Academic Librarianship, 40*(5), 541-549. https://doi.org/10.1016/j.acalib.2014.06.011

Mallapaty, S. (2019). Australian budget fails to impress scientists. *Nature, 568,* 152-153. https://doi.org/10.1038/d41586-019-01071-3

National Academies of Sciences, E., Medicine. (2018). *Open science by design: Realizing a vision for 21st century research*. The National Academies Press.

National Research Council. (2000). *A question of balance: Private rights and the public interest in scientific and technical databases*. National Academies Press.

Newman, P., & Corke, P. (2009). Data papers—Peer reviewed publication of high quality data sets. *The international journal of robotics research, 28*(5), 587. https://doi.org/10.1177/0278364909104283

Nguyen, M.-H., Duong, T. M.-P., Nguyen, Q.-L., La, V.-P., & Vuong, Q.-H. (2024). In search of value: The intricate impacts of benefit perception, knowledge, and emotion about climate change on marine protection support. *Journal of Environmental Studies and Sciences*, In Press. https://doi.org/10.1007/s13412-024-00902-8

Nguyen, M.-H., Jin, R., Hoang, G., Nguyen, M. H. T., Nguyen, L., Le, T.-T., . . . Vuong, Q.-H. (2022). Examining contributors to Vietnamese high school students' digital creativity under the serendipity-mindsponge-3D knowledge management framework. *Thinking Skills and Creativity, 49,* 101350. https://doi.org/10.1016/j.tsc.2023.101350

Nguyen, M.-H., La, V.-P., Le, T.-T., & Vuong, Q.-H. (2022). Introduction to Bayesian Mindsponge Framework analytics: An innovative method for social and psychological research. *MethodsX, 9,* 101808.

https://doi.org/10.1016/j.mex.2022.101808
Nguyen, M.-H., Le, T.-T., & Vuong, Q.-H. (2023). Ecomindsponge: A novel perspective on human psychology and behavior in the ecosystem. *Urban Science, 7*(1), 31. https://doi.org/10.3390/urbansci7010031
Nguyen, M.-H., Nguyen, T.-P., Nguyen, H.-S., La, V.-P., Le, T.-T., Nguyen, P.-L., . . . Vuong, Q.-H. (2023). Mindsponge-based investigation into the non-linear effects of threat perception and trust on recycled water acceptance in Galicia and Murcia, Spain. *The VMOST Journal of Social Sciences and Humanities, 65*(1), 3-10. https://doi.org/10.31276/VMOSTJOSSH.65(1).03-10
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716
Pfeiffenberger, H., & Carlson, D. (2011). "Earth System Science Data" (ESSD)-A Peer reviewed journal for publication of data. *D-Lib Magazine, 17*(1/2). https://doi.org/10.1045/january2011-pfeiffenberger
Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From open data to open science. *Earth and Space Science, 8*(5), e2020EA001562. https://doi.org/10.1029/2020EA001562
Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on integration of data and publications*. https://epic.awi.de/id/eprint/31397/
Sanchez, D. L., & Sivaram, V. (2017). Saving innovative climate and energy research: Four recommendations for mission innovation. *Energy Research and Social Science, 29,* 123-126. https://doi.org/10.1016/j.erss.2017.05.022
Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance, 22*(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
Space Studies Board, & National Academies of Sciences, E., Medicine,. (2018). *Open source software policy options for NASA earth and space sciences*. National Academies Press.
Succi, S., & Coveney, P. V. (2019). Big data: The end of the scientific method? *Philosophical Transactions of the Royal Society A, 377*(2142), 20180145. https://doi.org/10.1098/rsta.2018.0145
Teixeira da Silva, J. A., & Vuong, Q. H. (2023). Editors with multiple retractions, but who serve on journal editorial boards: Case studies. *Epistēmēs Metron Logos, 9,* 1-8. https://doi.org/10.12681/eml.33935
Thelwall, M. (2020). Data in Brief: Can a mega-journal for data be useful? *Scientometrics, 124*(1), 697-709. https://doi.org/10.1007/s11192-020-03437-1
Tollefson, J. (2023). How the US debt-ceiling crisis could cost science for years to come. *Nature*. https://doi.org/10.1038/d41586-023-01717-3
Uhlir, P. F., & Cohen, D. (2011). Internal Document. Board on Research Data and Information, Policy and Global Affairs Division. *National Academy of Sciences, 18*.
Van Noorden, R. (2023). Medicine is plagued by untrustworthy clinical trials. How many studies are faked or flawed? *Nature, 619*(7970), 454-458. https://doi.org/10.1038/d41586-023-02299-w
Vuong, Q.-H. (2017, 12/20). Open data, open review and open dialogue in making social sciences plausible. *Scientific Data Updates*. http://blogs.nature.com/scientificdata/2017/12/12/authors-corner-open-data-open-review-and-open-dialogue-in-making-social-sciences-plausible/
Vuong, Q.-H. (2018). The (ir)rational consideration of the cost of science in transition economies. *Nature Human Behaviour, 2*(1), 5. https://doi.org/10.1038/s41562-017-0281-4
Vuong, Q.-H. (2020). Reform retractions to make them more transparent. *Nature, 582,* 149. https://doi.org/10.1038/d41586-020-01694-x
Vuong, Q.-H. (2022a). Editor: The demanded but underestimated role in scientific publishing. *Learned Publishing, 35*(3), 418-422.
Vuong, Q.-H. (2022b). *The kingfisher story collection*. https://www.amazon.com/dp/B0BG2NNHY6
Vuong, Q.-H. (2022c). *A New Theory of Serendipity: Nature, Emergence and Mechanism*. De Gruyter.
Vuong, Q.-H. (2023). *Mindsponge Theory*. Walter de Gruyter GmbH. https://www.amazon.com/dp/B0C3WHZ2B3/
Vuong, Q.-H., Duong, M.-P. T., Nguyen, Q.-Y. T., La, V.-P., Nguyen, P.-T., & Nguyen, M.-H. (2024). Ocean economic and cultural benefit perceptions as stakeholders' constraints for supporting conservation policies: A multi-national investigation. *Marine Policy, 163,* 106134. https://doi.org/10.1016/j.marpol.2024.106134

Vuong, Q.-H., La, V.-P., Nguyen, M.-H., Jin, R., La, M.-K., & Le, T.-T. (2023). AI's humanoid appearance can affect human perceptions of Its emotional capability: Evidence from self-reported data in the US. *International Journal of Human-Computer Interaction,* 1-12. https://doi.org/10.1080/10447318.2023.2227828

Vuong, Q.-H., Le, T.-T., Jin, R., Khuc, Q. V., Nguyen, H.-S., Vuong, T.-T., & Nguyen, M.-H. (2023). Near-suicide phenomenon: An investigation into the psychology of patients with serious illnesses withdrawing from treatment. *International Journal of Environmental Research and Public Health, 20*(6), 5173. https://doi.org/10.3390/ijerph20065173

Vuong, Q.-H., Le, T.-T., La, V.-P., Nguyen, T. T. H., Ho, M.-T., Khuc, Q., & Nguyen, M.-H. (2022). Covid-19 vaccines production and societal immunization under the serendipity-mindsponge-3D knowledge management theory and conceptual framework. *Humanities and Social Sciences Communications, 9,* 22. https://doi.org/10.1057/s41599-022-01034-6

Vuong, Q.-H., Nguyen, H. T. T., Pham, T.-H., Ho, M.-T., & Nguyen, M.-H. (2021). Assessing the ideological homogeneity in entrepreneurial finance research by highly cited publications. *Humanities and Social Sciences Communications, 8,* 110. https://doi.org/10.1057/s41599-021-00788-9

Vuong, Q.-H., Nguyen, M.-H., & La, V.-P. (2022). *The mindsponge and BMF analytics for innovative thinking in social sciences and humanities*. Walter de Gruyter GmbH. https://www.amazon.com/dp/B0C4ZK3M74/

Vuong, Q. H. (2015). Be rich or don't be sick: Estimating Vietnamese patients' risk of falling into destitution. *SpringerPlus, 4*(1), 529.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Zastrow, M. (2020). Open science takes on the coronavirus pandemic. *Nature, 581*(7806), 109-111. https://doi.org/10.1038/d41586-020-01246-3