Zoe Y G Dilles. Data journals: Where data sharing policy meets practice. A Master's Paper for the M.S. in L.S degree. May, 2023. 157 pages. Advisor: Todd Vision

Data journals incorporate elements of traditional scholarly communications practices—reviewing for quality and rigor through editorial and peer-review—and the data sharing / open data movement—prioritizing broad dissemination through repositories, sometimes with curation or technical checks. Their goals for dataset review and sharing are recorded in journal-based data policies and operationalized through workflows. In this qualitative, small cohort semi-structured interview study of eight different journals that review and publish research data, we explored (1) journal data policy requirements, (2) data review standards, and (3) implementation of standardized data evaluation workflows. Differences among the journals can be understood by considering editors' approaches to balancing the interests of varied stakeholders. Assessing data quality for reusability is primarily conditional on *fitness for use* which points to an important distinction between disciplinary and discipline-agnostic data journals.

Headings:

data sharing

data publication

data reproducibility

scholarly communications

DATA JOURNALS: WHERE DATA SHARING POLICY MEETS PRACTICE

by
Zoe Y G Dilles

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

May 2023

Approved by:

_____

Todd Vision

TABLE OF CONTENTS

INTRODUCTION

Scientific datasets are scholarly outputs worthy of our concern and attention in the library and information sciences (LIS). In the scientific ecosystem, datasets undergird a large fraction of research findings. A movement toward data sharing arose in response to academies and disciplines that have historically relegated datasets to the background, making them inaccessible to researchers, policy makers, and the public. One consequence of this movement has been the advent of journals that specialize in publishing datasets as primary scholarly output. These journals have processes in place to review the quality of data described in published data papers. For the purposes of this study, we focus on data journals, defined here as any journal that has a process or procedure that operationalizes data review as set by quality standards.

Here, I ask what the relationship is between policies that promote data sharing, practices that make sharing possible, and the role of motivating factors like reusability, data quality, and verificability / replicability / reproducibility at data journals. To accomplish this, I have analyzed interviews of data journal editors. While the cohort I analyze is small, the unique position of these editors within the data publication ecosystem makes their insight and perspective on interpreting and implementing data policies invaluable at a moment in which, at least in the United States, such policies are being written and published often without tools and recommendations for how to enforce or apply them in a workplace.

This study is a timely analysis of interviews that are several years old because there is still a relative paucity in the literature about data journals' practices and no in-depth qualitative interview studies that capture the wide range of perspectives that editors have on data review. Additionally, the interviews were conducted at a crucial time when the lessons from recently founded data journals were fresh in the minds of editors who were often involved in their advent. Our findings may serve as a window into the internal processes at the journals for authors, data managers, and data repository workers that could help them navigate different journals' policies and procedures.

Our work is situated against the backdrop of data journals that focus on data as a scholarly product, many of which voluntarily impose high levels of verifiability on authors. Our study identified three different types of review at data journals (for more on this, see the literature review section on Types of Review and Editorial Role in Review as well as the findings discussed in the section on Reviewing to Different Ends).

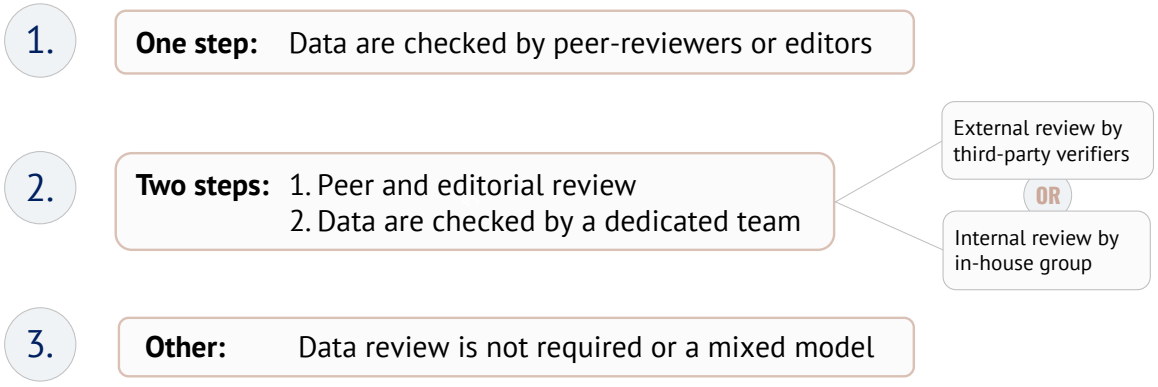

Figure 1: Data journal review models identified by participants in this study.

They use journal-based policies that align with various popular data standards and data sharing agreements that are part of open science discourse. Incentives and factors that might be driving these data journals to uphold rigorous standards include their editors' concerns about their reputations and the validity or replicability of their

published materials. Recent work investigating how a variety of different research funders view scholarly incentives for data sharing echoes these editor's responses in that there is a general concern over the lack of incentives and broadly a lean towards carrots over sticks (Anger et al., 2022). Additionally, it is worth considering more commercial or financial incentives. These could include a desire on the part of publishers to expand the scope their publication portfolios (e.g., the number of journals they put out or disciplinary range).

This work builds directly on previous work about the degree of correspondence between academic editors' and authors' understandings of data policy requirements (Christian et al., 2020). We clarify how a subset of specialized editors not only understand their policies, but how they operationalize them in their day-to-day work with a variety of stakeholders. Additionally, the workflow gaps, pain points, and incentives that editors' surface could yield actionable insights about which technical solutions could be developed and deployed to facilitate this type of work, perhaps of particular interest to scientific software engineers and cyberinfrastructure scholars. These interviews of data journals, conducted in 2018, are an important insight into how data journals work and how data journal editors see their work. As previously addressed in the findings and discussion, data journals' practices were evolving at the time of the interview and have continued to evolve in the years since. Some of the respondents are still in the same role at the same journal as of May 2023 while others are not. Most data journals are relatively young and many of the journals in this sample were founded or changed their data policies to incorporate data review around 2013 to 2015. This means that our insights about the workflows and practices at these journals are an important benchmark in

tracking how these journals policies and practices evolve as part of both the scholarly communications and data sharing worlds. The past successes, directions of growth, and direct recommendations of these journals could be of use in policy-drafting by data managers, academic grant managers and funders, and academic administrators as well as other academic publishers and scientific societies. This is the most in-depth study to date of how editors at data journals describe their working relationships, technical practices, and motivations behind data sharing, data publication, and data review.

LITERATURE REVIEW

Data as Scholarly Output, Data Publication, and Data Journals

Research datasets lay the factual groundwork for much of scientific thought and discovery but have not always been thought of as scholarly products in their own right like journal articles (Callaghan et al., 2012). Researchers and data stewards have sought to elevate the status of datasets and formalize data sharing practices through the language of publication (K. Li et al., 2020). Talking about datasets as publishable has helped incorporate data into current scientific frameworks, although some prefer a less commercial conceptual framing (Parsons & Fox, PA, 2013) . Silvello (2018) contends that our era is marked by data-intensive scientific discovery in which data is as important as a research artifact as a traditional paper. This framing aligns with the fourth paradigm view of scientific research in which computational, even algorithmic analysis, is integrated into the complimentary frameworks of eScience and cyberinfrastructure (Hey et al., 2009). In 2010, De Schutter made a compelling case for data publication as a means to review the quality of data separately from any analyses or interpretations using journal reviewers who understand the detailed data acquisition and curation procedures. This paper argued that such a model could serve to credit people who work with data and to reduce the number of publications with weak analyses yet strong data.

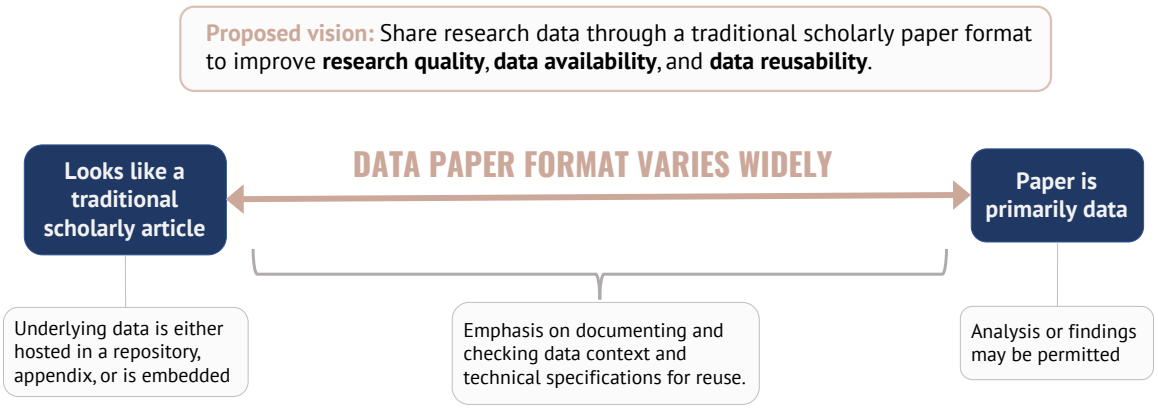

Figure 2: Data paper format ranges from looking very similar to a traditional scholarly article to a paper that primarily consists of data. All data publishers put an emphasis data documentation and availability. This availability requirement varies as well in our sample but is something more than "data unavailable" or "data upon request."

The work of Anderson et al. (2007) warns against the potential perversity of incentives in academic settings. A scarce resource mindset might generally help maintain high standards for what constitutes valuable research but can come at the cost of free and open sharing and even lead to sabotage, which may be fundamentally antithetical to the stated or underlying goals of data publishers.

The data paper has been identified as a means of sharing raw but clearly described, findable datasets to facilitate future reuse (Chavan & Penev, 2011). The contested status of rawness and the impossibility of "raw data" as a unified class or state of being has itself been the subject of scholarship (Bowker, 2005; Gitelman, 2013). A desirable level of "raw"-ness can be seen as contingent on a given dataset's intended use: data simply "provides a starting point for drawing conclusions" (Barrowman, 2018). A data paper may have new affordances for accessibility and reusability of its subject matter with respect to traditional, narrative research papers.

Without facilitation through cyberinfrastructure and collaboration with libraries and data centers, the work of data preservation and sharing primarily falls to authors. Researchers often face barriers in a technical data sharing environment as well as a

general lack of time, making initiatives and incentives to counteract these all the more crucial (Pampel & Dallmeier-Tiessen, 2014). One of the stark realities of data availability historically and to date is degradation of access over time. Work by Vines et al. (2014) tried to quantify the underlying threat of data loss over time, demonstrating that even publicly-archived data are 17% less accessible year over year post-publication. Work by Piwowar & Vision, (2013) found that publicly-archived datasets experience more reuse—a modest but statistically significant citation advantage—compared to non-publicly deposited data.

While literature abounds that attempts to quantify the benefit of open science to authors, publishers, the scientific community, and the public, relatively little has been published on the impact and role of data journals within the scholarly communications and open science ecosystem. The work of Kong et al. (2019) situates these changes in how the scholarly community thinks about data and publication:

> *"As a new model of data sharing, data publication ensures the publishing of high-quality data through some basic procedures including data submission, peer-review, data release, permanent data storage, data citation and data impact evaluation, which maximize the use of scientific data resources. The review process of data publication is very different from that of traditional academic journals. Although many data journals have proposed their evaluation criteria, peer-review of data papers, especially of the data itself, is still at an initial stage."*

Another such paper, Jiao & Darch (2020), looked at data papers and data journals and found that only about half of the citations *of* data papers indicate reuse, although there are other forms of reuse that would not necessarily result in traditional citations. Answering the question of the prevalence and range of disciplinarity in data journals proves to be somewhat complex. No existing index of data journals is comprehensive. In a recent overview of the data journal landscape, Walters (2020) compiled a list of 169 data

journals in which he categorized 19 as those that publish data papers on a regular basis. Candela et al.'s 2015 count of 116 was more than half health sciences; seven, in total, were considered “pure”—publishing only data—and 109 publish datasets as well as other types of papers (See: Appendix C. Sample Overlap with Previous Studies).

There are three clear summary papers that discuss the current operation and status of data journals. There is a relatively recent, comprehensive content analysis paper available that describes publicly available information about the goals of data journals, the incentives for scholars to use data journals as well data journals’ general characteristics and challenges (Walters, 2020). Two other papers about data journals published in *Science Editing* in 2020. One describes the general conditions inwhich data journals arose and performs a content analysis on data paper templates or guidelines from 24 different data journal; the other describes the system of peer-review, evaluation criteria, and editorial board structures as described on each journal’s websites for nine journals (J. Kim, 2020; Seo & Kim, 2020). These studies leave open questions about the social and technical roles of an editor and how they facilitate the different stages of review and what their relationships are to different stakeholders.

Data Quality, Sharing, and Reusability

As the scientific community broadly turns towards increasing data availability and accessibility, there is more community engagement around standards, quality, and workflows. There is no stable definition for data quality. Like data itself, data quality is best described as a relational category in which the value of a dataset arises from its potential usefulness (Leonelli, 2015). Data may be used to assess the reproducibility of

original findings, for reanalysis (e.g. demonstrating novel methods of analysis), or used to explore novel questions (Palmer et al., 2011). This study offers an opportunity to explore characteristics of data quality that may apply across disciplines and ask how such characteristics relate to data sharing practitioners' perceptions of data's reuse potential.

Research data may be used in numerous ways and are tied to the concepts of replicability and reproducibility. Different disciplines define these terms differently as well as related terms like repeatability and reliability. For example, according to Nosek & Errington, (2020), computational reproducibility is retesting the original claims of research with the original data. However, this is not a universal definition and because this study spans diverse disciplines, I am using both replicability and reproducibility to talk about research *that is able to be checked* (National Academies of Sciences, 2019). If it has been checked to the point that every bit of original code was rerun to confirm the original results, outputs it will be referred to as computational replicability for the purposes of this paper.

Both the meaning and quality of data is situated or relational. Clear description and contextualization of data itself is essential to enable and ensure reusability. Studies like Faniel et al., (2019) offer broad insights from data reusers across disciplines about what kinds of contextual information to incorporate into data management and digital preservation practices. Similarly, work by Atici et al. (2013) is vital to understanding the range of interpretive flexibility when data reuse actually happens. Bridging the gap between annotation, description, and formatting for anticipatory reuse does not necessarily map one-to-one with eventual reusability or utility. The call for detailed,

structured metadata, sometimes defined as “data about data”, is common in service of improving reuse potential and interoperability over time. Yet, metadata remains ill-defined. Some scholars view metadata as a process that is distributed amongst different individual data workers and scholars, is often ad hoc and does not necessarily reduce friction in collaborative knowledge production settings (Edwards et al., 2011). As scientific datasets proliferate, meaningful data sharing that enables reuse by both humans and machines will be supported through careful development and deployment of data management plans, metadata schemas, and software and tools to wrangle and parse datasets (Cousijn et al., 2022). Detail-oriented, labor-intensive data review, verification, and metadata development has long been identified as key in harnessing the unprecedented flow of data for wide use and social benefit (Borgman, 2012; Staunton et al., 2021).

One of the great hopes of data sharing is that data quality will naturally increase over time as open research data leads to more exposure, critique, and scrutiny of datasets by wider audiences which will in turn lead to crowd-sourced improvement (O’Hara, 2014). To achieve this, data should be meaningfully open enough to allow inclusive access so that diverse stakeholders can assess the quality, scope, and utility of datasets for themselves (Verhulst & Young, 2022). Wicherts et al. (2011) explored the relationship between error reporting and data sharing finding that in psychology research, authors are reluctant to share data due to the threat or fear of errors being exposed in their analyses. Work by Berberi & Roche (2022) found that, despite claims that opening data has the potential to increase data reuse and lead to error corrections or even article retractions over time in a more accountable and transparent science, there was no such detectable

change after journals implemented requiring data sharing policies. In fact, they point to poorly annotated code and unverified datasets as a challenge in seeing the full potential of open research data come to fruition, even with policies in place.

Data Standards and Requirements

Frameworks and agreements for publishers, funders, and the research community at large including the FAIR Guiding Principles for Scientific Data Management and Stewardship (Findable, Accessible, Interoperable, and Reusable), CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, and Ethics) and the Transparency and Openness Promotion Guidelines (TOP) encourage and recommend policies and best practices for opening science and data sharing (Carroll et al., 2021; Nosek et al., 2015; Wilkinson et al., 2016). Many of these frameworks and standards are adapted into journal-based policies, which are increasingly being adopted across diverse disciplines (Lee, 2022). Although these policies should, "obligate researchers to engage in specific activities to ensure that the research materials underlying published materials are discoverable, understandable, and reusable," the work of Christian et al. (2020) demonstrated that there is a great deal of variability in authors' and editors' understandings of the policies as compared to the text of the policies themselves.

Recent work by Hrynaszkiewicz et al. (2020) is part of the new Research Data Alliance (RDA) standards for robust data policies. These include data availability, data formats and standards, embargoes, and peer-review of data, which is only required at the two highest levels of review. In establishing criteria for data paper review and a data quality evaluation index for data journals, Kong et al., (2019) also help articulate the

relationship between data papers and existing data management frameworks, guidelines, policies, and mandates. Although data papers have a standardized format like traditional research papers, they are distinct because they primarily link to research data rather than interpreting or presenting validated findings. Data papers and data journals blur the lines between traditional knowledge organization domains and academic publishing roles. Because of this, they are perhaps more squarely aligned with the FAIR and TOP Guidelines than traditional papers (Kong et al., 2019; Schöpfel et al., 2019; Wilkinson et al., 2016). The scholarly communications literature is as yet unclear on how audiences use data papers and how embedded they are in existing scientific data discovery networks (Schöpfel et al., 2019).

Types of Review and Editorial Role in Review

Academic publishing practices including peer-review and editorial review have increasingly been the subject of scholarship and scrutiny in the last two decades. From a theory-grounded perspective, peer-review practices are a form of scientific communication that is a process of in-group, disciplinary social judgment by small groups or individuals. Peer-review is the generation of verified knowledge through assessment and selection (Bornmann, 2008). Peer-review is a long-standing quality standard across academic publishing. However, it has had its share of controversy including critiques like the following that should inform our analysis of editor's perspectives on data review practices, "The criteria used to measure review and manuscript quality are subjective, and these outcomes cannot predict an accurate assessment of data quality, novelty, and mechanistic insight by the reviewers" (London, 2021). Peer-review policies vary widely across different disciplines, communities of

practice, and publishers. Open peer-review policies are a relatively recent development within the broader ecosystem of open science practices and policies. Open peer-review can refer to either peer review with open disclosure between authors and reviewers or sharing peer reviews publicly (Karhulahti & Backe, 2021).

Peer-review is sometimes applied to datasets directly and understanding the scope of peer engagement is part of the purpose of this work. Peer-review of data is a traditional scholarly practice applied to a non-traditional scholarly product (e.g., not an analysis or narrative paper). In 2014, when data publication was first gaining traction, a survey explored what scholars in the sciences and social sciences expected of peer-review for data publication. These scholars identified several features of peer data assessment including: appropriate methodology, properly standardized metadata, “technical details check out”, “enough metadata to replicate” and “data is plausible” (Kratz & Strasser, 2015). Projects like the Peer REview for Publication & Accreditation of Research data in the Earth sciences (PREPARDE), funded by Jisc, brought together publishers and data managers to incorporate technical developments and policies into procedures and workflows for data publication (Callaghan et al., 2014).

Another scholarly research output that is increasingly subject to review is research code. A key development in recent years is a variety of scholarly tools aimed at checking the internal consistency of code and computational workflows including the CODECHECK initiative that are closely tied to review processes that are the primary subject of our work (Nüst & Eglen, 2021). Just as data journals grew out of a desire to see research datasets reflected as first-class research outputs on par with analytical findings,

there has been a movement towards research software journals like the Journal of Open Source Software (JOSS) founded in 2016 (Smith et al., 2018).

In recent years, several qualitative studies have sought to characterize journal editor perspectives on their roles and professional challenges in the scholarly communications ecosystem. Though focused on not on data journals but traditional or narrative journals, these studies can ground our approach towards analysis of editor perceptions of peer-review, transparency, editorial standards, and open access. Work by Glonti et al. (2019) thematically explored editors' perceptions of peer-reviewers at a biomedical journal. This study asked editors to reflect on what they think peer-reviewers' tasks are and should be as well as probing the position of peer-review is in the broader scholarly community. This work is particularly useful for our research as they examine how each journal's context, reputation, and operational constraints impact the editorial decision-making process. In another editor-focused study, Maggin (2022), surveyed journal editors across special education and school psychology journals to probe and assess their understanding and incorporation of open science practices and how they do or do not prioritize reproducibility in their workflows. They found that considering the relatively high burden and operational costs of applying new standards in a disciplinary community that does not already have wide adoption of open practices, the journals publishers and editors needed to carefully think about targeted, desired outcomes.

The Current State of Data Sharing Policies

Now, perhaps more than ever, diverse scholarly communities are grappling with how to develop and enforce data sharing policies and when to incorporate or adapt disciplinary or technical standards. This study focuses on data journals within the larger

data sharing or open data scholarly ecosystem and the impact of implementing data policies. So, the abundance of studies that highlight the lack of transparent, reproducible datasets and code in various disciplines and the imperative for data sharing practices and mandates are relevant for our purposes. Even in the few years since Christian et al.'s 2020 study, multiple papers have put out calls for more access to transparent, reusable datasets, and code. These include the work of Culina et al., (2020) examining the availability of code in ecology, Hamilton et al., (2022) which highlights the lack of policy compliance with data and code-sharing in cancer research, and the work of Raittio et al. (2022) which explores the language of "data not shown" in dental research. In disciplines like genomics, there have been increasingly nuanced discussions about the benefits and drawbacks of data sharing of the pressures on researchers and publishers in the scholarly communications environment (Choudhury et al., 2014). A key trend has been increases in calls for and implementation of policies allowing openness of data throughout the COVID-19 pandemic and appreciation for the role of open data for public and global benefit (Cheifet, 2020).

Data sharing is also increasingly required by governmental funding bodies, private funders, institutions of higher learning, and, of course, by academic journals themselves through their journal-based policies. These federal mandates are raising serious concerns and even often from strong open science and data sharing advocates because of their lack of enforceability and unclear attendant budgetary increases (Goodey et al., 2022; Health, 2020; Hrynaszkiewicz et al., 2020; Kozlov, 2022). In 2020, the United States National Institutes of Health (NIH) released a final policy for "Data Management and Sharing" that took effect in January of 2023 (Health, 2020). Just in

August of 2022, the U.S. Office of Science and Technology Policy (OSTP) released a memorandum recommending that all federal agencies: (1) "make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release", (2) "Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies" and (3) "Coordinate with OSTP to ensure equitable delivery of federally funded research results and data" (A. Nelson, 2022). Some authors including Musen & Musen, (2022) insist that, absent minimum standards, metadata requirements, and FAIR Guidelines adherence, most data would still be undiscoverable. With these concerns in mind and given how labor and time intensive data quality review can be, these policies may be at high risk of failure unless implementation is informed by people who are already doing this work.

Data sharing will only be more in the spotlight in the coming months and years due to external motivators for researchers and publishers alike including broad mandates from funding bodies, regulatory bodies, and even data repositories requirements. For example, badges for open data and open materials have been proposed and implemented as signals of participation in practices like data sharing and replication / verification / reproducibility. Badges are one example of a relatively simple, low-cost intervention and external motivator intended to promote transparent and open science practices across scholarly communications and data sharing environments (Crüwell et al., 2023; Kidwell et al., 2016; Radha et al., 2021). As the scientific research culture writ large changes its data sharing behavior, normative pressure could become another external motivating factor for scholars and publishers to promote data sharing (Y. Kim & Stanton, 2016). At

this moment, the perspectives of specialized professionals who sit at the nexus of the sociotechnical practices of working with data and the administrative function of scholarly communications as academic currency might allow many types of stakeholders to better understand how to make data sharing policy into practice.

RESEARCH QUESTIONS

Data journal editors are deeply embedded in the daily work of facilitating detailed data review, up to the point of verification for computational replicability. Relatively little work has been done on the successes and challenges of data publication in a scholarly communications context. No literature to date captures data journal editors' perceptions and experiences of their work in their own words. In 2018, The Odum Institute for Research in Social Science at the University of North Carolina, in partnership with the Dryad Digital Repository, conducted interviews with editors of data journals that publish peer-reviewed data papers describing research datasets and mechanisms for accessing these datasets. These informal interviews are part of a larger research project funded by the Robert Wood Johnson Foundation (#OAR 74419) that aimed to develop evidence-based models for data policy that increases access to quality research data. Christian et al (2020) analyzed (1) the language of data policies of academic journals and (2) survey responses of both the editors and authors of said journals to assess perceptions of their data policies. These authors followed up these findings with semi-structured, in-depth interviews with editors of data journals, journals in which datasets are a primary scholarly output of interest. These interviews have not been coded or published until now and are the subject of this second phase of the project.

In a qualitative analysis of these interviews, we examined editors' perspectives to understand the practices and implementation of data sharing and review policies rather than the language of the policies themselves. The purpose of this study is to document

how these journals apply quality standards for dataset publication in the growing landscape open data practices and policies. Our research questions are:

1. POLICY: How do editors describe their journals' data review policies?
2. DATA REVIEW STANDARDS: What do data journal editors perceive are the successes and challenges of verifying the quality of research data?
3. IMPLEMENTATION AND WORKFLOW: What review mechanisms do data journals use to standardize data evaluation?

Finally, our intent is to analyze these journal editors' recommendations for implementing policies and practices given their perceptions of the impacts, costs, and benefits of data review. We thematically explore what standards can be applied to publishable datasets. The data journals we examine come from different disciplinary contexts and include different types of academic publishers. The open questions of how to perform data review and integrate policy into scholarly communications workflows are all topics that our study design is well-equipped to address.

METHODS

This work is primarily phenomenological, focused on individuals' understanding and lived experience of a particular phenomenon. It is also pragmatic as we are trying to see how editor's personal perspectives on open data practice and policy relate to the data sharing world as it is today and what their recommendations or forecasting of the future are based on their professional repertoires. By performing qualitative data analysis of semi-structured interviews of data journal editors conducted in 2018, I will (1) be able to better contextualize the conclusions and recommendations of Christian et al, 2020 about data sharing policies and (2) contribute a novel perspective to the literature about the tools, techniques, and perceptions of people who perform data validation for formalized data sharing as a scholarly communications practice.

Within the world of scholarly communications research there is a subset of research on publications themselves sometimes called journalology or publication science, which is closely related in this study to metascience. Here, the focus of such scholarship is *data journals*, a subset of academic journals that have processes in place to review the quality of data described in published data papers. In publishing datasets as a scholarly product, data journals engage in *data publication*. As such, datasets can become subject to *editorial review*, assessment by an editor in the peer-review process to determine whether the submission is relevant and rigorous. Quality, as discussed in the background literature review, is a nebulous entity and we are in part investigating which

standards and metrics may or may not be applied to instantiate consistent evaluation and treatment of data as a scholarly output. Amongst such standards in the open data and data sharing community are the Transparency and Openness Promotion (TOP) Guidelines from the non-profit Center for Open Science (COS) that were originally published in 2015; TOP Factor (Factor was previously called Level) describes the extent to which journals either recommend or require submitting authors to include or attest to open sharing of all associated research artifacts for their submissions and to what extent they ask for "independent verification of computational reproducibility using the artifacts to reproduce reported results" (TOP Factor III as described by Christian et al, 2020). Research artifacts or scholarly products include datasets, detailed analytic methods such as codes or scripts, and research material such as codebooks and readme files (Appendix D. Policy Terms).

Positionality / Researcher Role

This secondary research analysis follows the work of from Christian et al. 2020 in which they identified that, "The next steps of this study are to synthesize these findings into a TOP Level III data policy model that offers standardized language to articulate policy requirements as well as guidance for editors and authors on policy implementation and compliance, respectively. This policy model may be informed by the findings of the current study as well as qualitative data from in-depth interviews, to be reported separately, that aimed to better understand the experiences of editors and reviewers implementing data review policies. In doing so, editors—along with members of their stakeholder community—will have the benefit of evidence-based guidance to support the development of an effective data policy that promotes research transparency as part of

normative research practice." I view myself as a junior research member of the UNC team that published in 2020 (Christian et al). My primary role in this study is to summarize, synthesize, and contextualize the as-yet unpublished interview findings so that the development of a cohesive data policy, as described above, model might be possible in the future.

I have never published a paper in a scholarly communication, but I have been involved in the scientific research process in many other capacities over my academic and professional life. I have worked with scientific data in many capacities: collection, transcription and cleaning, analysis, and sharing. From my vantage as a library and information professional, I find myself aligned with the interests of people who are engaged in creating sharing knowledge but often for very different reasons. Researchers in their disciplines are often not aware of the complex connections between different institutional and corporate scholarly communications mechanisms. Research data sharing is complex and takes place in a variety of venues including through digital academic libraries and related services like institutional repositories. Another venue are external disciplinary repositories are sometimes partners to library data work. However, academic journals' motivations to share data, publish data, or review data are not always aligned with the motivations of individual researchers or scholarly communities of practice. Academic libraries and academic publishers are dependent on one another within the scholarly communications ecosystem but fill overlapping or competing roles in research data sharing.

Sample / Research Participants

The population I am studying consists of editors of journals that publish datasets. This is a subset of the population of all journal editors who do so and a smaller subset of those who apply data standards to datasets that are not their own (e.g., repositories, external databases, and data centers). Per the existing interview guide I had access to, I developed research questions in which allowed me to analyze and highlight the differing ways that editors' experiences and perspectives.

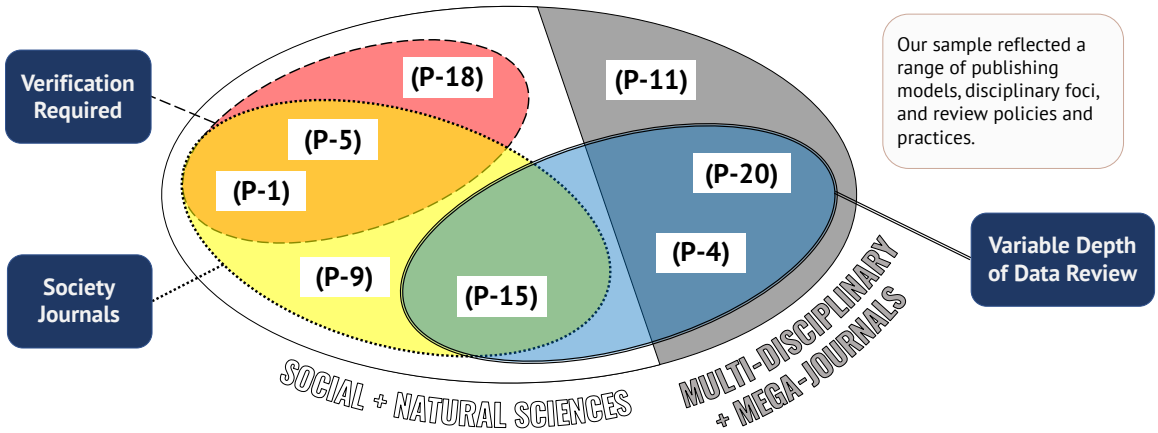

Figure 3:The eight study participants, here anonymized, can be divided based on several characteristics including their publishing model, disciplinary foci, and their review policies and practices.

This study employs non-probability sampling techniques. The initial selection of interviewees was purposive sampling both based on representativeness and maximum variation to cover the range of disciplinary and multi-disciplinary journals that engage in a variety of depths of review. These include journals that satisfy TOP Level, Factor III, or another high level of standard for computational reproducibility as well as others in which the depth of dataset review is shallower, less prescriptive, or entirely optional. As of 2018, the Center for Open Science (COS) only recognized 16 journals at TOP

Level/Factor III as of November 2016 and as of December 2022, they recognize only 13 that use shared data, code, and other materials to reproduce computational findings in manuscript submissions prior to publication (Center for Open Science).

The eight journals that are the subject of this study are the American Journal of Political Science (AJPS), Biodiversity Data Journal (BDJ), Data Science Journal (DSJ), Data in Brief (DIB), Earth System Science Data (ESSD), Ecological Archives (ESA), Scientific Data (SD), and State Politics and Policy Quarterly (SPPQ) (see Appendix A. List of Journals) and can be summarized thusly: two in political science and social sciences, one in physical or earth science, and two in biological sciences for a total of five disciplinary data journals with an additional three general-topic or multidisciplinary journals.

Data Collection Methods

We identified the journals from our list with a data policy that: 1) required authors to provide or describe access to data, code, and/or other research materials associated with the article and/or 2) included a component of dataset review in their workflow. Based on the editors' 2018 survey responses (Christian et al., 2020), we developed an interview instrument to better understand the editor's perspectives on data sharing policies and data review practices (see .

Table 1 and Appendix G. Interview Guide Instrument). The interviews were conducted via web conferencing software in 2018.

Human subject research was approved by the University of North Carolina at Chapel Hill Institutional Review Board (IRB#18–0295, 17–2143). Electronic consent

was obtained from study participants via an online survey instrument (see Appendix F. Letter of Consent). An invitation to participate (see Appendix E. Invitation to Participate**)** in the interview was sent via email to 15 data journal editors for whom the research team were able to locate email addresses from publicly available web resources or other networks. Eight respondents participated in one-on-one virtual interviews that each lasted 10 to 45 minutes in September to October of 2018.

Table 1: Interview questions posed to respondents.

Interview Guide	
Policy Clarification	Please tell me about your journal's data review policy and how it differs from that of journals that publish traditional articles that describe research findings.
	What are your thoughts about the data policy?
	How has the review policy and workflow evolved over time?
	How was the content and language of the current policy determined?
Policy Implementation Workflow	I would like to get a picture of what the data review workflow is like for your journal. Could you describe it for me?
	What have you found to be the most useful in implementing the data review workflow?
	In what ways has this contributed to the success of the policy?
	What, if anything, would you like to have changed about the current manuscript and data review workflow?
	What do you think about those challenges?
	What would have to be done to overcome or eliminate these challenges?
Data Review Standards	What are the standards by which data are evaluated during peer review?
	How have these standards evolved over time?
	How might they evolve in the future?
	What are common issues that prompt a revise and resubmit? rejection?
	How often are submissions rejected due to these data issues?
Comments	Based on your experience as an editor of a data journal with a data policy, what else do you think we should consider as we develop a model for data policy implementation for other journals?

Data Analysis Methods

The interviews were transcribed previously using an automatic transcription service embedded within the web conferencing application Zoom that was used to

conduct the interviews. These transcripts were then imported and verified for accuracy using MAXQDA qualitative data analysis software platform. We applied a coding scheme in MAXQDA to the text of the interviews. We used process codes—words or phrases that capture actions—to characterize how they do their work and their relationship to data sharing more broadly (Saldaña, 2016). I developed a set of primary codes reusing applicable portions of the original coding guide from the 2018 study and adapted from the text of the interview guide. After seeing how these applied to the language of the interviews, I inductively developed a preliminary set of codes which were further refined through two rounds of secondary coding (Saldaña, 2016). I performed the coding of all interviews independently and they will be recorded by a secondary coder, likely Thu-Mai Christian, prior to submission for publication.

I am leaning on the work of Heaton, 2022 to better understand the practical, ethical, and epistemological dimensions of qualitative secondary data analysis as I did not conduct the interviews nor was I involved in the initial design of the interview instrument. My assumptions about their contents are not directly informed by having previously participated in this work. I anticipated that negative case analysis—the practice of intentionally focusing outliers or edge-cases in the process of constant comparison used in inductive coding—would prove extremely important for a small but potentially heterogeneous set of responses (Hanson, 2017). I developed a list sensitizing concepts to help describe myself as the human instrument that will be conducting analyses in a grounded theory approach to understanding a social phenomenon (Bowen, 2006) (See

Appendix H. Preliminary Codes and Sensitizing Concepts). I am drawing guidelines from the work of Guest et al., (2006), Malterud et al., (2016) and J. Nelson, 2017 to determine saturation in the coding process. With my small, purposive sample size and phenomenological research approach, I have reached saturation through several rounds of coding. The study population, journals that have use a component of dataset review, is specific enough in scope and the aim of the study is narrow enough that 16 semi-structured questions yielded enough information for analysis to reach saturation.

TRANSPARENCY AND ETHICAL CONSIDERATIONS

To ensure credibility, I have transparently stated my relationship to the research topic, and I worked with other researchers throughout the project to analyze the data and hone my research practice. Our results include examples of participants' actual statements in the form of formatted quotes, anonymized as necessary, in addition to the inductively coded findings that emerge through analysis. If the terms of this grant-funded project allow, I will be able to walk through the data analysis process and interpretations with a peer as-needed to gain insight from another perspective, especially one who is not as intimately involved in an ongoing research project. I am currently using and adapting existing study instruments—codebooks, interview guides, and survey questions and responses—from the previously published work on this project to ensure continuity and rigor. I have gone to some lengths to contextualize the limitations of my work, especially given the gap in time since these interviews were originally conducted in 2018. I have also tried to ensure that my findings, analysis, and discussion are transferable by describing the research design and methods in their entirety and basing my designs on rigorous studies by other scholars in this discipline.

To ensure confirmability, I maintain an audit trail of my decision-making process, writing process, and data analysis tools, codes, and interpretations. I have also disclosed my personal positionality as well as that of the team that has participated in this work previously. This allowed me to build dependability in my results, discussion, and conclusion portions of my final study in which I can describe how the design process

evolved from the draft proposal stage onwards. This also entails exploring in detail the successes and challenges of the study design as I initially predicted it would be implemented. I sought perspective and insight of several other researchers to help me examine the codes development process.

The research conducted in 2018 that is the basis for this study was funded by the Robert Wood Johnson Foundation (RWJF). This means that some contingent of member(s) of the original research team, Thu-Mai Christian, Amanda Gooch, Todd Vision, and Elizabeth Hull, were being paid, at least partially, by the sponsor to conduct the study. Additionally, Todd Vision, a co-investigator on this study and my faculty advisor, was affiliated with Dryad Digital Repository and, at the time of publication, participates in unpaid activities such as consulting, serving on boards, giving speeches, and writing reports.

Another potential conflict of interest is that two of the sample cohort, AJPS and SPPQ are paying clients of the Odum Institute. The journals outsource their data verification processes for reproducing analytic results of a given submission to the Archive Staff at the Odum Institute. After a successful verification, the paper is released and the AJPS Editorial staff forward the manuscript on and retain final authority over the verification provision.

Although the 2018 letter of consent identified that there are only uncommon or previously unknown risks or discomforts associated with participation in the study, there are risks to the participants due to the potential disclosure of personally identifiable information that might jeopardize or damage the personal or professional reputation of participants. I do not believe that my positionality poses any additional threats or

conflicts of interest that could or would damage the safety or reputation of the participants of this study, but my relative lack of experience as a social science researcher means that I need to be particularly careful to follow best practices in protecting the anonymity and confidentiality of the participants. Some of the information in these interviews could pose a threat to the participants in that their candor or honesty about their practices in their place of work could jeopardize their current or future work prospects. I endeavored to minimize any threats by following the protection guidelines designated in the study's grant proposal in 2018. I am using pseudonyms or generalized references such as "an editor" and have replaced proper nouns in direct quotations to reduce the risk of deductive identification of study participants. Any data that cannot be de-identified were redacted from public-facing contact information continues to be stored separately from data. All study files are stored on an institutional server with protected access controlled by the Odum Institute.

FINDINGS

Data Review Policy, Standards, and Workflows

In the interviews, editors used a rich vocabulary which both converged and diverged on matters of policy, standards, and workflows. In this section, I present high-level themes that emerged across these interviews that highlight how editors at data journals conceptualize, operationalize, and justify data publication including data review.

A *workflow*, sometimes called a pipeline, is both the series of steps that authors go through when submitting their work for scholarly review and the series of steps that editors, reviewers, curators, and publishing teams use to check these submitted materials. There is no unified model of how research datasets are submitted, reviewed, and published. A workflow for a given dataset may include the activities of the author, the stakeholders responsible for review, and the stakeholders responsible for the dataset's storage, preservation, and publication which may each happen on different platforms.

Editors often described workflows in term of technical objects used to direct different stakeholder activities like templates, checklists, and policies or guidance documents. These objects, along with the platforms and software used in communication between stakeholders and to store and share datasets, are tools that can be described as *facilitating*, *implementing*, or *operationalizing* the data publication workflow. For all respondents, the workflows they describe are at least partially if not entirely digital and are often mediated with specialized software.

Stakeholders

There are many stakeholders in the data review process at these academic journals. Here, I summarize their relationships to the respondents.

Editorial Staff

Editorial work touches every process at these journals. Editorial work is *about* managing processes and other stakeholders. Some editors also have a role in developing journal data policies and guidelines. All stakeholders can clarify their journal's overall expectations regarding data given their positions of power and oversight. Some are involved in working groups, committees, or indeed are scholars publishing about issues related to data sharing and data publication. Some identified themselves in the interview as administrators, as scholars, or as publishers. Others make the distinction between themself as the *editor in chief* and the *publisher* or *publication staff*. One interviewee self-identified with the role of *data curator* and *peer-reviewer*. Others describe their work as one or more degrees removed from looking at datasets or manuscripts. One editor self-identified specifically as a *data publisher* and *informatics team member*. Another editor self-identified as a *data manager* and *curator.* For at least one other journal, the roles of *data curation editor* or *data curation manager* are separate.

Academic journals often have a large team of editors with different titles and purviews. Respondents frequently talk about their relationships to other editors and editorial teams. Topical or expert editors have subject or domain specific knowledge that is useful in reviewing certain types of research. Managing, associate, section, or handling editors help manage the different portfolios at a publishing house or the day-to-day operations within a journal. The collaborative nature of editorial work is variously

described through the framework of the editorial office, editorial board, or editorial team. The work of these teams may be complimented with an additional advisory panel or publishing ethics team. Some journal's editorial teams are large and physically dispersed which causes its own issues, "…so catching [these issues] has been really difficult because we have such a big board, and when it's going to 10 different editors… no one sees the trend until you like pull back and I was like, 'Why did we publish three papers that all look exactly the same?' So, we had to do a lot of retractions." (P-20, Pos. 65).

Data Repositories

The editors mention a variety of different digital data repositories, both disciplinary and general. These included specific mentions of Figshare, Dropbox, Dryad, and instantiations of the Harvard Dataverse. Some respondents reflected on the common mission of data sharing amongst data journals and data repositories and how these institutions work together to publish and disseminate scholarly datasets online. This includes integrating data repository standards into the researcher submission workflow or editorial workflow This is further explored in the section on Broad Availability. Workflows that are specifically partnered or integrated with repositories may be based on a shared disciplinary focus or, in one case, proprietary ownership. One respondent indicated that while, "We don't specify what the digital repository should be because most of the time, well, it'll either be domain specific repository or in an orphan works repository" (P-4, Pos. 30).

Reviewers

Seven out of eight respondents volunteered information about how peer-review works for data review at their journal. Within these interviews, the people conducting peer-review are called variously “peer-reviewers”, “referees”, “subject-specialists”, “experts”, or “scholars” by the respondents. We refer to them throughout this work as *reviewers*. Reviewers’ tasks and roles within the overall data review workflow are dependent on specific journal policies can set benchmarks for the scope and extent of data work required of different stakeholders. For at least one journal, “Some editors [may] choose to do the peer-review themself” (P-20, Pos. 33).

Data Curators

Data curators are sometimes described as individuals performing a particular type of data-oriented task but other times they are generalized workers affiliated with a particular team or organization. Some editors refer only to data curation as a function but not to individual workers with specific job titles.

Data curation teams or organizations may be either external or internal to the publishing house or journal. External data curation is either referred to by the name of organization doing the work (e.g., Odum Institute) or as simply the “third-party entity”, “third-party verifiers”, or “replicators.” Editor-provided descriptions of internal data curation work varies while one journal calls these “processers” that handle compliance, others call it an automated team or supplier that performs a “technical check”, the “team that checks licenses”, or “an in-house team of trained individuals.”

One editor remarked that, “the issues surrounding compliance are basically handled by the processers and are somewhat invisible, apart from the fundamentals of: "is

this good, sound data? Is this reasoning upon that data good and sound?" (P-9, Pos. 53). Here, the implication is that the processers are a separate group, but it is unclear if they are peer-reviewers, data curators, both, or neither.

Scholarly Communities

Each of the editors talked about their relationship to the broader community in their interview, describing it variously as the scholarly community, the academic community, the research community, or the scientific community. As previously noted, two of the journals are affiliated with specific scholarly societies. Additionally, two other editors used first-person plural "our" and "we" language extensively when speaking about specific disciplinary scholarly communities that indicated their close personal and professional ties. Some respondents talked about how different subject areas or specialties have localized practices, standards, frameworks, policies, and levels of adoption. Within our sample population, the degree of removal or cooperation between editors or publishing houses and scholarly communities ranges from being entirely in concordance (society journals) to ambiguous or even tenuous relationships for at least one of the multi-disciplinary journals.

Authors

In this study, *author* applies exclusively to researchers or scholars who are participating in data sharing and data publication. Authorial practices and their degrees of satisfaction with a journal's policies or workflows came up in nearly every interview. How much contact each respondent with individual researchers submitting their work varies and was sometimes ambiguous in context. One editor talked about directly soliciting authors for more or clarifying information, and another mentioned "hear(ing)

from your researchers", which implies direct communication. This is explored further in the section on Editorial Role. Many references to authors are collective and reflect particular workflow tasks — "authors revise", "authors make changes", "authors failed", "authors describe", and "authors get the green light" — or perceived feelings — "authors have been very happy", "authors get confused", and "authors are always very satisfied."

Publishers and Publication Staff

A key stakeholder is the publisher or publishing house and its publication staff. This distinction between publisher, editor, and curator is artificial for some journals due to overlapping duties. Depending on the size and hierarchical power distribution within the organization with one editor going so far as to refer to their publication supervisors at the professional society as, "our nominal bosses." The relationship between editorial office publishers or publication administration is clearly an important and many respondents described in detail. For some, it shapes their role, policy, and daily workflow. One editor noted, for example that, "The publisher is almost as known as I am, for example, within our world. And that's unusual. If we look at…most other publishers…they're there on the periphery to some degree, but they are less involved within this community." (P-9, Pos. 87). Some publication staff are quite integrated into the workflow, taking on tasks or roles in the editorial process, a topic that is further explored in the following section: RQ 3: Editor's summaries of their journals' data evaluation workflows.

Some editors have faced pushback or resistance from their publishers when trying to publish data as they see fit, "We've found that the publisher actually is refusing to

publish larger files" or "It's made clear that it's not a priority for the company." Depending on the publisher's funding model, there may be additional, implicit stakeholders in the data publication process, though this was only brought up in one interview: "At the end of the day, we're still a publicly owned, investor-owned company. So, we can't just hire people willy nilly." (P-20, Pos. 101). For one society journal, the data journal is published directly by the society so the editors also help act as publishers.

Data End Users

While the scholarly community includes contemporary and future academic users of published or otherwise shared data, this is not the only potential end user population. While much of the focus of these editors' interviews is about the means of implementing data accessibility, broad availability, and anticipatory usability, the experience of the individual data user, whether real or hypothetical, came up relatively rarely in these interviews. The difference between the frequency of data-centered versus user-centered language was marked: only one editor talked data as being "accessible to the public" or "available to the public." One other editor referred to it as "public data" versus "open data" (n=4) or "open access" (n=5). By contrast, all editors talked at some length about their journals' data availability requirements, rights, and licensing policies. The theme of broad usability and utility was cross-cutting. These issues are further explored in subsequent sections.

Considerations of end users were most often non-case-specific and described in the third person: "think of the user", "easier for the user", and "intended for different users." One editor mentioned the imperative to edit writing for, "a broader audience" while another emphasized that, "It's often about educating authors and the reviewers

about, there are multiple audiences for their data and then often data is not necessarily just aimed at the audience that they had in mind." (P-9, Pos. 80).

Peer-reviewers also see themselves as "potential users of these datasets" according to one editor, which they indicated influenced their willingness to be participate in labor-intensive peer data review practices (P-18, Pos. 44). Only one editor used first-person language when speaking about data users (P-9, Pos. 91). Interestingly, one respondent used second-person language when considering potential end users: "when you, as the next user in sequence…" (P-18, Pos. 24). Only one journal in this study engages in open peer-review practices—those that allow non-anonymous or public review. This journal's editor remarked that, "We find very often people just use their name because they're potential users of these datasets and they don't mind." (P-18, Pos. 44). One editor highlighted real examples of data reuse from their journal's publication:

> *"I know we have reviewed—and other submissions to other journals have looked at—is there any bias in reporting, is there p-hacking going on? And that's only been possible because we have made and required authors to make their data and replication code available."*
> *(P-5, Pos. 77)*

Another editor stated that his publishing house is launching a new article type at his data journal that is "designed to help us highlight examples of data reuse…As the journal goes forward, we are creating new types with new standards which we think help sort of set our data paper, which is still very much the focus of the journal. So that's been the main evolution, is some of these sort of ancillary formats." (P-11, Pos. 67).

Intriguingly, one multi-disciplinary journal editor makes mention of data reuse and reusability but does not highlight reusers as stakeholders. Another multi-disciplinary editor made the singular mention across all the interviews of non-academic engagement

in scholarly data publication: “We get citizen science stuff. We also will randomly get papers from high school classes. Measurements of butterflies and stuff which are kind of cute.” (P-20, Pos. 73). However, this editor did not directly engage with the topic of a broader public, including secondary-school students, as potential data users.

Landscape of Data Publication

Benefits of Sharing Data

Data publication is only one facet of data sharing. Although these publications’ stated data policies and practices vary with respect to open access and open peer-review transparency, all editors shared relatively similar stances on the broad benefits of sharing data for the broader scientific community or greater good. This is perhaps not surprising as their work is directly tied to these practices. Borgman, (2012) highlighted several reasons to share: to reproduce research, to make publicly funded research available to the public, to enable others to ask new questions, and to advance the state of research and innovation. Editors’ other motivations, apart from the general goals of data sharing and data publication writ large are examined in the following section: RQ 3: Editor’s summaries of their journals’ data evaluation workflows.

Broad Availability

Four editors spoke at some length about something like a fundamental premise of opening data availability and accessibility, usually for use by other researchers. The purpose of this data accessibility is tied to a larger, shared understanding about value of data in scientific research. Data sharing, in this sense, is an opportunity to expand or

extend scholarly engagement. The grand project is that data publication and data sharing, "beneficial to the scientific community and everything that we're trying to do building upon knowledge and furthering and advancing knowledge for others" (P-5, Pos. 29) or in the words of another, "building that more interconnected knowledge base about the natural world." (P-9, Pos. 69).

The possibility of reuse and reanalysis including replication and verification was also a strong theme through these interviews. A potential deficit of broad availability—scooping—only came up in one interview:

> *"I don't know of any case of someone taking the data and beating someone else to publication because there's such high start-up costs with what we do. You have to become familiar with the data, you have to understand the data, you have to do quite a bit of research, you have to have a theoretical underpinning. Data in and of itself is not enough for science. You also have to have the substantive, theoretical aspects. And so, it's not just as easy to pick up data, run some analyses, and produce something meaningful for the scientific community. So, I think we sort of laid that fear to rest."*
> *(P-5, Pos. 29)*

One editor was particularly clear about a goal of data sharing at their journal: "What we want to do is help you get that product ready for wider sharing, which we do through publication." (P-18, Pos. 54). This editor, along with one other, noted that that some data sharing is categorically better than no data sharing expressing that data deposition even in an uncurated or unreviewed platform is a net positive. One editor articulated her stance on the value of sharing research data sharing independent of scholarly publishing, saying, "I'd rather have datasets in Figshare than languishing on a CD in a desk drawer somewhere." (P-4, Pos. 46). All the respondents are deeply engaged with issues and trends in data sharing—its possibilities and dimensions. The distinct advantage of data publication, as described by several of these editors, is as a mode of

data sharing is the opportunity to add more context and structure about research datasets that they contend may or will make them more usable long term.

Comparisons to Other Models of Sharing and Publication

In each interview, our team asked a version of, "Please tell me about your journal's data review policy and perhaps how it differs from that of traditional journals." This prompted all editors to speak about the distinctive features of their journals in contrast with other publishers and journals, especially their model or models of data publication. They often spoke about the scope of data publication and level of strictness or depth of the review practices at their journals. Additionally, their comparisons often put an emphasis on their values including verificability, replicability, or reproducibility and their data availability policies or requirements. These themes, along with their perceptions of the overall effectiveness of the policies and practices, are further explored in subsequent sections.

Some editors used this question to reflect on the limitations of their own experience. One volunteered that they had only been in their position for a little over a year and so their knowledge of the processes at the journal was, "limited as well as my limited understanding of what other journals are doing" (P-5, Pos. 21). One editor contrasted their publishing house's collaborative approach to data publication with the scope of other journals:

> *"We have an editorial board…and an advisory panel. This is led by…an active, academic metadata researcher and we engage with them quite often. And our advisory panel includes librarians and data managers and some…non-traditional, well, for natural science-focused journal. There may be people that are a little bit…outside of the usual scope for research journals."*
> *(P-11, Pos. 42)*

This idea of data publication as an opportunity to redefine the scope what is possible or worthy of editorial, curatorial, and peer-review is cross-cutting amongst journals and is further explored in the section on Scope of Data Publication. One editor was very clear about how he thought his journal's scope is one of their distinguishing features in the world of scholarly communications: "We know that the physical and natural sciences have to keep detailed lab notebooks of the work that they're doing, and now it's just holding us in the social sciences to the same standard. Now of course, I think we've gone above and beyond what many in the physical and natural sciences have…we're asking our scholars to share their lab notes with the entire scholarly community." (P-5, Pos. 29). One editor, in a series of rhetorical questions, explored the challenge of both scoping and running a data-centered journal, "How do you make a data journal something that's different? Because if you relax too much you have regression to the mean: data journal just becomes normal journal." (P-11, Pos. 57).

Novelty and Difficulty of Data Publication

A common theme was the unique position of data review and publication models within the scholarly publishing ecosystem. Some are even unique within a given editor's publishing house—the theme of being the only one or the first comes through in their descriptions of other types of publication and data sharing. One editor described the data publishing landscape as maturing and evolving: "It wasn't maybe less sophisticated, pre-acceptance of FAIR…the repository landscape was rather different." (P-9, Pos. 33). Most of these journals are less than 10 years old as of 2023. When editors were asked about how their journal's practices have evolved, several employed figurative language like "evolution", "revolution", "pushing boundaries", and "exploration." For example, one

editor describes data journals as, "…pushing the boundaries of data sharing across geographic range, across cultures, across languages" (P-18), Pos. 59) and later, "We're exploring the different ways that the data could be presented." (P-18, Pos. 85).

The process of adaptation and iteration is a common thread throughout editors' remarks on policies and workflows. One editor expressed the nuance of their journal's approach given the current state of data publication as:

> *"So, we're using a system that was designed for papers and is falling apart to get attribution and credit for datasets and datasets aren't a good fit. Datasets are a lot more complicated than articles. So yes, the processes are going to change and the entire environment around these processes is going to change. I mean, there's already been a shift towards data as a first-class research output, which means that there's a bit less pressure on people to kind of wrap a dataset up as an article and publish it that way." (P-4, Pos. 58)*

This editor's usage of "first-class research output" is unique within this study's recorded responses. It speaks to this particular editor's position as an active researcher in scholarly communications.

Several editors spoke about the increasing popularity of data publication. Editors reflected on the popularity that their journals had seen within certain disciplinary communities or subject areas, one remarking, "I think this will evolve but still there's not that many journals doing what we do. And because so far at least we have a very good reputation for quality, we tend to be a fad journal…And maybe this is good. Maybe it's bad." (P-18, Pos. 62-63) and the other added, "[Some] communities have really picked up fast on the [journal's] concept …and we didn't do anything differently with them. So, we're trying to figure out why." (P-20, Pos. 61). Community adoption was, for one editor, very much a challenge of the past, feeling that their journal's distinctive practices have, "become fairly normalized" (P-9, Pos. 69). Strikingly, one editor who manages a

portfolio of journals with varying data review policies, noted that they have receive very few data paper submissions during her tenure, indicating low rates of community adoption to date. More than one editor alluded to or mentioned there are complaints from authors or misunderstandings with other stakeholders because their journals are so different from traditional publications.

Difficulty of community adoption is only one challenge posed by the novelty of data publication overall. Another theme was the difficulty of data review and working specifically with data. More than one editor called data review for publication difficult, hard, “not particularly easy”, or “it’s not something casual.” The workflows can be challenging for reviewers, editorial staff, and publishers alike, one editor saying, “Well, the hard part is there's not much to compare (the workflow) to.” (P-20, Pos. 53). The review processes are different enough from normal journals that at least one editor found that peer-reviewers “actually use” the checklists and advice provided to them (P-11, Pos. 60).

Only one editor remarked on the benefit of novelty and evolving state of data journals as a positional advantage for their journal in the publishing landscape: “It's our job to be alert to, informed by, and competent in these evolving technical standards so that when the next set of authors come in with a new important dataset, we can adapt to what they're currently using.” (P-18, Pos. 52).

One editor of a multi-disciplinary mega-journal lamented that scientific societies sometimes see the popularity of their journal and propose spinning off their own society-based data journals. This competitive relationship of some publishers with scientific societies was most evident in this interview and was absent in others.

Scholarly Incentives and Motivations

One editor was specific about the history and strategy of using the data journal to disseminate scholarly work and the power of using the paper format:

> *"Journals came about as a structure to publish and disseminate the results of academic work. The problem became the fact that the only true research output, as far as tenure and promotion committees were concerned was the paper. We, as data managers and curators, essentially tried to shoehorn data into an article shaped box and then package it up and pretend it was an article so that it could go through the journal publication processes in order to get the attribution and credit that the person who created the dataset would have got if they'd written a paper instead, right?"*
> *(P-4, Pos. 58)*

Many of the editors spoke in similar terms about using the existing rewards and accreditation systems within academia and scholarly communications to incentivize authors including citation or attribution, impact factor, and visibility connected to the status or reputation of the journal. Some editors also spoke explicitly about dangling the metaphorical carrot of citations for authors.

Incentivization of data sharing through data publication was a common theme that each of the editors brought up. This included not only how to build incentives for researchers to engage directly with their journal but how to promote the model of data sharing through data publication to other journals or publishers. Some contrasted their model of data citation with traditional research journals in that they can give credit directly to the data with a publishable reference that would be potentially useful in navigating tenure and promotion rewards structures. As one editor put it, their journal's data citation and attribution model should be seen as a boon to researchers: "To the extent that you're using these metrics to fight your way through your career and promotion, tenure and promotion systems…our product is an ally in that." (P-18, Pos. 79). Overall, half of editors highlighted the citation, attribution, or crediting advantage for authors that

choose to publish through their journals. The advantage of sharing through data publication is some combination of credit and credibility in one editor's estimation, "Our perspective…is anybody can publish data anywhere, but when you put a peer-reviewed scientific publication behind it, it has a lot more credibility." (P-18, Pos. 89).

Editors were specific about their roles in implementing these incentive structures and promoting their product, one leading off with, "We make it worth your while as a researcher." (P-18, Pos. 79) and another adding, "We just…try to give out this value statement and you're basically getting a twofer when you publish the separate data article or methods article…You're helping highlight what would essentially be like your lost supplementary materials." (P-20, Pos. 17). The "twofer" was a unique reference to that journal's submission model, but another editor noted that, "If an author uses the dataset that was published on Dryad, but the paper announcing that dataset is in [our journal]: it's, two citations should be being made and we're finding one is being made most of the time. So, from the publisher standpoint, we're really worried about that citation aspect and making sure everybody gets their credit and all of the journals get their citations." (P-18, Pos. 85). This highlights the technical dimension of publishing and citing datasets as a primary scholarly output and trying to use systems not built to accommodate them: three editors independently brought up the retrieval role of Thompson/Clarivate/Web of Science or Mendeley and the challenge of making sure data papers are accurately indexed. These editors expressed concern that datasets, when not embedded within a narrative document or when hosted on an outside repository, may not be retrieved and therefore searchable in aggregators.

Several editors acknowledged that not all citations are created equal, highlighting the role of a journal's reputation with respect to visibility, impact factor, and perceived credibility or quality. At two journals that engage in strict computational reproducibility practices, one editor sees their high impact factor as a motivator for submitting authors while the other journal's editor cites their practices as the cause of their increasing impact factor. The latter journal also engages in badging as not only a motivator for their published authors but as a call to action for other journals: "We hope the incentive is enough to get a publication in [our journal] …I think that we can only continue expecting that if other journals also join up and that that becomes the new status quo. We also try to incentivize by the little badges that we put on articles that send a signal to readers that this has been verified and replicated…" (P-5, Pos. 81). The role of the editor in mediating the overall reputation of the journal is further explored in the section on Editorial Role.

One editor briefly mentioned the role of European external data sharing initiative Plan S. Another mentioned funding mandates as external motivators for authors to share data. Additionally, two editors mentioned incentivization through affordability, one in reference to the supporting role of their publishing house and the other for promoting data publishing to other journals. In two instances, editors highlighted intrinsic motivators for authors "sense of success" that authors might experience by going through the hard work of data review while another described the potential benefit to authors as an opportunity to go more into detail about their work.

RQ 1: Editors' descriptions of their journal's data review policies

How do editors describe their journals' data review policies?

When asked about how their journal's paper's policy is different from that of a traditional journal, editors' responses ranged widely. One editor affirmed that they are organized like a normal journal while another asserted that, "a normal journal couldn't do what we do."

Editors talked about their journal's data availability and data review guidelines, informal practices, and written policies. But before we can analyze how the policies may or may not be shaping the processes for data review, what do these editors think the policies say?

Some editors poked the utility of generic or broad policies a given the heterogenous characteristics of research datasets and different stakeholders. One remarked, "You can't please all the people all the time, so don't even try. I'd say if you can get 80% of the situation sorted out, then you're doing really, really well." (P-4, Pos. 70). Another acknowledged their aspiration for, "an across-the-board policy" (P-18, Pos. 33) but made it clear that their publishing house does not have clear or consistent internal policies for data review. Some editors manage more than one journal and so they are managing multiple policies and guidelines for each, some of which overlap. One editor implied that there is perhaps no lower limit threshold of what makes a policy effective: "I think some policies is better than no policy and some attempt to do it even if it's in house is better than nothing." (P-5, Pos. 73). This editor also expressed a unique concern about the potential threat of venue-shopping and how consistent policies across journals might be a bulwark:

> *"…Given that other journals don't have as strict as requirements as [our journal], it seems that—I've heard anyway at conferences and such—that there is some venue shopping that happens, right? So, scholars who may not have done as good of a job of documenting how they produced their analyses or their data or their perhaps there's still*

> *some concerns that they sort of go to other journals. So, I think having some consistency across policies within discipline would matter."*
> *(P-5, Pos. 73)*

Data journals sit at the nexus of academic publishing, disciplinary communities' standards, and data repositories. So, these editors views of data journals policies often emphasize their interconnectedness. For example, the transferability of a policy and the scalability of the workflows practices journal are linked in this quote from a multi-disciplinary journal editor: "I think the big issue for the academic community is for each sub-specialty... defining for itself what their data policy is going to be. They're all doing the same thing right now they're all at the same level, just building these frameworks and these working documents." (P-20, Pos. 121).

Overall, when editors talk about policy, they talk about many different things: policy development, tools that they as editors use to clarify the policies, and the necessary conditions for review. For example, data needs to be available for it to be reviewable. One editor, when asked about the journal's specific policies for data review immediately began explaining in detail what their reviewers *do* to review datasets. This specifically demonstrates how the journal's policies shape the workflow for data review Others did not connect their journal data review policies and practices in their interviews.

One editor described the "painstaking work" that the original editor of the journal had taken in developing the original data review policy continues to serve as "structure and framework" for the editorial and data curation staff at their journal. As one editor put it: "It's easy-ish enough to write a policy. Where the real complications then start creeping in is in the implementation." (P-4, Pos. 70). These nuances are further explored

in the following section: RQ 3: Editor's summaries of their journals' data evaluation workflows.

Policy Development

Just has the field of data publishing is evolving, these data journals' policies are in different stages of development. Policy development includes how these policies were initiated, how they have grown to date, and their current status including the stability of the existing policies. The journals policies fall into two camps, either being in a state of flux (n=6) or generally stable (n = 2). They range from being long-standing to, "My journal doesn't actually have a formal, written down data review policy as yet," only stressing the diversity represented amongst even such a small cohort study. Congruently, when asked about evolving data review policies and workflows, the editors' responses covered a huge spectrum from: "it is evolving right now" to "to be honest it has not evolved, it's been really frustrating."

Most of the journals' policies have changed over the years. The editors describe this as updating, tweaking, and iteration, sometimes calling their policy "a work in progress" while two others described a process of guess and check for effectiveness. However, for one journal the data review standards themselves are amongst, "the most stable elements of our policy", according to the editor (P-11, Pos. 66). Another discipline-focused journal has had a strong open data goal since its founding that the editor says they have not erred from, but the editor did not reflect on the specifics of a data review policy. One society-funded-journal's editor took a very strong stance that the journal's data review policy, which he helped author, and which has been in place since its inception, is highly effective and that in that respect, it might be difficult to improve

upon (P-1). The other society-funded-journal editor also remarked that, “the actual underlying data paper itself hasn't changed too much in 10 years.” (P-18, Pos. 77). Some publishing houses have their own set and structure of policies. An editor described a publisher-wide timeline for increasing the transparency and openness of their journal’s practices: “It started out basically as every journal had… it's called ‘Level One’, there's four different levels. It's only recommendations and for some journals, it's every six months they go up another level. Some journals, the editorial boards have decided to stay at one level or another. All I've seen evolve really is attitudes.” (P-20, Pos. 41).

Motivations for changing their policies varied. Several journals have adapted their policies to manage specific cases that arise. For example, one respondent brought up an interaction in which a particular university was sending in multiple versions of the same submission. In response, the journal, “implemented this policy that now anyone from that school needs to submit along with it a signed affidavit verifying the authenticity of the work, that it's their own work, and that it was commissioned by the university” (P-20, Pos. 73). Editors are often heavily involved in this aspect of policy updating as they oversee the workflow, so this is examined further in the section on Editorial Role. Most of the editors are involved in either clarifying the language and scope of the policies or building tools and guidelines that also help communicate the policies expectations. Two editors briefly mentioned that they have updated their policies over time to be stricter to reduce the overall number of submissions through deterrence as these journals deal with high volumes.

Relationship to External Guidance and Community Standards

The process of policy development for many of these journals has been in concert with external or disciplinary community standards. The editor of a journal with strong disciplinary-community ties noted that their policies have changed, "in relation to the accessibility of the repositories themselves and the conditions that they impose" with approximately half of the journal's standards arising from their domain specifically and half from outside (P-9, Pos. 33). Specific disciplinary community standards that editors mentioned included DarwinCore, the Taxonomic Databases Working Group (TDWG) which is now called the Biodiversity Information Standards, and Topologically Associating Domain Knowledge Base (TADKB), though many referred to their journal's relationship to unspecified domain-specific standards. Only two editors mentioned how their journals incorporate or lean on community-established standards or policies, one discipline-specific and one multi-disciplinary. The pragmatic editor of a multi-disciplinary journal said, "There's never going to be one standard to suit everybody and that's okay, but I'm kind of coalescing around domain specific standards are definitely a useful thing and will help with an awful lot of processes, not just academic publishing ones," adding that the journal reviews datasets for whether variables are "named appropriately according to domain specific ontologies and controlled vocabularies." (P-4, Pos. 50-54). Another editor of a multi-disciplinary journal added that, "most journals will…require data according to community standards" (P-11, Pos. 16) and that his journal's peer-reviewers often look to community standards in their review practices, "And then, you know, tends to be some debate within the community. Some communities

have very specific standards and so that could fall into it. Like have you complied with the standard? Could you do that better, etc." (P-11, Pos. 81).

Repository standards came up in three interviews, editors nodding to the conditions, requirements, or policies of third-party hosting sites. One editor mentioned that while their journal doesn't have control over any repositories policies, they do in fact rely, to some extent, on the fact that author's go through an acceptance process for "a data center with a strong curation," implying that they are outsourcing their own policies to these repositories. The other two editors highlighted that they have changed their policies to meet data repositories' standards and even cited them or adapted such standards for their own uses.

A few respondents mentioned external guidelines and initiatives how they have been incorporated or how they have shaped policy development. Surprisingly, only one journal explicated the influence of the FAIR principles at all. Their journal's policy predates FAIR but, "as things like FAIR came along, we stopped to cite them and adopt them" (P-9, Pos. 37). Only one interviewee mentioned the TOP Guidelines and that they underly their data review policies, "We try to keep the high, TOP-level standards sort of fixed and primary but underneath you need a lot of flexibility." (P-18, Pos. 52). A different editor brought up the Force 11 working group on data citation principles as well as PREPARDE and their work dataset publication, repository requirements, and peer-review of data (P-4).

The responses of these editors have not coalesced around any set of core external standards that their journal's policies are built off. Many of them explore how existing academic publishing data guidelines are incorporating community-established standards

and best practices as needed. By contrast, some journals see the uniqueness of their policies as a distinctive and even advantageous feature of their publication in a branding or marketing sense.

Data Policies as Branding

The data policy can be a branding tool for some journals. As a record of the journal's mission, it helps them form the journal's identity. One multi-disciplinary editor put it this way, "It helps us be different. It also makes us, I think, a bit unfamiliar." (P-11, Pos. 57) later remarking on the tradeoffs of unfamiliarity, "There's some of these arbitrary decisions that…have become part of who we are…, but anything that's different will always be a bit of a challenge for someone." (P-11, Pos. 57). One editor highlighted that the flexibility built into the different active journal data requirements within their professional society's journals allows them to appeal to a larger market of researchers: "I think that's what makes our society, very, very good for most people is we have a home for everyone, depending on what they want to do." (P-18, Pos. 33). This editor was not alone in describing that a wide variety in options within a single publishing house, whether a society journal or large press, could prove attractive for authors. As another editor put it, "[now we have] clearer policies on how institutional repositories can be used, because we don't try to list all of them, which gives us a much broader scope and allows people to submit a much wider range of datasets to the journal." (P-11, Pos. 33). While this may make the journal more marketable to a broad range of authors with a wide range of datasets, it does not make the workflow review process simple, as we will see in further sections.

Clarity and Communicating Expectations

Many of the editor's responses to our question about their journal's policies complicate the idea of a single data policy, much less a concrete data review policy for a given journal. One editor immediately replied with, "What do you mean by policy? Because we have several different structures..." later indicating that, "…authors get confused about what options are available to them and we have to continually remind ourselves what we're processing because of the different policies." (P-18, Pos. 33). Another editor described their journal's policy environment by saying that the publishing house provides, "a family of policies that we stand on top of." (P-11, Pos. 42). Respondents often mixed references to organization or publisher wide data policies, external data standards, and other guidelines. In fact, several editors brought up how complex and potentially confusing the policies are in their estimation for editors, reviewers, and authors alike. Similarly, another editor of a large publishing house that has multiple journals: "We've tried to streamline our message to authors, because our data story is pretty convoluted. I think we'll pretty readily admit that, like when you go to—I could show you any given [Publisher] journal, when you go to submit a paper, you're basically fed like four different stories on sharing your data." (P-20, Pos. 17). Communicating expectations through policy can be complex given the multi-layered policy structure at some journals. The role of the policy itself as tool or structure to communicate with stakeholders or resolve misunderstandings came up in several interviews. One journal with a particularly detailed review process observed:

> *"But having an actual policy, having guidelines for authors that has, I think, set the tone and structure for the process. So that's been really, really beneficial. Because it's easy to go back to when there's something that falls outside the norm or even something that falls inside the norm. If there's questions that come up, it's sort of easy to go back to and say,*

> *"This is what we mean. This is what our expectation is." So having an actual written policy—and a very detailed policy—I think I have looked on other journals' websites and they sort of have an overarching framework and a requirement that you upload the data."*
> *(P-5, Pos. 33)*

Several interviewees spoke at length about what tools they use to clarify the intent of their journals policies to different stakeholders. These include checklists, information on their websites, and published papers and editorials. Two journals have historically leveraged the position of the editorial office to clarify what their institutional goals, standards, and policies are regarding data through published editorials. For one, the journal's original editor in chief initially framed their policy with an editorial (P-18, Pos. 45) while the other mentioned that they would be soon featuring an editorial about the purview of peer-reviewers in examining datasets and that they had previously put out another, "outlining our path for clinical data" (P-11, Pos. 34).

At one natural sciences journal, they had initially published the guidelines for reviewers as a paper in the first issue of their journal that, "basically set out what the journal was aiming to do, under the conditions in which we would publish." (P-9, Pos. 33). They had published these guidelines as a paper in another peer-reviewed journal within the same publishing house as well. At the time of the interview, they had also recently published their current guidelines as a paper in a different journal. These guidelines, "describe all the conditions…under which data is public, why data is published, how it is reviewed, and what kind of processing, it might go through." (P-9, Pos. 29). The other biological sciences journal editor also highlighted that they had a separate set of peer-review guidelines for data papers (P-18, Pos. 97).

These editors were not the only ones who pointed to the utility of reviewer guidelines in clarifying their policies. These may take the form of—as previously mentioned—peer-reviewed articles as well as internal-usage policy-supporting instructions, templates, and checklists. As one editor put it, “This is well-described on our website…there's under our policy section…called ‘For Referees’ that goes in quite some detail through our peer-review criteria. And there's actually nine questions that we give to our peer-reviewers that that really goes through in detail what our peer-reviewers are supposed to check for.” (P-11, Pos. 60).

One editor noted how their journal tries to help authors comply with their policies by talking about it in their acceptance letter as well as putting details about required documentation and procedures on their website: “We also have PDFs on our journal website for…how to prepare replication files, how to prepare the code.” (P-1, Pos. 17). Additionally, three different journals operationalize their policies regarding the required documentation and description of datasets for authors using a template in their submission system. One editor mentioned briefly “a specifications-type table” in their template though, in practice, this appears not to work universally as authors can and often leave these fields blank (P-20). Conversely, one editor spoke in detail about their journal’s complex template system. This system includes predefined paper format templates for different data paper types ranging from something like single-species descriptions to an R module. However, the editor also added that there are limitations to templating datasets, “…(if) you're publishing a generic dataset for which there is no regular standard then the template can't be very prescriptive about what it needs to write because frankly we don't know.” (P-9, Pos. 42). For the other journal that takes this

templated approach, the editor mentioned how unique their template is in his estimation and how this is, as discussed in previous subsections, both a branding tool and point of comparison with other journals.

At the only journal that did not have a written data review policy at the time of the interview, the editor herself serves as a main means of clarifying the data review criteria as it is, “still very much an evolving process” and she is highly engaged in the academic literature and community of data sharing and data publication in her other professional capacities. This journal was primarily relying on traditional peer-review expectations to clarify what data review should constitute: “It's a standard kind of review: comments, come back, and revise, resubmit…” (P-4, Pos. 30). One other editor did not mention any tools that they use to clarify the data review process for different stakeholders though the editor did talk in detail about the review criteria which are also outlined on the journal’s website.

In sum, most of these editors mention the use tools other than just the language of the policy itself to clarify and communicate expectations including putting out editorials, authoring peer-reviewed work, being active scholars in data publication, making guidelines for reviewers, and creating templates for authors to help facilitate policy-compliance in the submission process. A brief examination of existing documentation on each of the journals’ websites reveals that there are variety of scaffolding documentation and tools that they make public for the benefit of authors and peer-reviews.

Table 2: Summary of Publicly available Guidelines for Peer-Reviewers and Authors at Eight Journals that Publish Data as of 2023

Journal	Guidelines for Peer-Reviewers and Authors
American Journal of Political Science	The journal's current website features documents like "Guidelines for Preparing Replication Files" and "Verification Checklists" designed to help authors.
Biodiversity Data Journal	The journal's current website features documents like "Guidelines for Preparing Replication Files" and "Verification Checklists" designed to help authors.
Data Science Journal	This website includes links to repositories and external protocols that authors can use in preparing their submission as well as the guiding questions for reviewers.
Data in Brief	This website includes a submission checklist, a frequently asked questions section, video tutorials about appropriate data repositories and the submission and many other guides for specific cases to help authors. The section for authors points to the publishing houses general guidance for review practices and a page explicating the scope of journal and which data are eligible for review.
Earth System Science Data	This journal's website includes extensive description of the submission process including links to external standards and checkers. They also include guidelines to proof-reading and copy-editing and lists of criteria for various aspects of the publication process including a breakdown of the review guidelines.
Ecological Archives	This society-journal has a society-wide data policy, and their website does not have a specific section for the data journal. However, they do include a frequently asked questions area within the data policy section for authors. The section for reviews includes a multi-page document that covers reviewer guidelines for multiple of the societies' journals and is not specific to dataset review.
Scientific Data	This website includes a section for authors with submission guidelines and separate information about data description and manuscript templates. Their section for reviewers is general and does not include any checklists or links to support tools.
State Politics & Policy Quarterly	This website includes a section for authors that includes a codebook guide, a data construction guide, and frequently asked questions. They also include instruction for peer-reviewers that also incorporates frequently asked questions as well as a how-to guide.

Levels of Data Availability

Data cannot be subject to any level of review if it is not available or accessible at some level: every interviewee addressed how their journal deals with this. Journals had different requirements as to when the dataset must be deposited and at what point it may or must be shared with the reviewers. At one extreme, there was an open data journal that does not actually require data availability pre-review: "For the actual data review there is not a lot of that…We're not requiring the data be made available until after the paper's accepted, the peer-reviewers and the editor may not actually see the underlying data until, you know, the paper's published and online and the data had to be made available." (P-18, Pos. 49). This editor, who works for a society that has multiple journals mentioned that the different journals have different levels of data availability required, "We have a journal that has an actual data policy, two journals, that require all the underlying data. We have one journal that does not. And then we, but that is actually our journal that publishes data papers. Call that interesting, right?" (P-18, Pos. 29). Later contradicting herself, she added that, "the metadata is available for [reviewers] to mark up as they need to and give full-reviewed comments. And all the data files are also available for them to review." She critiqued the journals' complicated and somewhat contradictory policies herself, explaining the inconvenient knock-on effects in the workflow. For one, because they don't require the data until after acceptance, the author is often, "scrambling to get their data in a format that is easily readable, that a data center with a strong curation will actually accept."

At the other end of the spectrum, some journals require that the data be deposited prior to submission as it is included in the submission package (P-9, P-18, P-4). One of

these editors noted that part of their vetting process includes looking for an active permanent identifier (DOI) for the dataset—by the time the manuscript makes it to the editorial office, about 20% still do not have a DOI and because the manuscript must pass this test, the journal then has to wait until it is deposited which "can add months to the process." Another added that if it the data is not already in a "recognized digital repository," it is her job as editor to reject it indicating that there are varying degrees of leniency across these journals if a submission comes in without free, or openly available data. Some journals however require the dataset at some point after submission, usually after the article or manuscript has been accepted, "So once the article has been accepted, we instruct our authors to upload their data" (P-1, Pos. 17). This journal requires the submission of data and code for external review through a third-party verifier. One editor added that because funding bodies are requiring deposition, they are getting less pushback from authors about requiring data submission although they did not have a specific data availability policy: "We just try to do as much as we can to encourage people to deposit their data and… we've moved from accepting the dataset within the paper itself to requiring that it be deposited in some repository." (P-20, Pos. 17).

One editor was critical of other journals that choose not to require or check data availability prior to peer-review: "And then really think about how you're going to actually do these checks before peer-review," going on to say, "Obviously, asking for data after peer-review still allows you to release it alongside the final publication, promoting reuse and enriching the final paper. This is the dominant way that data sharing is required by journals, and it's extremely important. It's just not useful in terms of enriching the peer-review process." (P-11, Pos. 89).

The comprehensiveness, completeness, or rawness of what data is or is not required to be shared varies as well. As one mega-journal editor put it, "you have to share all of the data. And now this is, of course, open to some interpretation." (P-11, Pos. 16). His journal included deposition in the right place as one of the criteria for review and that they specify that the data described in the manuscript should be provided "down to a raw measurement level." The other mega-journal editor added that each of their publisher's journals has a different data policy at that they range from "suggestion to requirement, and there's different iterations and different types, and within each sector" (P-20, Pos. 17).

Rights and Licensing

Levels of data availability at these journals is also shaped by restrictions that come with the data including copyrights, licensing, and proprietary data. Some journals do allow caveats to their open data availability requirements, others do not. Two journals that engage in dataset replication using a data curation team manage the tension between their goals of open access and replicability with data restriction in different ways. One works closely with authors who have proprietary data restrictions, first asking them to get a confirmation from the data source before transferring it to their third-party verifiers after which it is not made publicly available. This means that proprietary data is subject to review just like other papers but in some instances, they cannot get confirmation at which point they still allow the author to publish the work but have make the authors write and include, "a detailed note of how they gained access to the data, their process for constructing the data in the hope that in the future if there are interested researchers or readers that they are able to follow those procedures and produce the same data." (P-5,

Pos. 25). The other journal maintains that the standard for accessibility applies regardless of proprietary formats and will, in fact, reject on this basis: “We will occasionally reject for a technical reason. Look, you came in with ARCGIS. It's entirely proprietary... it might be a good product but nobody else can use it because it can't be shared, unless you buy this license…” (P-18, Pos. 55). However, the same editor acknowledged the need for flexibility in their policies and standards regarding the goal of open access through open licensing:

> *“We keep the goal high but the real world, has somewhat eroded [it] or we always find exceptions…One is national policies: a set of researchers from the UK will submit their data and everything will look good. We can read it. It's useful…but we find out that their data center requires registration, and we say, ‘Whoops, that's wrong, we can't have that,’ and they say, ‘Well, it's not us not up to us as the researchers, it's a national policy…’ So, so we still hold that initial policy—free and open access—primary but we also admit that if you're trying to get data from [other places] we understand that there's going to be a registration step. We understand that they tried to keep it neutral and anonymous.” (P-18, Pos. 36)*

This flexibility does not hold for at least one other journal that, by the editor in chief’s account, regularly is rejecting submissions due to commercial data restrictions. The editor acknowledged that despite the possibility of those restrictions being “inherited, maybe it's not the author's fault…but it's still a problem for us in terms of what we can do.” (P-11, Pos. 70). So, the primacy of data availability not only for the purposes of the review process is important for this journal but also the availability of the final data paper for a broad audience. The editor summarizes their approach thusly:

> *“We're not just sort of an open access journal. We're an open data journal as well. So, one of the first thresholds before peer-review will even begin is that you have to really be agreeing to release your data openly. And we actually… have a team that checks licenses, so we don't allow commercial restrictions — it has to be open in the, you know, the stricter sense of that as the community, the research community understands it.” (P-11, Pos. 15)*

A different editor echoed these sentiments when speaking about her journal's commitment to openness: "If you want them to publish your dataset, you want to publish it properly, which means making it open access. Nothing puts off data users than having to go and register for access to a dataset, especially if they don't know if it's the dataset that they'll actually be able to use or not." (P-4, Pos. 62).

For this journal, the focus on open access including licensing extends to proprietary formats—a concern brought up by three editors independently—showing how the open data, open access, and open-source movements are often interwoven in these journals' data policies and practices. As one editor put it, "If it's not in an open format, I'd say go stick it in an open format because again, longevity. If it's locked into a proprietary format, you either need to provide the software that will open it, which generally won't happen, or an emulator or some way of converting it." (P-4, Pos. 62). This focus on the connection between long-term data preservation and data availability came up in only one other interview, "In truth, I've never had a situation where someone said, 'Ah, the data is no longer accessible' except in situations where there was some temporary screw up and things are slightly gone awry." (P-9, Pos. 78). This acknowledges the challenge of trying to anticipate long term usability and even the instability of the idea of interoperability.

Sensitive Data

Another impediment to open data availability that three editors surfaced was sensitive datasets. This is a complex issue that each editor clearly felt differently about. At one journal, the editor viewed dealing with sensitive human datasets as a direction in which their policies and practices might expand and grow: "What we're moving toward

is: if you can't provide the dataset, you need to state explicitly in a letter why you cannot share it, like patient data…" (P-20, Pos. 17).

Another editor came out with a more universal stance that perhaps these data cannot or should not be reviewed at data journals: "Another thing that I would ping back is…if the dataset was restricted and people wanted to publish it. I'd want to know why they wanted to publish it and why they couldn't make it open. So, it might be the case like if it was sensitive or commercial data, then they wouldn't want to make it open. In which case, you'd be looking at them going, 'Well, why do you want to publish it in the first place?'" (P-4, Pos. 62).

The third editor to bring up this issue expressed a high degree of ambivalence about the status of the journal's handling of sensitive data and how it might evolve. Like some other editors' responses about proprietary data, he mentioned that they try being flexible enough to accommodate the needs of authors with sensitive, human-derived datasets. However, he also added that, "We're constantly thinking about how we handle human-derived data, sensitive datasets" and that it requires more work in the form of conversations with authors and dealing with the technical challenges of guaranteeing anonymized access to reviewers of these data. Their journal was, at the time of the interview, taking the stance that, "(If) the data are so sensitive that our peer-reviewers can't see it—we won't touch it. That's…challenging for us because it means we turn away important clinical datasets, but to have a data journal that's publishing datasets when no one has put eyes on the data—that makes me very, very uncomfortable." (P-11, Pos. 34). In fact, this editor went so far as to warn against trying to peer-review sensitive data for other journals that are considering it, "We are trying to peer-review sensitive human data,

so it's definitely doable. I just wanted to point out that it is very challenging, and there are much easier steps that most journals can adopt, which could have a huge positive impact on clinical data sharing." These recommendations from the editor require authors to, "describe how others can apply for access to the data," like the procedures described at another journal and how they deal with proprietary dataset issues. The other piece of advice he had was that public data use agreements might be an easier thing to check for and that that alone, "would be a sea change in clinical research." (P-11, Pos. 90-91).

The other ethical issue with data availability policies that came up, not only in this interview but in one other, was how to maintain anonymity in the data review procedures. This is one of only two places in which data security or data sovereignty was alluded to by any of the respondents. This editor draws out the tension between requiring data access and protecting anonymity, the degree of openness required for review with the degree of closedness required for ethical practices:

> *"If we want a data journal that publishes data papers we need someplace where we can host the data that reviewers can get to it anonymously. And that becomes a problem for us because if the author hosts the data, we can't ensure that a reviewer reaches it anonymously and different data hosting sites all have different policies, and we can't police all of those. And when the data is too big, we can't host it ourselves, so we end up in a little, you know, a little problem. We've had a Google Drive setup which will hold some data, but the more data papers we receive we're going to hit our 15-terabyte limit at some point and not be able to host everything that's in peer-review at any given time." (P-18, Pos. 93)*

In this quote there is a strong overlay of the technical challenges of working with datasets as an object of review. Other data-specific concerns like the degree of integration with data repositories, managing growing file sizes, and the importance of reliable digital

identifiers in tracking a dataset throughout review are further explored in subsequent sections.

The other unique case that came up in the interviews regarding data privacy and security was at a larger journal that manages their manuscript submission review workflow through a high degree of delegation between different stakeholders. Previously, they had, “an entire team that we, like, call it automated, but it's really their suppliers that work out of India, and they're the ones who are manually pulling the PDFs apart and submitting on behalf of authors…” (P-20, Pos. 37). The editor mentioned that they had recently run somewhat afoul of a data protection regulation in these practices which had prompted them to change their logistics sooner than they had anticipated.

RQ 2: Editors’ perceptions of data review standards

What do data journal editors perceive are the successes and challenges of verifying the quality of research data?

One editor spoke about her hands-on involvement in reviewing data like this, “I tend to do a quick check of the dataset in the repository…we had a case where somebody submitted a data paper and I said, ‘You need to upload the dataset to a repository somewhere’ and they uploaded the dataset, and it was a scanned copy of their MSc thesis. So, I bounced that one, unsurprisingly—I do check on these things.” This is not the only example of an editor personally reviewing a dataset to check its compliance with a given journal’s policies. Still, it is an excellent example of a few related themes that came up in these interviews when we asked how review standards work for datasets: how strict or to what degree is data reviewed?, who is reviewing data?, what criteria are datasets being held to?, and what—in fact—is a dataset that is worthy of review through the eyes of an editor? Even when we specifically asked respondents about their journals’ data review

policies, they often took quite a lot of time and care to explain and clarify whether and how their publication uses standards for review. These issues, and consequently our research questions, are inherently connected especially in editorial work.

Scope of Data Publication

An important theme that emerged was editor's trying to define data eligibility for review and the scope of relevant datasets for each journal. The respondents offered different definitions of what constitutes a data paper. They often define their scope of publication with respect to other publication models; scoping can engage several different types of stakeholders. For example, at one journal, the editor describes their review process as "multi-layered" but "relatively typical for journals that have academic boards and active in-house teams." Here, for example, the editor in chief ultimately has discretion overboard members if something is "really outside the scope" (P-11, Pos. 48).

The locus of power in determining the scope of review is somewhat flexible but usually centers on the editorial office with most labor for actual application of standards being contributed by authors themselves, data curators, peer-reviewers, and, in some cases, publication staff. For another editor, he views the role of the editor in defining scope as not catering to any particular disciplinary perspectives or approaches but rather, "to actually to have an eye on the wider, those bigger picture issues and not just sort of thematically, maybe, what's correct in, say, lepidopterology research, say someone working on butterflies, often has a particular perspective on things and I need to kind of broaden that out. (P-9, Pos. 83)

"Is it data?"

In interviews with each of six non-political science journals, the editor tried to define their data paper with respect to data itself. The other two political science journals both engage in data review as a computational replication and reproducibility step that is appended onto what is an otherwise standard peer-review process. Their papers are topical, narrative papers, not data-focused data papers per se. However, most of the other journal editors we talked to do review data papers and so the question of "what constitutes a dataset?" came up in editors' responses to a variety of our interview questions.

Editors' descriptions of how they scope their data papers are sometimes in contrast to other things. Per the example that this section began with, a scanned copy of a thesis is not in scope for that journal and that case serves as an edge case or boundary case to explain what does qualify. Another editor of a multi-disciplinary mega-journal contrasted his publication's scope with the "just traditional research articles" that other journals publish. For him, his journal has a different mindset, a different expectation, "You're not just writing a research paper where you're giving just enough information for someone to understand your conclusions. Ideally, you're providing enough information for someone to be able to really use and actually…reproduce all your data processing steps." (P-11, Pos. 79). The same mega-journal editor explained that "The advantage of data journals is, 'what the data are' tends to be better defined because the *paper is about the data*." (P-11, Pos. 16). For this journal, the primary scope criteria are that "We are reviewing a manuscript about data."

Another editor, this one at a disciplinary journal, also defined their scope with respect to "research journals" explaining that they are not an analysis journal and that, "(the) distinction between how a research journal reviews its papers—and I review a lot for them— and how a data journal reviews its papers, its products…is crucial." (P-18, Pos. 24). This statement implies that what defines the scope of a data journal is not *what* they review but *how* they review. Then, the editor immediately went on to define the primary data scope criterium for the journal as, "we (…) only accept datasets." In another instance, a disciplinary society-funded journal defines data papers as those that are "meant to be just data" (P-18, Pos. 33), according to its editor.

Beyond this "just data" criterium, two editors mentioned that they ask whether the "dataset exists" or even "Is this data to start with?" as part of the review process amongst editors and reviewers. As one editor pointed out, "some datasets are essentially just textual descriptions (…) of a dataset that might exist in somewhere like Dryad, for example." (P-9, Pos. 44). This editor referred to datasets repeatedly as having a stable standard of minimum viability for publication, though did not define this concept with much granularity (P-9, Pos. 78). Perhaps this is a version of a criterium that another editor summarized as: "We just make sure that it looks…like science." (P-20, Pos. 81).

Adherence to standards: flexibility, strictness, and in-depthness

Datasets being heterogeneous and data publication being context-dependent means that there must be flexibility in adherence to standards for review. The in-depthness of review and the strictness of the standards application by whomever is perusing the data varied across our sample group. As previously mentioned, one editor described their standards as having been "eroded" by the real world and that exceptions

for any possible rule mandated flexibility, not only in specifying technical requirements for authors, but that reviewers must remain flexible as well (P-18). This is particularly acute for this journal as they set a relatively high barrier to entry with respect to the level of attention to detail required of their peer-reviewers:

> *"When we review a dataset, we ask the reviewers, and these are very hard, to find to actually, we call it test drive, the dataset. We want them to open it. We want them to download it. We want them to look at the format. We want them to try and reproduce figure three. We want them to look at the uncertainty terms. So, the point is that when you, as the next user in sequence, decide that you want to use that data over North America over Europe or globally, whatever it is, you don't have to do all those tests, right? Because you trust the journal, just like you trust a research journal, to have done the quality control. So, we do a very high level, very tough test drive of these datasets." (P-18, Pos. 24)*

Two other journals have a review standard of computational verification but unlike this journal, which specifically asks their peer-reviewers to be data curators, the others use a third-party in a two-step process. The editor of one of these journals summarized it thusly, "Prior to publication, all the code and data is analyzed by that third party for verification purposes and then is all deposited onto Dataverse. As far as I know, there are only a couple other journals that engage in that kind of verification." (P-1, Pos. 5) The other such journal in our sample describes the process as one in which, "…all of the data files used to produce the analyses in their article as well as all the data replication code that they used to produce those analyses…are verified and those code files are replicated by a third party entity…until they are successfully replicated and verified" (P-5, Pos. 21).

By contrast, several of the other journals leave the degree of in-depthness and the strictness of the review of the dataset up to the reviewers. One of these journal's editor stated firmly that, "We are reviewing a manuscript about data. You can call that data review if you want—and I think that that's a useful shorthand—but we do not treat our

peer-reviewers like data curators. If they want to look at the raw data, they can." The same editor also clarified that, "…our experience shows quite, quite clearly that reviewers will engage with data—we don't require that they do—we just make sure that it's accessible to them. And that it uncovers substantial issues, in a small number of cases, that would not be caught by traditional peer-review." (P-11, Pos. 46). Another journal asks reviewers to look at the data but not to do something as in-depth as a test drive: "It's not kind of getting in depth and trying to recreate the analysis or recreate the dataset themselves. It's more kind of looking at it to see if it's plausible, if it makes sense, if it's adequately documented—that sort of thing." (P-4, Pos. 30), the last clause perhaps putting an emphasis on a movable adequacy benchmark. Finally, one editor described journal's standards for scientific review as being aligned with traditional research journals:

> *"Well, I'm going to say in our case it's not actually very different. We publish peer-reviewed articles that are your standard article with an introduction and methods and materials and discussion. For our data review policy, we ask our reviewers to be just as strict as they would with one of those research papers and give it a full peer-review process…the metadata is available for them to mark up as they need to and give full-reviewed comments. And all the data files are also available for them to review. So, we're really asking for a very in-depth review for our journals that will be as strict as any other scientific review."*
> *(P-18, Pos. 21)*

Observing that reviewers must be particularly adaptable when working with datasets, one editor remarked, "I think there's a bit of an open question that's sort of down to reviewers," while another editor reflected on inherent heterogeneity in how reviewers apply standards, "(it) is always a little bit of a variable thing because a lot of it is down to individuals and their particular perspectives." (P-9, Pos. 49). The subjectivity of review again came up in a different editor's account, "Research is hard, I mean, there's no way

around that. And no one likes to be rejected. But we try to make those rejections to be based as much as possible on objective grounds." (P-11, Pos. 71). Perhaps the tension between reality and ideals, and the fundamental pragmatism of a project like data review and publication, if it can even be called a cohesive process, is summarized best by the following excerpt:

> *"The ideal is that you have a fixed standard, and all datasets meet it. But the real world is that everybody's dataset is different, has different features, is intended for different users. So, it's not…possible to have one standard that meets all."*
> *(P-18, Pos. 47)*

Reflecting on Quality

One editor, speaking about the variability of datasets and what might be eligible for review made an interesting comment: "The original view of (our journal) …was that basically: all data is good data and that there pretty much is no minimum viable publishable unit. Basically, whatever the author determines that to be. But that is not a universally held view." (P-9, Pos. 77). Here, the locus of power of who determines the scope of review has shifted uniquely to the author. In trying to define the scope of review here, she opened a new conversation about the worthiness of data for review: how good—and by extension—how bad can data be and who gets to make that call? There were only a dozen mentions of the term *quality* across five of eight total interviews. So, what do these editors talk about when then talk about quality?

Two journal editors, spoke about quality in terms of control. In one instance, it referred to the cleaned and collected dataset that might be submitted journal, offering the example of: "The historical work or historical record of quality-controlled earthquakes around the world." (P-18, Pos. 37). In another instance, data quality control is the internal process of review that a dataset would be subject to after submission, "We're taking a pile

of data that have been let's maybe not deliberately hidden, but certainly not deliberately exposed and now we're bringing them into exposure and usage and quality control…" (P-18, Pos. 62). In congruence with the latter sense, a different journal's editor made a claim about data quality control as one of the core functions of data review alongside data management and other checks (P-4, Pos. 50). In fact, the first editor contended that the data quality function that their journal provides was not only something worth fighting for but something that ought to be, "extended to or be partnered with the research journals, because the research journals are just too busy, and they don't have the time to do this right." (P-18, Pos. 84). This editor also implied in another statement that data quality is akin to the correctness or accuracy of a dataset.

Interestingly, another editor, in describing how their journal's review standards incorporate the quality of data description, indicated that that quality is more akin to completeness or comprehensiveness than correctness or accuracy. "Data quality" is included in the instructions for reviewers at another journal alongside the following criteria: "Metadata presentation", "Metadata completeness", "Data organization", and "Data integrity" as well as clarity in methods and appropriateness or correctness of study design. This is not the only instance in which data quality is somewhat nebulously defined in these editors' accounts. Perhaps, then, data quality is a more subjective category. Some editors certainly seem to think so. One multi-disciplinary journal's editor answers the question of what data quality means head on:

> *"Peer-review, generally, for a paper can generally answer the question, 'Is this a good paper or not?' to a first approximation. Data is trickier because you can say 'Is this dataset of good scientific quality?' and it's a case of, well, what does that actually mean? (...) So, the quality of the dataset really depends on what you want to use it for. That being said, you can definitely say that a dataset is bad if you can't actually use it, right?*

> *So, it's not so much about, to a certain extent, it isn't about the data itself but also, it's more about the quality of the metadata, the completeness of the metadata, the understandability of the metadata."*
> *(P-4, Pos. 26)*

This model of data quality as fitness for use is important for this editor because she has experimented with reviewing and reusing datasets and in doing so, she found that documentation and explanation about the conditions of the dataset were the most important factor in anticipating reuse and instilling confidence in reusers. Another editor also challenged the issue of data quality that, even within this journal's relatively niche discipline, does not have a shared definition:

> *"We within in our community get very obsessed about quality: data quality. It is exceptionally difficult to do…But what we can do quite reliably is document the process that's been attached to our data and by flagging that, we can then be more transparent temporal users of our data—about what that data might be fit for."*
> *(P-9, Pos. 91)*

Another editor, who acknowledged the impossibility of universal standards offered, instead offers thoroughly documented uncertainty terms for analyses as a proxy for good scientific quality. He claimed that this helps his journal not only uphold a philosophical commitment to standard error-correction but that it practically informs datasets reuse conditions (P-18, Pos. 47).

Usefulness and Usability

All these journals have a predisposition towards dataset usefulness, utility, or usability and reusability, whether for future users or to enable the data review process in support of other goals like computational replication and reproducibility. Will the data be useful in the future? These editors address this in several ways.

One editor stated that they are moving to a more utility-centered review policy on the heels of overwhelming volumes of submissions despite their previous policy having

been, “We'll publish data…as long as it's reproducible and it's valid. It's not under our jurisdiction to decide what is and is not useful to other people.” (P-20, Pos. 81). For one editor, the format and rawness of the data are “equally relevant” when considering reusability (P-11, Pos. 61). This editor and another both at different points described their internal monologues when considering usability, one asking herself, “Could I use this dataset again if I wanted to?” (P-4, Pos. 26) and another remarking to himself in the review process, “That could be really useful to someone.” (P-11, Pos. 79). Sometimes the scope or granularity of the data throws its utility into question. An editor had recently received a dataset for review that he determined was not sufficiently useful for the target disciplinary audience and used this as grounds for rejection:

> *“I was dealing with that quite recently where actually it was being targeted at macroecologists, and I really felt the data wasn't sufficiently useful to macroecologists. I think, probably in the state that it was I would say vaguely interesting but was pushing the mark. And actually, this is the case where the handling editors had already accepted it. So, I always say [rejections] tend to be more about disputes about the viability of the data.”*
> *(P-9, Pos. 78)*

Another editor also pointed to tools that might help future users discern whether a given dataset might be of use to them including data visualizations like quicklook plots (P-4, Pos. 66).

Data Description and Interpretability

All these editors discuss the necessity of sufficient documentation, description, and context. This is in the service of interpretability and understandability for both data review and potential data reuse. This includes the comprehensiveness and completeness of dataset description as well as annotations, details, different types of metadata, and in some cases, specific inclusion of fields, titles, and units for submitted datasets. In fact,

one editor makes a pitch for data papers and data journals as a superior way to capture these types of descriptive information in ways that they think will make data more useful:

> *"Data articles are really, really good ways of capturing important metadata that doesn't necessarily get caught by the standard metadata schema that a data digital repository would request. So, you get more with the data paper, you can actually have the authors giving more of the story about how the dataset was created and the interesting things that went wrong while creating this and you get more of the context, and you get a richer context. Whereas, if it's just kind of the metadata schema for the repository: name, title, latitude, longitude, temporal extent, that sort of thing…you can miss out on an awful lot of detail, like calibrations and stuff like that."*
> *(P-4, Pos. 58)*

One editor described their review standards as being very simple: "…we want to make sure that everything in the codebook is documented and documented accurately" (P-1, Pos. 41). The editor freely acknowledges that this standard exists because there are persistent errors in submitted datasets including rounding errors, missing minus signs or coefficients, and "loose code" that authors might be able to catch in a proof-reading process, but he was sympathetic that, "Sometimes when you're working on something, your code is messy. And so, this [data review] makes it much more efficient and cleaner." (P-1, Pos. 49). Their review process often takes several rounds, something that several editors also remarked upon: "all of our manuscripts go through a single round of revision and very common things are people are asking for more details in terms of methods" (P-11, Pos. 75). This initial lack of clarity and the power of the data review process to catch errors is a common theme across several interviews. For one journal, the role of reviewers and editors is not simply catching typos but also to verify that nothing is missing when comparing the standards for review against the structure of the data paper (P-18, Pos. 81). Sometimes authors simply do not provide enough description or

metadata, as one editor bemusedly remarked: she felt that although good documentation in general makes data review easier, the inclusion of any descriptive information at all is good compared to a single sentence, something she had recently encountered in her editorial work (P-4, Pos. 42).

Another journal editor describes the standard for documentation as, “Basically they want to see how it was collected. They want to know the parameters around which, you know, like what's the indoors, outdoors, elevation, temperature…” (P-20, Pos. 81). “They” here likely is likely peer-reviewers or editors as opposed to data curators although this is unclear from context. One editor defines the scope of their data paper as “data-rich” with attendant metadata that clearly describes the data and provides information about how it might be used (P-18, Pos. 33). Another journal that similarly describes their data paper structure as being primarily a dataset also allows “a little bit of explanation or demonstration of how the data is relevant to research” (P-18, Pos. 24). Necessary specifications and parameters that these different journals require include descriptive data titles, README files, tables of contents, .do files, codebooks, variable names, data summaries, which software version was used, to what degree the data has been processed or cleaned, and clearly marked units of measure or analysis (Ecological Archives, P-4, P-5). This metadata might be included in the journal’s submission template or online manuscript management system or as a separate document in PDF format. The metadata themselves sometimes have their own criteria for review including completeness, quality, and understandability at one journal.

Data Accessibility for a Broad Audience

Only two editors brought up the topic of accessibility of language and research for a broad audience. Both are at disciplinary-focused although one in social science and one in the natural sciences. The first states the journal's goal is to make research, "…if not accessible to a layperson, at least accessible to a moderately advanced political scientist" (P-5, Pos. 45). The second explained that their journal intentionally edits submissions for usage of technical jargon as they hope that the writing will be for a broad audience. Still, this editor was insistent that they do not reject because of technical language usage or non-standard English submissions that might come from non-native speakers, explaining that "We… try and solve the language problems in the interest of getting that data available, rather than rejecting on the basis of the language problem." (P-18, Pos. 59). In general, these statements imply that the journal aims to make the underlying data as accessible as possible to as broad of an audience as possible.

Technical Criteria

We did not specifically inquire about the technical challenges of reviewing datasets. All eight editors interviewed surfaced technical data issues. Four of the eight editors brought up data's technical dimensions as a as a factor in the review process explicitly. Only one of the editors we interviewed mentioned the FAIR principles. Their disciplinary-journal incorporates these technical specifications in multiple ways (P-9), while another editor mention machine-readability as a review criterium (P-4).

Another editor talked about technical challenges and the importance of having technical data management expertise and capacity to run a data journal, expressing the ongoing challenge of how to, "keep up in terms of your administrative processes and

your technical expertise?" (P-18). He spoke about how much their journal has learned about processing different formats and how they have grown and adapted alongside technical standards. Both multi-disciplinary mega-journal editors that we interviewed have in house processes to check the technical characteristics of the datasets they review. In one case, they have a section of the author submission system called "technical validation." They have a team that checks licenses and availability as well as a data curation editor or manager that facilitates relationship with repositories (P-11). In the other case other, the data, which is compressed and attached to the manuscript package in their proprietary workflow management software, is checked "just to make sure it conforms to the standard format" and a technical check step is performed by a publisher team member in India with the title of publishing content specialist (P-20).

In the following subsections, I explore the technical criteria and standards that different journals employ in the pursuit of data accessibility and availability of data. This includes the ways standards are and are not used to manage different data files and formats as well as how to reliably identify datasets.

Formats and Interoperability

While some issues of format and interoperability that editors brought up were addressed in the previous section on licensing and rights, mentions of dataset formatting came up in myriad ways throughout these interviews. For some editors this was in discussions of dealing with data file formats including proprietary formats, software versioning, and the contemporary best practices and affordances for preserving databases and other updatable datasets as static files. This included specific mentions of data formats, languages, and technical standards (e.g., Excel, NetCDF, CSV, PDF, GIS, SQL)

and how these characteristics of data are or are not incorporated into the data review process or standards. Editors also discussed file size challenges.

Software compatibility, processing speeds, and computing requirements can pose technical challenges for different stakeholders. Sometimes issues arise that are local, for example, one editor cited compatibility issues that might be just, "specific to the disk drive that's not more generalized" (P-1, Pos. 49). Another editor connected that these types of issues more directly with their impacts on the external data curation team that their journal partners with:

> *"I'm sure it's a challenge for the third-party verifiers when you have authors who are using special software programs that aren't commonly known and accessible or maybe RAM size or computer speeds that are necessary or perhaps different computing networks or structures that are necessary to even replicate some of these advanced analyses."*
> *(P-5, Pos. 45)*

In another instance a different editor surfaced a similar data review standard for their peer-reviewers: "Is there software provided to support other people using this data? It's even simpler things like: …Is it Excel formatted or is it comma separated variable? So, is it open-source format or not? Is it a standardized, community specific format or not?" (P-4, Pos. 54). Another editor cited their journal's requirement for comma separated values (CSV) versus Excel and specified other examples of unacceptable software or formats in their interview because they are not open source. Their standard is NetCDF versus MATLAB, they prefer open GIS to ArcGIS, and are trying to transition to mySQL as an open database format. These standards arise directly from a consideration of potential end users as the editor emphasized when he walked through a hypothetical example of a user trying to download a dataset, "If a) the link doesn't work, b) the format: they think is NetCDF but it turns out it's not annotated correctly, and then

it turns out it's not gridded correctly and they can't actually bring it into their model. Then…it's not a useful product" (P-18, Pos. 28). This editor also highlighted that their journal's archives will inevitably look different over time, "as data types change, as new tools come into place" (P-18, Pos. 51). As previously discussed, one journal relies on the FAIR principles, but they also incorporate the Darwin Core file standard to guarantee technical compliance of diverse data type submissions as the standard, "imposes some quite rigid structures…regarding the need to add metadata, machine-readability, open licensing." (P-9, Pos. 29).

Relatedly, the primary difficulty that another journal was facing with file formats due to file size. For this editor increasing file sizes is as much an issue of convenient access for data users' post-publication as it about the challenges of finding server space:

> *"The larger file sizes are problematic for people who are, say, out in the field or on a slower connection. And we've found that the publisher actually is refusing to publish larger files. They have such trouble with download speeds and complaints about that on their platform that they tried to implement a 10-megabyte file size limit. And we pushed back very hard because 10 megabytes is ridiculously small when you're talking about files and especially some of the data files out there. So, we're, we're now, we're trying to put up files that are one hundred and two hundred megabytes, to see how many complaints we receive. And if we receive a lot of complaints, we're going to have to tell those people to go to an outside repository that can handle files of that size, like KNB (Knowledge Network for Biocomplexity). We had to send someone there because they had two terabytes of data."*
> *(P-18, Pos. 50)*

The trend of increasing file sizes poses a direct challenge to their peer-review protocol as addressed in the previous section on sensitive data. Server size is clearly a concern for this editor and at least one other who brought up that their journal was, "increasingly having problems of giga-, you know five gigabyte files."

The rawness of the data, in terms of the required resolution, granularity, or degree of processing of the data, only came up in one interview. At this journal, per the editor's

account, peer-reviewers of datasets often ask, "'Can I have the data at a different level?... it looks like you did a normalization, why haven't you shared the normalized? That could be really useful to someone.' or, 'Actually, I'd like to have the data as it comes right off the machine. I know that specialty...'" (P-11). The conditionality of what constitutes a more "usable" standardized format came up in a different context within two other interviews, namely, how to facilitate non-static formats for objects like databases as well as versioning for datasets that they have published previously. These editors used slightly different language to describe what I will call live updating and the challenges of trying to capture, preserve, and store updating or version-able datasets in static formats. For one editor, overcoming the complex challenge of "effective annotation" for "living, breathing datasets" is in service of a more seamless, interconnected user experience (P-9). The status quo for annotation and live updating requires publishing it anew, duplicating the amount of work for authors and reviewers alike at both these journals. One editor proposed increasing efforts to link data using structured data and two-way interaction. Another journal has approached this challenge with a snapshot in time model for database archiving but acknowledged that this is an imperfect solution that "doesn't emulate all the features" and that this does not help them update the publication as the database evolves post-publication (P-18). Both editors were trying to explore and solve a problem for the benefit of the broader scientific community.

Identifiers and Findability

Versioning and updating dataset identifiers are particularly applicable for database formats with version updates. As one editor described their workflow, each time something is published or updated, "You still do a DOI captured snapshot of each one"

(P-18). Just as formats and interoperability are a key component to accessibility and availability, so too is the characteristic of findability. This is mediated for these journals through a variety of identifier tools and standards like DOI, HNDL, and permanent URLs. The theme of identifiers and findability also came up in references to links to data, bidirectional linking, dead links, data citation, and descriptive supporting information. At its most simple, a reliable dataset identifier can guarantee that, "there's a path for reviewers to get to the data." (P-11, Pos. 47). In total, six of the editors directly mentioned that their workflow engages a permanent dataset identifier at some point. The high degree of repository integration of the other two journals implies that they may have similar requirements as a step of submission to the Harvard Dataverse. One editor lamented the inflexibility of using permanent identifiers in their workflow:

> *"An open DOI needs to follow that whole process because if you're a reviewer and…you click on the DOI and it fails, then immediately you give up. 'Why am I bothering with this paper if I can't see the data? So, and that's a tricky bit because some data centers don't want to assign a final DOI until the paper is already published and, in some cases they're happy to submit the DOI, what if the there's a substantial revision? You have to go back…to the data archive, establish a new version, get a new DOI."*
> *(P-18, Pos. 44)*

The power or capacity of any permanent identifier is perhaps stretched to its absolute limit in one journal's unique workflow. Although their general policy is that every submitted dataset is subject to peer-review, they have a separate policy for papers that are co-submitted alongside original research as this editor manages multiple journals at one publishing house. If the original research has already been accepted, they forgo separate peer-review of the dataset, relying instead on "the knowledge that this has already been more or less peer-reviewed" (P-20). This level of trust in a different journal's workflow is represented in the instantiation of that paper's DOI in that the

editors, “don’t necessarily send it for peer-review if they feel like, ‘this looks good, this was published in [*journal name*]. I trust this scientist, obviously that paper has already got a DOI, it's already fine.’” (P-20, Pos. 9). After the dataset is published, it gets a separate DOI with bidirectional linking.

Reviewing to Different Ends

So, if data can be reliably tracked through the review process using standards for identifiers, if it is in a format that allows it to be reviewed, and it is within scope for a journal, what other standards might apply for these journals? Anticipatory reuse, utility, interpretability, and broad accessibility matter to some editors and not to others. Data quality, per these interviews, is a desirable but elusive characteristic. Quality is usually identified as being context-dependent or related to fitness for use, maybe to the extent that standardization is functionally useless. Still, there are other characteristics that these editors spoke about tied to the different goals and uses of dataset as a scholarly object. Some of these editors try to incorporate review criteria at their journal that operationalize these other characteristics.

Data Novelty and Originality

Novelty or originality is a metric that only briefly came up across five of these eight interviews. The primary commonality amongst these comments is that application of any standard for novelty would be performed by an editorial office or peer-reviewer but not by a data curatorial team member, whether internal or external. Two editors stated that their journals try to be even-handed in assessing the suitability of datasets for review vis-à-vis the dataset’s potential to make a significant contribution, both contrasting their practices with other traditional publication approaches. One editor puts it like this:

> *"We do not apply an impact threshold. So, we really tried to be as open as possible to small and large datasets. I think every journal, you know, and our board, there's going to be an implicit assumption that whatever is being submitted needs to be, you know, meritorious enough to deserve peer-reviewer's time…one of the goals of a data journal is to publish what people call 'negative data'. At a proper data journal, in my view, there shouldn't be anything—that shouldn't even be a word—because you never have —there's no interpretation of whether it was negative or not."*
> *(P-11, Pos. 62)*

He later added that, "We actually asked people to remove amazing new findings from papers" (P-11, Pos. 62). However, this editor also lamented having to reject important datasets because of sharing restrictions, indicating that either he or the journal have somewhat conflicting views of whether impactfulness, importance, or novelty should be part of the review criteria. The other editor broadly writes novelty off as something dependent on the view of the associate editor reviewing a dataset and that the journal tries not to engage in novelty assessment. However, when they do, he as the editor in chief is the one to do it, "I intervene when there are questions about novelty or utility or... that sort of minimum viable set. Interestingly, occasionally, it can go both ways. More often than not, my role as editor in chief is actually to explain that scope often to editors to allow them to be a little more lenient. But just occasionally, it goes the other way." (P-9, Pos. 78). This hints at the complexity of trying to define any minimum impactfulness standards for datasets as objects of review.

One editor works at a journal with separate review policies for original research papers and data papers but briefly mentioned that originality is also part of the editorial process for the datasets contrasting this with validity, utility, and reproducibility (P-20, Pos. 9). Another editor referred to scientific novelty as something that is assessed by dataset peer-reviewers alongside plausibility, error-checking, and catching falsification (P-4). Finally, one journal editor who manages different journals with varying data

availability requirements and standards for peer-review listed all the criteria that peer-reviewers are instructed to examine for data papers: novelty or impact is not on the list, although data quality is (ESA).

In a final somewhat separate example, a different editor who did not otherwise bring up novelty as a standard of review implied that he perceives an incentive on the part of authors and on the part of editorial teams and reviewers to increase the visibility of particular types of datasets using publication. He cites a hypothetical in which, "a graduate student works (…) finds a long record of something in China, trees, meteorology, I don't care, says, oh, this would be really cool data to submit to the community…" (P-18, Pos. 59). Importantly, he emphasized that the journal would invest significant time and effort to help the student make those data available even though it's, "really a moving target and if we're opening up access to some of these long hidden or even forbidden datasets, then that's very positive for the community and for the journal." (P-18, Pos. 60). This willingness to invest precious resources into what is a potentially unique, new, or impactful dataset is framed here as a commitment to a broader ideal of innovation in data sharing through the form of a data paper.

In short, those editors we interviewed that engage with data review of a data paper format this is like a narrative-style analysis paper did not bring up the concept of novelty or originality for datasets at all. This is likely because they employ a bifurcated two-step review process in which the quality of the research output is assessed first and then the attendant or supporting data is subject to review with a complimentary set of criteria applied by a third-party. For those with a less traditional paper, more akin to a pure data paper, some use a review process in which novelty is intentionally not a standard that

they try to employ precisely because they are looking at datasets as a different type of scholarly object that is being primarily reviewed for a generalized reusability. Other data journals that look at novelty alongside other review criteria use a combining peer-review step.

Data Soundness, Veracity, and Rigor

Many other metrics came up related to whether the dataset is correct or valid, whether it was reasonable or plausible from a scientific merit standpoint, whether it appeared that it was noise and not signal. For example, one editor when talking about noise and signal takes a clear stance that identifying noise in review is not only possible but essential: “If it's a failed experiment, then you're just measuring noise. Well, that's a failed experiment and we won't publish that.” (P-11, Pos. 62). Another editor directly contradicts this absolutist idea of what might be significant for reuse in stating, “I have in the past worked with datasets that were absolutely full of noise, as far as I was concerned, but yet another researcher who was looking at a different aspect of that atmospheric phenomenon that I was studying would think that what I called "noise" they called "signal" and vice versa.” (P-4, Pos. 26).

The idea that soundness, veracity, and rigor are stable concepts is at odds with these different editors’ descriptions of their procedures and standards, even down to the question of determining whether the dataset is measuring something real. Some editors look at correctness or validity as a function of whether the final contents of the dataset match the analytic output—a form of internal consistency—and whether they hold with the experimental design. Making the latter determination, whether by a peer-reviewer or editor, relies on a scientific reasonability or plausibility test that is inevitably extremely

context dependent, likely varying across sub-disciplines, different methodologies, usage of certain analysis techniques in data processing or even the instrumentation used in data collection.

At one journal the technical compliance issues are, “basically handled by the processers and are somewhat invisible, apart from the fundamentals of: ‘is this good, sound data? Is this reasoning upon that data good and sound?’” (P-9, Pos. 53). This indicates that soundness is one of the standards for review at this journal. In another interview, one editor counted his data journal amongst a larger yet undefined cohort of, “sound science journals.” (P-20, Pos. 61). Coming back to the first editor, he later focuses on his role as an editor overseeing the process of review itself and scoping it in terms of transparency as opposed to compliance with specific criteria of review or standards: “We are not responsible for the actual content. We are responsible for the processes by which that data has gone through…that degree of transparency of process.” (P-9, Pos. 91). This was the only interview that mentioned transparency explicitly.

The two journals that outsource their verification function to a third-party look to their processes as a type of check that a dataset is, “actually true in the sense that there is not a result of any errors or accidentally running the wrong model and misspecification” (P-1, Pos. 9). The other added that veracity is a key function of replication but that it cannot guarantee or replace rigor within a dataset:

> *“I think we're trying to guarantee that the results match up to the findings in the tables that are reported in the article, but again in no way can we verify that that was the best way to do it or that they included all the controls possible, that they always support the substantive or theoretical conclusions that authors are drawing. And I don't think that's the end goal anyway. I think that's for the peer-reviewers to do, that's for other scholars to check.”*
> *(P-5, Pos. 29)*

Here, checking data for replicability is a test of internal consistency but not necessary for scientific merit or quality.

Two multi-disciplinary journals take quite different approaches towards rigor, validity, and correctness. One states that one of the most important things that their peer-reviewers do is check for experimental rigor as represented in an annotated dataset or data manuscript. This includes the controls of the experiment that produced the data and enough information about the conditions of its production to show that, "you're measuring what you say you're measuring" (P-11, Pos. 60). They want the peer-review process to be rigorous. They ask reviewers to check that their computational dimensions of the experimental data are not flawed and evaluate them using publishing house's broader "level of rigor" (P-11, Pos. 71). The other mega-journal, other than counting itself as a "sound science journal" is much more laissez-faire in its approach, asking their peer-reviewers to check if the dataset looks valid and that "The conditions under which this was performed seem valid." (P-20, Pos. 81).

Editors of the remaining four journals ask relatively similar, general questions about dataset veracity in their data review practices. For one natural sciences journal, peer-reviewers are directed to assess if, "…the study design appropriate, and correct?" (P-18, Pos. 73). At a second natural sciences journal, the editorial office looks for full uncertainty analysis and rejects in cases where "people just didn't do this correctly." (P-18, Pos. 51-55). A third journal asks reviewers to look for "scientific plausibility" and "if it makes sense" (P-4, Pos. 30).

Data Verificability, Reproducibility, and Replicability

Finally, the practices and values around verifying, reproducing, and replicating datasets are completely interwoven throughout these editor's description of their standards for data review. Six of the eight journals, all except the biological sciences journals, talked about these ideas in their interviews. These concepts are part of the scope of each journal that engages in data review, whether the workflow engages in computational verification (P-11, Pos. 79, P-5, Pos. 29). One editor, after quickly mentioning that they require information about how the dataset was collected and its parameters, his main point was that "The difficulty sometimes, you know, you can't, it's not always... Reproducibility, I think, is tough. Like, you can't just ask an author to re-study something because it's like, this is your data. A lot of times it's negative results." (P-20, Pos. 81). For another editor, the connection between theory, practice, and policy are interwoven through their replication work. The journal requires data replication files for each submitted dataset:

> *"Then there's also the data replication files need to be made in an accessible platform. We generally require them to be made available in either Stata or R. There are others as well, so long as that they're accessible to the general political science community. Those are required. So those are some of the standards. Of course, our verification and replication policies lay it out much more clearly than I've actually articulated here. Essentially, we're asking authors to provide everything that an interested reader would need to follow step by step…*
> *(P-5, Pos. 5)*

At a journal with a computational verification-type workflow, the editor defines the replicability standard as, "We just want to make sure that the results that are reported in the paper are, in fact, the results you're going to get by running author's code with the author's data." (P-1, Pos. 41). The cross-cutting theme here of documentation in service of future reuse, in this case safe-guarded with a built-in replication process within the

journals review workflow, came up in interviews for journals that do not themselves require reproducibility or replication by a data curator either internally or externally. The editor for one such journal clearly states how reuse and reproducibility are connected in her accounting of why description works and how some level of required documentation could prove useful.

> *"You could have a dataset that was absolutely riddled full of holes, but if each of those holes were documented and there was an explanation given for why those holes were there, then that could still potentially be a useful dataset in the future, especially if it's (…) something like historical measurements that we can't reproduce."*
> *(P-4, Pos. 26)*

One editor wanted to highlight some of the same themes as above, noting that despite his journal's emphasis on full computation replication of all analyses and code, "I think actually the data curation, in my opinion, is far more important and far more useful to other researchers." (P-5, Pos. 29). This indicates that he is separating the value of replication workflow as separate from metadata provisioning function that comes along with. Both journals that engage in third-party verification stressed the iterative nature of their work: few if any datasets are verifiable upon submission. The review and replication process are designed with this in mind. One said, "I think there's probably a handful or maybe six or seven cases where authors had their materials successfully replicated and verified on the first try, out of (…) around 200 or so." (P-5, Pos. 57). The other editor affirmed this: "I have not seen any articles come back that haven't had any glitches" (P-1, Pos. 21), going on to joke that even their founding editor in chief had an issue with his initial submission.

RQ 3: Editor's summaries of their journals' data evaluation workflows

What review mechanisms do data journals use to standardize data evaluation?

To examine diverse dataset review and evaluation workflows, we looked at how data journal editors talk about the process of editing, processing, and managing datasets for publication. This can, but need not, include steps or levels of editorial review, data curatorial review, and peer-review.

Policy directly informs the data review workflow for most of the journals but in very differing ways. For one editor, who manages a portfolio of journals subject to different review policies and practices, she explained it as follows:

> *"Especially for the journal with the open data policy, because we're not requiring the data be made available until after the paper's accepted, the peer-reviewers and the editor may not actually see the underlying data until, you know, the paper's published and online and the data had to be made available. Which I, you know, I have a little control issue as a publisher so that gives me pause. I like, I would like to know that it was actually peer-reviewed."*
> *(P-18, Pos. 49)*

This hits on a few key elements that will come up in the following subsections, namely: the enactment of policy in practice and multi-faceted role of the editor in managing the workflow, including reputational management and their trust in other stakeholders. Some editors reflected on their workflows with respect to traditional journals' practices, "just like the very traditional paper, [reviewers] will also submit a text or description of what they think of the manuscript, where the problems were. I think the main difference is that they also, most reviewers do actually comment directly on the paper." (P-9, Pos. 49). This commenting practice in dataset evaluation is enabled by this journal's use of a custom manuscript submission software, which is just one means of technically facilitating dataset review.

Operational Constraints: Time, Money, and Expertise

Each of these journals has operational constraints that can be mostly broken down into the related classes of time, money, and expertise. These factors can serve as rate-limiting steps or bottlenecks in the workflow that constrain not only the existing rate of publication and condition of operational efficiency, but potentially the future scalability of data review for these journals and others.

In total, four editors brought up money or expenses in their description of their operational costs vis-à-vis workflows. These sentiments were all extremely congruent, despite how different some of these journals are from one another. The general sentiment was perhaps best summarized by editor's matter-of-fact statement: "Everybody's busy, funding is tough" (P-18, Pos. 77). Both journals that engage a third-party for data and code replication were clear about the perceived tradeoffs. One explained, "It adds more work, more time, without a doubt there's costs associated with it, financial as well as time costs." (P-5, Pos. 29). He acknowledged these valid logistical and financial reasons that other journals might choose not to use a third-party review step, though he felt that at his journal the reputational benefits outweighed the costs for his journal. The other expressed that sufficient staffing at the third-party data curation center might, in the future, prove to be a rate limiting step though it had not been an issue to date and that the workflow was working well, overall (P-1, Pos. 53).

Managing high volumes of submissions came up directly in two interviews. At one of these journals, the editor stated that this had prompted him to scale up their review operations. This editor remarked that his staff were getting pushed to their limits, "We were borrowing staff from other groups, and that's when we realized, okay, we need to

actually hire people." (P-20, Pos. 101). Even this journal, which is part of a large, well-resourced publishing house is constrained financially; the editor does not have ultimate say in how funds and personnel can be appropriated. In fact, the editor stated that he was struggling with the weight and complexity of managing his portfolio, "I either need fewer journals or more money, or both." (P-20, Pos. 109). The issue of capacity came up in another interview in which the editor ascribed some of the issues she sees at her data journal to broader trends in scholarly communications:

> *"Academic publishing has problems in the fact that we're producing so much, (...) people don't have the attention span anymore, we need filtering. The publish or perish mentality means that there is just simply too much coming out and nobody is capable of keeping on track of it. Which means we're missing things and we're losing things and there's an awful lot of effort going into it, which is not needed."*
> *(P-4, Pos. 50).*

This acknowledgement of the limitations in expertise was personal for one editor who identifies primarily as a scholar and not as a data curator. While he acknowledged that he had adapted to the journal's unique workflow and "learned it on the fly", he generally prefers, "to leave things of this nature to the experts." (P-5, Pos. 29). In another interview, an editor stated that he does not want to entrust peer-review of data to academic editors because they are all so busy. At this journal, they have implemented a socio-technical solution in that they now use an "in-house team of trained individuals" and that they have, "manipulated our submission system in some very, very simple ways that allows people to input information about their datasets before they're submitting. And submission will actually fail unless they have input information about their datasets." (P-11, Pos. 51).

Five of the eight journal editors indicated that having a data review step policy in practice extends the period from author submission to final publication to some extent,

sometimes by weeks. One editor put a number on it: 50 to 75 days on average (P-5, Pos. 45). This delay as well as the time and expertise necessary to do this type of work might serve as a deterrent for other journals adopting similar workflows and policies, according to a different editor, “Nobody else is doing that because it's too hard, takes too much time.” (P-18, Pos. 24)). Embodying the competing demands and pressures in these complex publishing environments, an editor described their editorial review process as “really fast” but then later in the same interview points out that if, “you do one thing wrong and suddenly you've got 100 papers that have been in queue for like months.” (P-20, Pos. 81-97). This consideration of timeliness and rate came up in other interviews.

> *.”. the reason (our workflow) is so fast and because basically it's all structured from the start. So there is never, there's none of this, "I need to upload it to a particular place", "I need to now make that data accessible", or "I need to remove the embargo" or "I need to find out the dataset identifier and “I don't have that": the kind of things you'd normally have with the normal publishing process that we don't get so much with [our journal] because essentially it is all structured from the beginning. At least that's the theory.” (P-9, Pos. 49)*

This editor also considered the effects of making it authors for authors to comply with his journal’s standards, which incorporate community-set standards: “Ease of publishing also requires then some of the like checks and balances that we might like in the community to have, to essentially be absent.” (P-9, Pos. 72). For him, making the data review model scalable might mean sacrificing quality checks, prioritizing data dissemination over in-depth review. The effects of having a data review step in a publication workflow can be mapped in terms of efficiency which may impact the scalability of these models. One editor called for more industrial process in review more generally in academic publishing due to the time, money, and financial constraints and

increasingly high volumes of paper and dataset submissions, “I think we need to move from a situation where it's a pair of reviewers looking at a dataset and writing individual things to a more structured, industrialized process of data management and data quality controls and checking.” (P-4, Pos. 50). For her, to make data review scalable, it needs to become less labor intensive, prioritizing data dissemination over in-depth review.

Conversely, a different editor proposed *expanding* his journal’s time, labor, and expertise intensive model of in-depth data review including computational verification. He said that, looking to the future, expertise might prove to be the biggest workflow bottleneck as he faces challenges in finding enough reviewers with domain-specific knowledge to serve as topical editors and peer-reviewers that also double as data curators, “As an administrator of a journal trying to manage a process, (…) how would you, say, efficiently process these through topical editors, through reviewers, through special issues—all the normal tools that we use—is tricky in the data world because (…) we're just turning over these datasets all the time and new topics are emerging.” (P-18). Aside from keeping up with the shifting technical changes and the growing disciplinary and topical range of submitted datasets, he has found that in more niche disciplines, finding any appropriate reviewer for the data is extremely challenging, exclaiming that everyone with relevant expertise is “already on that paper!” (P-18, Pos. 70).

Another editor mentioned a similar concern about scalability related to expert data curation staff’s capacity to maintain the desired publication rates, here once again prioritizing the in-depthness and comprehensiveness of review. This editor also was very clear that data curation and data review also has to be affordable, in some ways implying

that all stakeholders invested in the potential benefits data review ought to incentivize adoption by other journals by keeping costs down, “So, I don't deal with the budget and whatever else (…) but, you know, to get more journals to do this, it has to be affordable for them to do so.” (P-1, Pos. 53). This editor also proposed offloading some of the editorial responsibilities to the third-party curation teams, proposing a novel socio-technical approach in which the data curators are integrated more directly into the manuscript review software and correspond directly with authors about data issues through the iterative review workflow.

Workflow Management Systems

Although we did not specifically ask in every interview which software, services, or other technical tools are used to standardize or facilitate data evaluation, seven of the eight editors brought it up when explaining their workflows. These respondents were very specific in that each of their journals leverages either an editorial manager software, a journal publication software, or a manuscript submission software to track submitted materials through the review process, though they use them in different ways, and some feel quite strongly about the relative effectualness of these systems in working with datasets.

One editor centered this software as the place where editors and peer-reviewers make decisions about author’s submissions. The third-party verifiers are not integrated into the workflow. This bifurcation of communication in which the author’s communication with editors as well as peer-reviewers happens through the software and the editorial office corresponds separately with the third-party curators, is part of the journal’s policy and the editor feels that the existing system, “adds another level of

assurances and legitimacy to the process." (P-5, Pos. 29). Still, the editor expressed that the existing system might be improved and streamlined if the data curators were directly integrated and that it would reduce the need for the editorial office to do "a lot of handholding" and "shepherding the articles." He added, "It's fine, but it's just time consuming and there's always the risk that something could get missed by the editorial office." (P-5, Pos. 41). The other journal that uses a third-party verifier does have an unspecified "backend" technical publishing management platform for managing submissions, but it is not "through the Manuscript Central or anything else for this process. It's all off that platform" (P-1, Pos. 21-33). The third-party verifiers do not have access to this system and are only ever contacted directly via email by the editorial office. This editor, like the first, felt that it would be nice if the third-party data curation office was more directly integrated into the publishing software (P-1, Pos. 53).

Another journal's editor mentioned that her journal has an integration with at least one outside repository to help reduce manual data re-entry across platforms for authors and allows the journal to provide authors with a custom submission link, which, the editor claims, "really facilitates their submission process and it makes the data process much more streamlined for them (P-18, Pos. 61).

One journal in taking this integrated model to the extreme, has developed a bespoke manuscript submission system to manage data papers. Their templated submission system, which is hosted entirely online—they do not accept Microsoft Word submissions—was developed in consultation and through extensive engagement from relevant disciplinary communities to guarantee that the data platform was compliant with

their standards, according to the editor (P-9, Pos. 95). Putting himself in the shoes of a potential user, the editor explained the system's implementation.

> *"You might choose to publish using a local instance of (…) the Integrated Publishing Toolkit, which allows you to publish basically specimen references and then that data flows through there. Or you might also publish or include structured information about a particular species. So typically, that would (be)…geographic range or occurrence data of collected certain specimens. And again, that would be published in a form. To some degree, the template guides you through, but it will be published within the manuscript as part of (the) paper."*
> *(P-9, Pos. 43)*

This step, as described by the editor, means that standards compliance is embedded within the initial submission. This is an error -checking / prevention step that precludes certain types of issues from arising later in review. From there onward, the submission is reviewed entirely online through the system: editors, reviewers, and authors have the capability to read and write live edits in the document. Reflecting on the system's development and current effectiveness, "In many ways I think the biggest challenge was actually building such a system whereby those processes were built in." He also identified that it might be useful to build in even more services in the future to offer "more content-oriented services that would improve the quality of the data." In a hypothetical example he proposed that integrated data services could help authors identify synonymous taxonomic names (P-9, Pos. 72).

At another journal, peer-reviewers get access to the submitted manuscript and links to the data—which is deposited in an outside repository—through an unspecified submission system. This templated system has, like that described previously, aided in standards compliance and error prevention. One journal's submission system and workflow integration were referred to by the editor as, "a pile of sticky tape because that's literally what it is, because there's no other workflow within [the publishing house] that

even kind of resembles this." (P-20, Pos. 37). When directly asked about the complex technical aspects, this editor laughed out loud and called it, "a lot of systems": the journal has proprietary access to various commercial submission and manuscript software through the publisher that he, the publishing staff, and the in-house curatorial staff use but there is no unified submission service.

Only one editor did not mention whether her journal uses software or tools when asked about what has been most useful in implementing the data review workflow, instead focusing her response on the technical qualities of data itself that facilitate review.

Editorial Role

These editors' work involves several different facets of data review and their reflections on their positions and roles within the policy development, evaluation criteria implementation, and workflow management are deeply intertwined.

Workflow Metaphors

As editor's describe their high-level, supervisory perspective on journal workflows, they employ a range of figurative language that can help us understand the unique challenges and the inherently socio-technical nature of data review work.

Many descriptors were procedural. Their workflows are *multi-layered*, in which datasets and their stewards go through *tiers*, *steps*, or *levels* of review, in which the editor "stamps" or gives "the stamp of approval." Many editors' language focused on facilitation, integration, and streamlining communication, describing themselves as "shepherd", "go-between", in a role of "handholding", having a job where the editor "liaises" with authors, try to make "pieces fit together" or one in which they "mediate" the needs of different stakeholders. Other editors drew on navigation or directional

language to describe the data review process, one calling it a “traffic light process” or in other case using “green light”, “flagging”, or “raising a red flag” to communicate feedback to other stakeholders. For one editor, data curation is a “test drive” of the data. Data access is a key step for one editor as it provides “a path for reviewers to get to the data” and another editor tries to avoid sending reviewers to a “dead end.”

The challenging nature of their work comes through in sporting analogies like using the language of data policies to put different stakeholders on the “same playing field”, describing the criteria of data sharing through publication as “a moving target”, and setting up barriers to entry including different stages to review as “hurdles” that authors ought to navigate. One editor called communication amongst reviewers and authors “ping pong” where another just called this “a lot of back and forth in communication.” The complexity or perhaps stressful nature of their work also came through in medical and military analogies. One calling data publication and policy compliance “an uphill battle”, another calling their journals publication practices “revolutionary” where another called them something to “fight for.” In another instance, describing a step of review in which they rely on publisher staff to “triage” submissions.

Advocacy

Because editors are in a leadership role and often in contact with each type of stakeholder, they can advocate on behalf of the interests of others. Different types of labor and stakeholders are valued differently and may or may not have an opportunity to shape the workflow. The editor must think about the journal in terms of different

subgroups and, in so doing, “look beyond his or her personal perspective and incorporate the broader picture of the journal and its readership.” (McGinty, 1999, pg. 129).

Two editors took somewhat protective stances with respect to peer-reviewers. One stated that reviewers do not like to be treated like data curators, going so far as to recommend against other journals’ requiring their peer-reviewers to review data. He has either personally altered or overseen a tweak to the review workflow instructions so that peer-reviewers would not be required to review the minutia of each dataset: “If they want to look at the raw data, they can” but added later that “So we have occasionally tweaked language to say, look, you have the right to go through the data files and to the degree that you've checked them, what do you think about them? And we did a little survey of, I think, about a third of our peer reviewers. You can tell that they were going through the data files from their reports without asking them anything about it. (P-11, Pos. 66). Another editor framed the implementation of their journal’s pre-peer-review step of review by a topical editor to protect peer-reviewers, “What we’re trying to do there is, because reviewers are our most difficult, fragile, valuable asset, we're trying to keep from sending garbage to reviewers, okay?” (P-18, Pos. 44).

Four editors indicated their sympathy towards researchers through their advocacy for authors’ interests in different ways. One editor went to a workshop for his publishing house to voice his concern about the complexity of their data submission system (P-20). Another emphasized that authors have “done a lot of work” and “gone to the trouble to assemble a dataset” so it is the journal’s responsibility to help them get it ready for publication (P-18). This editor also displayed a sense of strong responsibility towards

datasets themselves, especially making data available that is out of scope for other journals and that might otherwise be lost or those that have not or perhaps would not necessarily be in the scope of publication for other journals: “It would be a huge, how you would say, positive accomplishment or would be huge credit to that journal that those datasets are starting to appear.” (P-18, Pos. 58). (P-18, Pos. 59).

Four editors took time to consider the accessibility and disciplinary interests of people at an indeterminate point in the future who might benefit from affordances that the editorial team and other stakeholders can implement in the present including mediating access for large file sizes, trying to anticipate what formats and descriptors will make a particular dataset usable.

One editor mentioned that he had to complain to the publisher that his staff who are essentially in-house data curators, were being overworked. He is trying to implement technical changes to the workflow that might make editors’, publication staff’s, peer-reviewers’, and authors’ roles easier and attends conferences and workshops where he surfaces these issues to, presumably, his superiors but his main takeaway was that: “It just takes a lot to justify your actions sometimes.” (P-20, Pos. 101).

Marketing

One of the editors’ primary roles promoting and maintaining the image their journal as a product or brand (also see section: Data Policies as Branding). Six of the eight respondents at one point or another in their interview used product-centered language or other turns of phrase that make the connection between their scholarly journal’s practices and commerce. One editor cited his publisher’s consistent branding on data availability requirements and at another point highlighted that his data journal works

with many data repositories to encourage fluidity and sharing, he used his publisher's data repository product as an example in the interview, "just because it's… proprietary." Product-centered language primarily took the form of using "product" as the unit of reviewable material that journals' work with as opposed to manuscript, submission, paper, or dataset. One editor described his workplace as, "We're a journal and we are in many ways, a very traditional product." (P-11, Pos. 46). One editor drew a direct connection between a unit of exchange in scholarly publishing—a citation—and money when referring to accidentally "shorting people on their citations" (P-18, Pos. 85).

Reputational Management, Trust, and Delegation of Authority

Another key theme that arose was the how much each of these editors is responsible for the reputational management of their journal. One editor explains:

> *"We believe that doing it this way is just absolutely essential to being a credible data journal. And, you know, that's not a criticism on anyone else, that's what we believe — this is what we need to do, the minimum that we need to do, in order to be credible." (P-11, Pos. 25)*

The two journals that engage in third-party data review emphasized the value of replicability as a safeguard against errors in publication. They describe this strict review standard as an assurance for the journal's reputation: "I know that I sleep easier at night with having third-party verifiers who are checking these things" (P-5, Pos. 29), the other adding they he it gives him, "peace of mind" (P-1, Pos. 9). Conversely, one editor expressed serious qualms about their journal's practice of dataset deposition in external repositories without prior editorial or peer-review vetting but concluded that ultimately, "…the proof is in the pudding. If the data had a problem, we're going to hear about it from other scientists later." (P-18, Pos. 49). This indicates an implicit trust in the scholarly communications ecosystem as a free marketplace of ideas, something that

another editor might have balked at as her description of the traditional publishing model was that it is "falling apart" (P-4, Pos. 58). This editor, however, relies on and trusts data centers to fulfill key parts of her data review workflow, "I'm more likely to trust the dataset if it comes from a data center that I know of in my field, (P-4, Pos. 46). Two other editors referred to repositories in terms of trust, one adding that the first step in their workflow is, "to get a manuscript into the system and to make sure the data are in a repository we trust."

This trust, or lack thereof, in other stakeholders is important because all the stakeholders in the relatively confined community of practice for both data review and data publication are interdependent. Some editors were more willing than others to trust other parties and delegate some of their authority due to the reputational, or other, risks. One extremely busy editor was happy to delegate the technical check to a new specialist team member. His teams are particularly large and distributed so, in general, he is willing to put higher levels of trust in other stakeholders. For example, he trusts that peer-review will have already been done peer-review for certain types at his journal due to their complex workflows (P-20). Another editor at a journal with a labor and time-intensive review process was pleased that his journal outsourced the verification functions to a third party as it directly alleviated time constraints on editorial offices and resulted in "less nightmares." (P-5). In another case, an editor lamented that peer-reviewers could not always be reliably trusted to follow advice provided for reviewing datasets and was adamant that academic editors should not be trusted to perform data curation functions. This editor, who works at a large journal, however seemed happy to outsource his formal data curator-editor relationship management duties to a new middle-managerial team

member (P-11). One journal was very pragmatic about recent changes to their review process in which the publisher does the first check looking for an active DOI: “We're trying to offload some of the editorial process to the publication staff” (P-18, Pos. 44).

Recommendations and Effectiveness

Editors’ attitudes varied significantly when speaking about the successes and challenges of implementing a data policy and review process. Some find their model to be highly effective. At the other end of the spectrum one editor went so far as to say, “I would recommend [laughs]—don't do it this way.” (P-20, Pos. 97). Both editors at journals with third-party verification polices were well-pleased, one stating, “I think the workflow is pretty efficient. It's definitely effective as is.” (P-1, Pos. 25). In fact, this editor added that, “I think it puts us at the forefront of political science journals in doing this.” (P-1, Pos. 57). Although, every editor identified areas for growth or development in either their policies, review standards, or workflow implementations, these two were both very satisfied with the effectiveness of their policy in practice.

Another satisfied editor, this one at an established data-centered, discipline-focused journal concluded that:

> *“I certainly think there's some good lessons from what we've done…which would be highly effectual to many other journals. I think they kind of challenge some of the precepts about having highly generic data journals because a lot of the stuff that we've been able to do is because we're quite thematically focused. I think working very closely with the academic community—who understand some of the more nuanced issues within a community—I think that is a key lesson.”*
> *(P-9, Pos. 87)*

This recommendation, to partner with and leverage the existing power and knowledge within a community of practice is like the recommendation of another editor, this one at a

mid-size, discipline-agnostic data journal in that they call for anyone trying to adopt or implement a data review policy to start by building relationships:

> *"If somebody high up endorses it then it will tend to percolate down through the system and people will actually think 'Okay, right. Well, we'd better do something about this,' and start thinking of ways to implement it."*
> *(P-4, Pos. 70)*

This advice highlights the role of a data journal editor in managing relationships as well as the importance of linking the ideals or goals of the journal with actual practices. Another journal editor, this one at an interdisciplinary mega-journal, hit on this issue in his response:

> *"Like any editor out there, I'm really tired of people promising stuff and they're not actually doing it. We all do it, right? … So, really think about what you want to achieve. And then, you know, don't make gigantic promises like saying you're going to review all data. It's hard. It's harder with a research manuscript than it is with a data paper."*
> *(P-11, Pos. 89)*

Here, the emphasis is on the role of an editor in helping to set realistic expectations and effectively communicating those to different stakeholders in a publishing environment with operational human constraints and technical constraints.

DISCUSSION

This study was motivated, in part, by our desire to gain new perspective on the practice of data sharing through data publication, focusing on editorial actions. These editors use a huge variety of terms to describe a relatively limited number of tasks in their workflows, indicating that there is a complex system of synonymy at work that has not been resolved by close communication or collaboration amongst different data journal's editors. All respondents shared an assumption that data sharing as facilitated through their journal's policies, standards, and practices will increase the possibility for eventual data reuse. Each respondent acknowledged that enacting their data sharing goals is complex in practice. This process is always mediated and constrained by both the social and technical nature of their work. Data journals sit at an intersection of traditional scholarly communication review practices and those used in data curation and data sharing institutions. This is academic publications work that is highly informed by developments in data repositories' technical capacities and data curatorial practices.

Data Journals

Data journals are here broadly defined as all those that engage in the processes and practices of reviewing datasets for quality. But how much data is enough to merit publication? What state should it be in? And what counts as data for this journal versus others? These questions do not have a unified or stable answer when compared with the range of responses of data journal editors.

Not all these journals publish "pure" data papers but each one has a process in place to manage datasets in their publication review process (Walters, 2020). Datasets, although heterogeneous within and across the purview of each of this study's different

journals, have categorically different characteristics than text-only documents, manuscripts, or datasets. This means that managing, working with, storing, processing, and sharing datasets for review poses distinctive challenges. In some ways, this aligns these journals more closely with the broader data sharing community than the traditional scholarly publications system.

In fact, the relationship of these editors to the broader academic publishing world is tricky and highly variable. Some editors appear to honestly believe in an open marketplace of academic ideas in which the wheat and the chaff will naturally be sorted by critical and mass consensus within the traditional scholarly communications ecosystem which indicates how the editor likely makes sense of his work—advancing knowledge, however incrementally (McGinty, 1999, p. 131). However, the depth of review and labor-intensive practices at some of these journals make a sharp contrast with traditional scholarly communications.

Journals are economic enterprises and editors have both a duty to and sometimes influence over the financial health of a journal within the publishing environment. The editors' usage of this product-centered, commercial, or marketing-style language may be a way that the editor can find meaning by tending the financial viability of the journal perhaps in service towards the scholarly community in general. As a gatekeeper for operations at the journal, the editor has a high burden of liability in decision making (McGinty, 1999). In general, this is directly proportional to their degree of purview at the journal. An editor's personal identity and professional reputation is coupled with that of their journal. The dual function of credibility for an editor's and a journal's reputation is a motivating factor for certain practices.

Although most of the editors acknowledged the impossibility of universal standards, they all indicated that their policies and practices have evolved, are changing, and will continue to grow and adapt over time. It is unclear whether data journal review policies will ever shape review policies at other, more traditional types of journals. However, it is possible that data journals' policies might become collectively more uniform through adoption and integration, perhaps with input from initiatives, scholarly communities, or societies, and governmental or funding mandates.

Data Policies and Review

Policies, standards, and practices are intertwined at data journals. Standards for availability are set by policies. They are realized through authorial compliance. Standards for replicability have been checked through an iterative process of review that engages peer-reviewers, editors, and a data curation team. The resultant granularity and documentation of the dataset may make it usable and useful to external end user. Iteration in review is often a feature, not a bug, when applying stringent standards for review and it is just part of the cost of doing business for these unique journals.

Accessibility of data is multifaceted at data journals. Two of the editors we talked to manage journals that do not have a universal Open Access policy meaning that the published papers will not necessarily be openly accessible to anyone. Two other journals take accessibility to include the language and jargon of the final product, prioritizing its readability for a broader audience.

The data review and availability models observed varied quite widely. All journals require something beyond "data not shown." There may be some technical minimum data characteristic or standard that is universal amongst these journals given

their uniting goal of predictive or potential data utility. Beyond that, many of the journal's qualities vary widely. Data gets reviewed in broadly one of three ways for a given journal: (1) One party conducts multiple types of review—for example, peer-review and a technical review, (2) Multiple parties perform multiple types of review—for example, two parties perform either a peer-review or a technical review, respectively, or (3) One party conducts a single type of review.

Stages, levels, or steps of review may be performed by different stakeholders and may be either in-house or external. This topology is imperfect. The key workflow and policy stakeholders that I identified were editorial staff, data repositories, reviewers or referees, data curators, scholarly communities, authors, publishers/publication staff and data end users. These are not intended to serve as exhaustive or mutually exclusive classes but ways to talk about and track the perspective, perceptions, and positionality of editors.

Availability of data for different stakeholders happen at a variety of points along the submission timeline. The rate of publication, which is dependent on multiple constraints like the volume of submissions, required depth of review, and labor and expertise constraints, ranges from quite slow compared to traditional publications to quite fast. The degree of establishment of the data policies ranges from unwritten to multi-layered and long-standing. They are, relatedly in different stages of evolution: some have not changed much if at all since their initial advent, some were actively changing at the time of the interview, and others were hoping to make more changes in the future regardless of their current status. We see, even in this small sample size, the full range of data review policies described by Hrynaszkiewicz et al.'s (2020) new RDA standard. This

range includes data availability, data formats and standards, embargoes, and peer-review of data. Similarly, the degree of these journals' differentiation from a traditional journal ranges from editor's feeling closely coupled to these models to feeling that they are distinctly and intentionally different.

Not all, but many, of these editors explored the tension between standards and reality, control, and flexibility. The utility of universally applicable standards and the limits of generalizability were common themes. Interpretive flexibility and interpretive diversity are natural in the human-implemented review practices that operationalize human-created review standards. As a compliment to these conceptual challenges, which came up in nearly every conversation, all the editors talked about their hands-on roles in managing edge cases, harnessing iterative or collaborative review, and weighing the potential of different stakeholders to meet the difficulties of data review.

None of the editors used the term interoperability at any point during their interviews. There are likely a few reasons for this. Firstly, we did not specifically ask about interoperability or the FAIR principles at any point in the interviews. Secondarily, many of these editors do not work with data format and file issues head-on—this work is often relegated to other stakeholders like data curatorial teams, both in-house and external, as well as publisher staff. The underlying data review objectives for these journals are not common amongst the sample. They are not all trying to enact types of compliance or set new transparency best practices. In fact, transparency only came up when editors tried to define what a dataset might be suitable for. No one mentions accessibility for accommodating disability specifically. There were only a few, fleeting mentions of information justice, equity, and security.

Data Work and Data Expertise

In their interviews, these editors expressed a high degree of comfort with both ambiguity and using their position of authority to make hard calls. For them, flexibility is as important as any standard. The flexibility they do and do not allow with respect to standards is acquired through repertoire and collaboration. While their attitudes towards their work ranged widely—from exacting to relaxed—these editors' work and experience touches every aspect of the research data lifecycle and can inform approaches to by many different stakeholders.

Each of these editors spoke about how instituting data review requires more labor which can translate to increasing time costs and personnel costs. So, increasing integration and facilitation are the two most common standards or goals that editors in this study wish to enact through their development and application of data policies and related tools. There is an underlying need for all of editors to justify the extra labor inherent in this work by mapping their journal's policies, standards, and practices to larger ideals or beliefs like the real or proposed benefit of data reusability and data reproducibility or replicability to a broad audience. To meet this need, every editor defined their journal's work with respect to other models including traditional publishing or disciplinary data repositories rather than data sharing. Some editors, in their effort to highlight their journal's model and its unique benefits pointed to hypothetical or real examples of data sharing in practice.

Money is perhaps the least surprising of the three primary constraints that emerged as running any organization requires funding: time and money are often inversely proportional in these settings and the availability of labor is dependent on both.

However, the unique challenges posed by the model of data review at these journals perhaps throw this into sharper relief and reveals that not only labor but workers with specific expertise is necessary to perform the types of analyses. Indeed, perhaps no journal trying to do this work will be constrained by money alone: data review experts themselves are at a premium. In the proliferation of publishable data and data publishing, to even attempt to dissemination, data journal editors themselves become specialized professionals (McGinty, 1999). We see this too in the case of data review experts and data curators in these editors' responses. As work by Thomer et al., (2022) indicates, data workers are often invisible in a workflow model. It is precisely this invisibility that makes classification of their role difficult. Bifurcated workflows may make these workers both indispensable and completely replaceable at the same time. It takes time, money, and skilled labor to operationalize data review that all the data journals are conducting, regardless of the depth of review. At one journal, the locus of valued expertise is placed solely on peer-review, subject-area experts who are supposed to check the scientific meritoriousness of the data—whether the right tests were performed, correct uncertainties etc.—as well as the technical dimensions of the data.

Data Quality and Standardization

Throughout these interviews quality is desirable but ambiguous. Data quality, per the interviews I analyzed, is an elusive but somehow omnipresent characteristic of data. Sometimes quality appears so context- or discipline-dependent that standardization of review criteria is functionally useless. While some editors appeared to rely on a vague category like common sense in peer review or an implied minimum standard of what is quality, it is important to recall that the scope of most of these journals is relatively

narrow in the landscape of broader publishing. Most of them are not particularly mature journals with long-established practices.

However, there is an almost uncanny reiteration of a call for *fitness for use* as the uniting thread of quality in these interviews. Perhaps a sound data journal is one that has certain practices in place as opposed to certain types of research or data in certain formats. These editors' references to quality being dependent, contextual, or continent on the reuse conditions are strikingly concordant given the diverse disciplinary contexts and publication models of the different journals in our sample. Editors of both a multi-disciplinary mega-journal—one with a large-volume of publications, that is relatively disciplinarily-agnostic—and discipline-specific journals highlighted the impossibility of determining how other people's data can, will, or should be used. Still, none of these editors has given up on the idea that there is some level of interoperability possible for aspirational broader or longer-term data use and reuse. The threshold for what degree of description, context, and annotation is sufficient for their version of interoperability is different for different data journals. The responsibility for describing work thoroughly and meeting these standards rests primarily with authors in these editors' descriptions. Different journals facilitate the data description process using domain-expert review and/or data-expert review along with socio-technical tools to make it easier for authors to meet research dataset description quality standards.

Some editors enumerated or described the criteria by which data is evaluated. Quality, in some of these descriptions, is distinguishable or separate from the methodological constraints of data's production. For these editors, it was something more intrinsically scientific or conceptual than mere comprehensiveness or completeness.

Quality must also, in that case, be distinct from dataset's formatting and its integrity. If the data is otherwise enriched and contextualized, data quality, may be a stand in for the reliability of the data, uniqueness of the data, or timeliness of the data.

It is worth focusing on whether there are commonalities in the technical aspects of working with data as an object of review precisely because these journals engage in different types of review of different types of data engaging different stakeholders who apply different standards. The distinction between the technical criteria of a dataset and its other scientific or conceptual qualities—which are described variously in these interviews—is blurred in journals data review goals and practices. As discussed in the previous section, the relative value or utility of a dataset may be broken down into a technical check and a scientific quality check. These are inherently related. If one cannot find, open, process, or store a dataset, its other utility and reusability dimensions are functionally moot.

LIMITATIONS

Interviews are extremely useful in quantitative research because of their capacity for in-depth data collection that affords exploration and flexibility. Interviews can be key to understanding how individuals understand and make meaning of issues (Flick, 2022) Like many qualitative studies, our purposive sample does not guarantee that our findings are generalizable or representative (Williams & Wager, 2013). We did not record self-identified demographic information of the study participants in terms of age, gender, career stage, years of experience, country of residence, area(s) of disciplinary specialization, and so on. Our sample does not cover all scientific domains. Christian et al (2020) intended to study journals across the biological, health, and social sciences, and as in that study, health sciences journals were vastly underrepresented in this study with zero health or biomedical journal editor participants, in contrast to the higher representation of earth and physical sciences. There is still not as much appetite and implementation of open data in the health sciences due to human subject privacy, confidentiality, and security concerns as well as the ethics of informed consent with unspecified future use of data (Tenopir et al, 2015).

There are many additional approaches that could augment or extend the proposed study but that I did not undertake due to the lack of time. In the original grant proposal, the primary investigators proposed complimentary interviews with submitting or published authors at data journals to help understand their perspectives. These, as well as interviews of peer-reviewers and data curators, could add another dimension to this study's findings.

My purpose was to characterize data journal editors' perspectives, expectations, understandings, and perceptions regarding their data review policies and practices and to synthesize their experiences. However, as with any study, I have more unanswered questions than answered questions. My research raised questions about the potential for automated tools to be proposed as technical solutions to the shortage of time, money, and expertise necessary to conduct detail-oriented review that I observed at these data journals. One of our study subjects proposed a more industrialized review model in 2018 but in in 2023, this seems like a direct appeal to artificial intelligence, machine learning, and algorithmic approaches to dataset and manuscript review optimization including systems like peer-review, which are already being proposed (Ghosal et al., 2022). Algorithmic approaches often propagate the existing subjective characteristics of human review processes, including bias, are often only made obdurate and invisible through automation, raising even more questions. In fact, calls by Gebru et al., (2021) for datasheets for datasets used in machine learning—emphasizing the power of context and documentation for ethical dataset use and reuse—echoes the calls for the transparent description of research datasets that data journals and their dataset review practices have tried to answer.

CONCLUSION

Interviews of editors at data journals demonstrate how data policy and practices are related and interdependent. Our detailed characterization of the verification work being developed and implemented at data journals can inform and improve data sharing practices and policies more broadly. All data journal editors in this study are carefully weighing the costs and benefits of reviewing the quality of datasets and are not only professionally but personally invested in the goals of data sharing. Editors' vantage as managers of other stakeholders, policies, and workflows makes them ideal candidates to report on the practices at these journals.

The technical aspects of working with datasets pose specific challenges in the data review workflows of these different journals. Editors are considering how to manage issues like proprietary formats, digital preservation of live datasets using static file formats, sensitive human datasets, and increasing file sizes that pose storage concerns. Most journals partner with one or more designated data repositories to make their workflows simpler and more efficient. This may or may not include integration or sharing of dataset descriptions across different platforms in an online manuscript submission program. Such software can allow different stakeholders to communicate about the application of review standards to data. Conventional tools like manuscript submission portals and editorial managerial software may or may not work for various aspects of dataset review depending on the format of the final data paper.

The idea of data quality is everywhere and nowhere in these interviews. Quality is directly tied to *fitness for use*. This is sometimes also tied to the potential for re-use and some editors view re-use as not only desirable but inevitable if certain quality standards

can be met. However, they acknowledge that the conditions of future use are unclear, so it is impossible to anticipate which criteria to apply for optimal interoperability.

Disciplinarity matters in the world of data publication and data review. Data journals that are more narrowly-focused within a discipline, scientific society, or community of practice seem to have an easier time generating community buy-in for their journal's unique, labor-intensive models. Indeed, many of these journals arose out of a particular disciplinary community's desire to see either certain types of datasets more widely shared or certain quality standards reinforced for the benefit of all. Because of the more narrowly defined subject matter and/or methodologies, these communities and journals can more easily incorporate existing community data description practices into their workflows than multi-disciplinary journals. Two of the multi-disciplinary journals in our study could be classed as mega-journals—those that deal with extreme volumes of submissions and may have more industrialized review processes in place —and face somewhat different operational constraints than the disciplinary journals in the natural and social sciences in our sample. For example, these journals from wealthier publishing houses were the only journals that used an in-house team to perform technical data quality checks, likely because such a model is not affordable at scale for smaller journals. The role of discipline-agnostic or mega-journals editors with respect to setting or enforcing quality standards for all data is ambiguous given that quality is primarily defined by fitness for use which has a disciplinary and community of practice dimension (Spezi et al., 2017; Thelwall, 2020; Wakeling et al., 2019).

Publishing and reviewing datasets is both difficult and time-consuming in these editors accounts. There are often tensions between the values that journal editors, publishers, peer-reviewers, and data reviewers are trying to enact at these journals which makes data journal operation perhaps even more complex than traditional scholarly communications. There is a categorical agreement amongst this sample cohort that trying to publish or review datasets is difficult, in part because the model is still relatively novel in the world of data sharing and scholarly communications. Many of these journals or their data-specific review and publication policies and practices are less than 10 years old as of 2023; as the field matures, new tools—like the custom review software that one of the publishers in this study has developed in-house to manage dataset submission and review—and standards will likely emerge. For some journals, their focus on fitness for use is immediate and is in the form of computation verificability. This practice of re-analyzing data as part of the journal's review protocol requires specialized knowledge and expertise of either the subject-matter or data curation. Data review expertise is a cross-cutting operational constraint; even journals with no shortage of time, money, and labor may struggle to find skilled data reviewers.

REFERENCES

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, *13*(4), 437–461. https://doi.org/10.1007/s11948-007-9042-5

Anger, M., Wendelborn, C., Winkler, E. C., & Schickhardt, C. (2022). Neither carrots nor sticks? Challenges surrounding data sharing from the perspective of research funding agencies—A qualitative expert interview study. *PLOS ONE*, *17*(9), e0273259. https://doi.org/10.1371/journal.pone.0273259

Atici, L., Kansa, S. W., Lev-Tov, J., & Kansa, E. C. (2013). Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. *Journal of Archaeological Method and Theory*, *20*(4), 663–681. https://doi.org/10.1007/s10816-012-9132-9

Barrowman, N. (2018). Why Data Is Never Raw. *New Atlantis: A Journal of Technology & Society*, *56*, 129–135.

Berberi, I., & Roche, D. G. (2022). No evidence that mandatory open data policies increase error correction. *Nature Ecology & Evolution*, 1–4. https://doi.org/10.1038/s41559-022-01879-9

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

Bornmann, L. (2008). Scientific peer review: An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture : Journal of the Sociology of Self - Knowledge*, *6*(2), 23–38. ProQuest Central; Sociological Abstracts.

Bowen, G. A. (2006). Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods*, *5*(3), 12–23. https://doi.org/10.1177/160940690600500304

Bowker, G. C. (2005). *Memory practices in the sciences*. Cambridge, Mass. : MIT Press, c2005. https://catalog.lib.unc.edu/catalog/UNCb4823178

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7*(1), Article 1. https://doi.org/10.2218/ijdc.v7i1.218

Callaghan, S., Tedds, J., Lawrence, R., Murphy, F., Roberts, T., & Wilcox, W. (2014). Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples. *International Journal of Digital Curation*, *9*(1), Article 1. https://doi.org/10.2218/ijdc.v9i1.310

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

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, *8*(1), 108. https://doi.org/10.1038/s41597-021-00892-0

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(15), S2. https://doi.org/10.1186/1471-2105-12-S15-S2

Cheifet, B. (2020). Open data in a deeply connected world. *Genome Biology*, *21*(1), 96. https://doi.org/10.1186/s13059-020-02010-6

Choudhury, S., Fishman, J. R., McGowan, M. L., & Juengst, E. T. (2014). Big data, open science and the brain: Lessons learned from genomics. *Frontiers in Human Neuroscience*, *8*, 239. https://doi.org/10.3389/fnhum.2014.00239

Christian, T.-M., Gooch, A., Vision, T., & Hull, E. (2020). Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLOS ONE*, *15*(3), e0230281. https://doi.org/10.1371/journal.pone.0230281

Cousijn, H., Habermann, T., Krznarich, E., & Meadows, A. (2022). Beyond data: Sharing related research outputs to make data reusable. *Learned Publishing*, *35*(1), 75–80. https://doi.org/10.1002/leap.1429

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in One Issue of Psychological Science. *Psychological Science*, *34*(4), 512–522. https://doi.org/10.1177/09567976221140828

Culina, A., van den Berg, I., Evans, S., & Sánchez-Tójar, A. (2020). Low availability of code in ecology: A call for urgent action. *PLOS Biology*, *18*(7), e3000763. https://doi.org/10.1371/journal.pbio.3000763

De Schutter, E. (2010). Data publishing and scientific journals: The future of the scientific paper in a world of shared data. *Neuroinformatics*, *8*(3), 151–153. https://doi.org/10.1007/s12021-010-9084-8

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. https://doi.org/10.1177/0306312711413314

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, *75*(6), 1274–1297. https://doi.org/10.1108/JD-08-2018-0133

Flick, U. (2022). *Designing qualitative research* (By pages 2-15). SAGE Publications, Ltd. https://doi.org/10.4135/9781849208826

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, *17*(1), e0259238. https://doi.org/10.1371/journal.pone.0259238

Gitelman, edited by L. (2013). *"Raw data" is an oxymoron*. Cambridge, Massachusetts ; London, England : The MIT Press, [2013]. https://catalog.lib.unc.edu/catalog/UNCb7884328

Glonti, K., Boutron, I., Moher, D., & Hren, D. (2019). Journal editors' perspectives on the roles and tasks of peer reviewers in biomedical journals: A qualitative study. *BMJ Open*, *9*(11), e033421. https://doi.org/10.1136/bmjopen-2019-033421

Goodey, G., Hahnel, M., Zhou, Y., Jiang, L., Chandramouliswaran, I., Hafez, A., Paine, T., Gregurick, S., Simango, S., Peña, J. M. P., Murray, H., Cannon, M., Grant, R., McKellar, K., & Day, L. (2022). *The State of Open Data 2022*. https://doi.org/10.6084/m9.figshare.21276984.v5

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough?: An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Hamilton, D. G., Page, M. J., Finch, S., Everitt, S., & Fidler, F. (2022). How often do cancer researchers make their data and code available and what factors are associated with sharing? *BMC Medicine*, *20*(1), 438. https://doi.org/10.1186/s12916-022-02644-2

Hanson, A. (2017). Negative Case Analysis. In *The International Encyclopedia of Communication Research Methods* (pp. 1–2). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118901731.iecrm0165

Health, N. I. of. (2020). Final NIH policy for data management and sharing. *NOT-OD-21-013. Vol NOT-OD-21-013. NIH Grants & Funding. Bethesda, MD: Office of The Director, National Institutes of Health*.

Heaton, J. (2022). *Reworking qualitative data*. https://doi.org/10.4135/9781849209878

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R., & Goudie, S. (2020). Developing a Research Data Policy Framework for All Journals and Publishers. *Data Science Journal*, *19*(1), Article 1. https://doi.org/10.5334/dsj-2020-005

Jiao, C., & Darch, P. T. (2020). The role of the data paper in scholarly communication. *Proceedings of the Association for Information Science and Technology*, *57*(1), e316. https://doi.org/10.1002/pra2.316

Karhulahti, V.-M., & Backe, H.-J. (2021). Transparency of peer review: A semi-structured interview study with chief editors from social sciences and humanities. *Research Integrity and Peer Review*, *6*(1), 13. https://doi.org/10.1186/s41073-021-00116-4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kim, J. (2020). An analysis of data paper templates and guidelines: Types of contextual information described by data journals. *Science Editing*, *7*(1), 16–23. https://doi.org/10.6087/kcse.185

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799. https://doi.org/10.1002/asi.23424

Kong, L., Xi, Y., Lang, Y., Wang, Y., & Zhang, Q. (2019). A data quality evaluation index for data journals. In J. Li, X. Meng, Y. Zhang, W. Cui, & Z. Du (Eds.), *Big Scientific Data Management* (pp. 291–300). Springer International Publishing.

Kozlov, M. (2022). NIH issues a seismic mandate: Share data publicly. *Nature*, *602*(7898), 558–559. https://doi.org/10.1038/d41586-022-00402-1

Kratz, J. E., & Strasser, C. (2015). Researcher Perspectives on Publication and Peer Review of Data. *PLOS ONE*, *10*(2), e0117619. https://doi.org/10.1371/journal.pone.0117619

Lee, J.-S. (2022). Setting up a checkpoint for research on the prevalence of journal data policies: A systematic review. *Information for a Better World: Shaping the Global Future: 17th International Conference, IConference 2022, Virtual Event, February 28 – March 4, 2022, Proceedings, Part I*, 100–121. https://doi.org/10.1007/978-3-030-96957-8_11

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. JSTOR. https://doi.org/10.1086/684083

Li, K., Greenberg, J., & Dunic, J. (2020). Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. *Journal of the Association for Information Science and Technology*, *71*(2), 172–182. https://doi.org/10.1002/asi.24226

London, B. (2021). Reviewing peer review. *Journal of the American Heart Association*, *10*(15), e021475. https://doi.org/10.1161/JAHA.121.021475

Maggin, D. M. (2022). Journal editor and associate editor perspectives on research reproducibility and open science. *Remedial and Special Education*, *43*(3), 135–146. https://doi.org/10.1177/07419325211017294

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, *26*(13), 1753–1760. https://doi.org/10.1177/1049732315617444

McGinty, S., 1958-. (1999). *Gatekeepers of knowledge: Journal editors in the sciences and the social sciences* (Davis Library). Westport, Conn. : Bergin & Garvey, 1999. https://catalog.lib.unc.edu/catalog/UNCb3272723

Musen, M. A., & Musen, M. A. (2022). Without appropriate metadata, data-sharing mandates are pointless. *Nature (London)*, *609*(7926), 222–222. https://doi.org/10.1038/d41586-022-02820-7

National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Board on Research Data and Information; Division on Engineering and Physical Sciences; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Analytics; Division on Earth and Life Studies; Nuclear and Radiation Studies Board; Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Board on Behavioral, Cognitive, and Sensory Sciences; Committee on Reproducibility and Replicability in Science. (2019). *Reproducibility and Replicability in Science*. National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK547537/

Nelson, A. (n.d.). *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. Memorandum for the Heads of Executive Departments and Agencies. Washington, DC: Executive Office of the President Office of Science and Technology Policy; 25 Aug 2022 [cited 12 Sep 2022]*.

Nelson, J. (2017). Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research. *Qualitative Research*, *17*(5), 554–570. https://doi.org/10.1177/1468794116679873

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691.

Nüst, D., & Eglen, S. J. (2021). CODECHECK: An Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. *F1000Research*, *10*, 253. https://doi.org/10.12688/f1000research.51738.2

O'Hara, K. (2014). Enhancing the quality of open data. In L. Floridi & P. Illari (Eds.), *The Philosophy of Information Quality* (pp. 201–215). Springer International Publishing. https://doi.org/10.1007/978-3-319-07121-3_11

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for*

Information Science and Technology, *48*(1), 1–10. https://doi.org/10.1002/meet.2011.14504801174

Pampel, H., & Dallmeier-Tiessen, S. (2014). Open research data: From vision to practice. In S. Bartling & S. Friesike (Eds.), *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing* (pp. 213–224). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_14

Parsons, M. & Fox, PA. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*, WDS32–WDS46. https://doi.org/10.2481/dsj.WDS-042

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Radha, S. K., Taylor, I., Nabrzyski, J., & Barclay, I. (2021). Verifiable Badging System for scientific data reproducibility. *Blockchain: Research and Applications*, *2*(2), 100015. https://doi.org/10.1016/j.bcra.2021.100015

Raittio, E., Sofi-Mahmudi, A., & Shamsoddin, E. (2022). The use of the phrase "data not shown" in dental research. *PLoS One*, *17*(8). Agriculture Science Database; Environmental Science Database; ProQuest Central. https://doi.org/10.1371/journal.pone.0272695

Saldaña, J. (2016). *The coding manual for qualitative researchers* (Third edition.). Los Angeles : SAGE, [2016]. https://catalog.lib.unc.edu/catalog/UNCb8697471

Schöpfel, J., Farace, D. J., Prost, H., & Zane, A. (2019). Data Papers as a New Form of Knowledge Organization in the Field of Research Data. *KNOWLEDGE ORGANIZATION*.

Seo, S., & Kim, J. (2020). Data journals: Types of peer review, review criteria, and editorial committee members' positions. *Science Editing*, *7*(2), 130–135. https://doi.org/10.6087/kcse.207

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917

Smith, A. M., Niemeyer, K. E., Katz, D. S., Barba, L. A., Githinji, G., Gymrek, M., Huff, K. D., Madan, C. R., Mayes, A. C., Moerman, K. M., Prins, P., Ram, K., Rokem, A., Teal, T. K., Guimera, R. V., & Vanderplas, J. T. (2018). Journal of Open Source Software (JOSS): Design and first-year review. *PeerJ Computer Science*, *4*, e147. https://doi.org/10.7717/peerj-cs.147

Spezi, V., Wakeling, S., Pinfield, S., Creaser, C., Fry, J., & Willett, P. (2017). Open-access mega-journals: The future of scholarly communication or academic dumping ground? A review. *Journal of Documentation*, *73*(2), 263–283. https://doi.org/10.1108/JD-06-2016-0082

Staunton, C., Barragán, C. A., Canali, S., Ho, C., Leonelli, S., Mayernik, M., Prainsack, B., & Wonkham, A. (2021). Open science, data sharing and solidarity: Who benefits? *History and Philosophy of the Life Sciences*, *43*(4), 115. https://doi.org/10.1007/s40656-021-00468-6

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

Thomer, A. K., Akmon, D., York, J. J., Tyler, A. R. B., Polasek, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The Craft and Coordination of Data Curation:

Complicating Workflow Views of Data Science. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 414:1-414:29. https://doi.org/10.1145/3555139

Verhulst, S., & Young, A. (2022). Identifying and addressing data asymmetries so as to enable (better) science. *Frontiers in Big Data*, *5*, 888384. https://doi.org/10.3389/fdata.2022.888384

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Wakeling, S., Creaser, C., Pinfield, S., Fry, J., Spezi, V., Willett, P., & Paramita, M. (2019). Motivations, understandings, and experiences of open-access mega-journal authors: Results of a large-scale survey. *Journal of the Association for Information Science and Technology*, *70*(7), 754–768. https://doi.org/10.1002/asi.24154

Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights the UKSG Journal*, *33*, 18. https://doi.org/10.1629/uksg.510

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLOS ONE*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Williams, P., & Wager, E. (2013). Exploring why and how journal editors retract articles: Findings from a qualitative study. *Science and Engineering Ethics*, *19*(1), 1–11. https://doi.org/10.1007/s11948-011-9292-0

Appendix A. List of Journals

Journal	Domain	ISSN	Publisher	Professional Society Affiliation	Access Type (as of 2023)	TOP Factor (as of 2023)	Website
American Journal of Political Science (AJPS)	Social sciences	Online ISSN: 1540-5907, Print ISSN: 0092-5853	Wiley	Midwest Political Science Association (MPSA)	Gold Open Access (Author Processing Charge)	11	https://ajps.org/
Biodiversity Data Journal (BDJ)	Natural sciences	Online ISSN 1314-2828, Print ISSN 1314-2836	Pensoft	NA	Gold Open Access (Author Processing Charge)	NA	https://bdj.pensoft.net/
Data Science Journal (DSJ)	Multidisciplinary	Online ISSN: 1683-1470	Ubiquity, De Gruyter	*Committee on Data of the International Science Council (CODATA)	Gold Open Access (Author Processing Charge)	NA	https://datascience.codata.org/
Data in Brief (DIB)	Multidisciplinary	Online ISSN: 2352-3409	Elsevier	NA	Gold Open Access (Author Processing Charge)	NA	https://www.sciencedirect.com/journal/data-in-brief
Earth System Science Data (ESSD)	Natural sciences	Online ISSN: 1866-3516	Copernicus	NA	Gold Open Access (Author Processing Charge)	2	https://www.earth-system-science-data.net/
Ecological Archives (ESA)	Natural sciences	Online ISSN: 1939-9170	Ecological Society of America	Ecological Society of America (ESA)	Hybrid (varying models across ESA journals)	0	https://esapubs.org/archive/default.htm
Scientific Data (SD)	Multidisciplinary	Online ISSN: 2052-4463	Nature	NA	Gold Open Access (Author Processing Charge)	8	https://www.nature.com/sdata/
State Politics and Policy Quarterly (SPPQ)	Social sciences	ISSN: 1532-4400	Cambridge University Press	American Political Science Association (APSA)	Hybrid model (partly Gold OA, author chooses)	NA	https://www.cambridge.org/core/journals/state-politics-and-policy-quarterly

Appendix B. Journals Policy and Guidance

Journal	Data Publication Policies	Guidance for Authors	Guidance for Reviewers
American Journal of Political Science (AJPS)	https://ajps.org/ajps-verification-policy/	https://ajps.org/wp-content/uploads/2018/05/ajps_replication-guidelines-2-1.pdf	https://ajps.org/ajps-reviewer-instructions/
Biodiversity Data Journal (BDJ)	https://bdj.pensoft.net/about#Datapublication	https://bdj.pensoft.net/about#Forauthors	https://bdj.pensoft.net/about#GuidelinesforReviewers
Data Science Journal (DSJ)	https://datascience.codata.org/about/editorialpolicies/	https://datascience.codata.org/about/submissions/	https://datascience.codata.org/about/editorialpolicies/
Data in Brief (DIB)	https://www.sciencedirect.com/journal/data-in-brief/about/policies-and-guidelines	https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors	https://www.elsevier.com/reviewers/how-to-review
Earth System Science Data (ESSD)	https://www.earth-system-science-data.net/peer_review/review_criteria.html	https://www.earth-system-science-data.net/submission.html	https://www.earth-system-science-data.net/peer_review/review_criteria.html
Ecological Archives (ESA)	https://www.esa.org/publications/data-policy/	https://www.esa.org/wp-content/uploads/2021/03/Instructions_For_Reviewers_ECY_ECM_EAP_Mar2021.pdf	https://www.esa.org/wp-content/uploads/2021/03/Instructions_For_Reviewers_ECY_ECM_EAP_Mar2021.pdf
Scientific Data (SD)	https://www.nature.com/sdata/policies/data-policies	https://www.nature.com/sdata/author-instructions	https://www.nature.com/sdata/policies/for-referees
State Politics and Policy Quarterly (SPPQ)	https://www.cambridge.org/core/journals/state-politics-and-policy-quarterly/information/author-instructions/preparing-your-materials#dataverse	https://www.cambridge.org/core/journals/state-politics-and-policy-quarterly/information/author-instructions	https://www.cambridge.org/core/journals/state-politics-and-policy-quarterly/information/peer-review-information/instructions-for-peer-reviewers

Appendix C. Sample Overlap with Previous Studies

Journal	Christian et al., 2020	Walters, 2020	Seo & Kim, 2020	J. Kim, 2020	Walters, 2020- Classification	Walters, 2020- Notes on Review Processes	Thelwall, 2020
American Journal of Political Science (AJPS)	Y	N	N	N	NA	NA	N
Biodiversity Data Journal (BDJ)	N	Y	Y	Y	Group 1: "Pure" data journals, for which data reports comprise at least half the papers in the journal	"After initial editorial review, each paper is sent to two or three nominated reviewers, who are expected to submit their comments within ten days; and to several panel reviewers, who may choose whether to comment. Authors' revisions are expected within one week, although extensions may be granted. Most revised papers are re-evaluated by the editors, although some are sent for another round of review."	Y
Data Science Journal (DSJ)	N	N	N	N	Group 2 Journals that publish data reports but are devoted mainly to other types of contributions	NA	N
Data in Brief (DIB)	N	Y	Y	Y	Group 1: "Pure" data journals, for which data reports comprise at least half the papers in the journal	"Six criteria: Is the data format in alignment with existing standards? Are the protocol/references for generating data sufficiently explained? Is the data description complete and is data well-documented? Do the authors adequately explain the data's utility? Are the data potentially reusable? Does the article adhere to the template?"	Y
Earth System Science Data (ESSD)	N	Y	Y	Y	Group 1: "Pure" data journals, for which data reports comprise at least half the papers in the journal	"Papers that meet the standards of an initial rapid review are posted to the journal's website. Readers are invited to submit reviews or comments, and the editors' decision accounts for both the solicited reviews and any additional remarks. If the paper is accepted, it is published with the referees' comments (anonymous or attributed), the readers' comments (attributed), and the authors' replies."	Y

Ecological Archives (ESA)	N	N	N	Y (Ecology and Ecological Research)	Group 3: Journals that do not actually include data reports a as a publications type; Ecological Applications requires public dissemination of the date used in empirical papers.	NA	N
Scientific Data (SD)	N	Y	Y	Y	Group 1: "Pure" data journals, for which data reports comprise at least half the papers in the journal	"Each paper is reviewed by one data standards expert and at least one subject expert based on 'the technical quality of the procedures used to generate the data, the reuse value of the resulting datasets and their alignment with existing community standards, and the completeness of the data description. [Acceptance] is not based on the perceived impact or novelty of the findings."	Y
State Politics and Policy Quarterly (SPPQ)	N	N	N	N	NA	NA	N

Appendix D. Policy Terms

Examples of terms used in policies to refer to categories of transparency requirements (Christian et al, 2020)

Data	Analytic Methods	Research Materials
data	protocols	materials
dataset	programs	other details
microarray data	computer code	description
sequence data	computer programs	readme file
genetically modified organisms and mutants	scripts	additional information
electron microscopy data	methods	related materials
genotype data	program code	metadata
nucleotide sequences	software	other useful materials
proteins sequence data	algorithms	other artifacts
raw data	models	explanatory file
certain types of data	statistical tools	codebook
stimulus norms	analytic methods	relevant description
supporting data	laboratory protocols	
	commands	

Appendix E. Invitation to Participate

Increasing the Value of Open Access through Open Data Publication Policies

INVITATION TO PARTICIPATE

Editors

Dear [NAME]:

The Odum Institute for Research in Social Science at the University of North Carolina, in partnership with the Dryad Digital Repository, is conducting interviews with editors of data journals that publish peer-reviewed data papers describing research datasets and mechanisms for accessing these datasets. These informal interviews are part of a larger research project funded by the Robert Wood Johnson Foundation (#OAR 74419) that aims to develop an evidence-based model for data policy implementation that yields the greatest degree of access to quality research data.

Because data journals have processes in place to review the quality of data described in published data papers, your insight is valuable to us and we hope that you will participate. The information you provide will help us to identify the most effective and efficient methods for implementing robust data policies that include data quality review.

The interview will be conducted via web conferencing software and will take approximately 30 minutes to complete. Your participation is voluntary; you may end the interview at any time. You will NOT be individually identified in any reports that are produced from the interview. While measures have been put in place to prevent confidentiality breaches, there is a chance that your information may be accidentally disclosed.

If you have any questions or concerns about the study, please contact the principal investigators at odumarchive@unc.edu or 919-962-6293. If you have questions about your rights as a research participant, you may contact the University of North Carolina Institutional Review Board at IRB_Subjects@unc.edu and mention study 18-0295.

If you are willing to participate, please suggest a date and time that suits you. I will do my best to accommodate your schedule. If you have any other questions, please do not hesitate to ask.

Thank you,

Thu-Mai Lewis Christian

Assistant Director for Archives

H. W. Odum Institute for Research in Social Science

Appendix F. Letter of Consent

University of North Carolina at Chapel Hill

Consent to Participate in a Research Study

Adult Participants

Consent Form Version Date: 08/07/2018

IRB Study #: 18-1711

Title of Study: Increasing the Value of Open Access Through Open Data Publication Policies

Principal Investigator: Thu-Mai Christian

Principal Investigator Department: Odum Institute for Research in Social Science

Principal Investigator Phone Number: (919) 962-6293

Principal Investigator Email Address: tlchristian@unc.edu

What are some general things you should know about research studies?

You are being asked to take part in a research study. Participation in this study is voluntary; you may refuse to participate or withdraw your consent to participate in the study for any reason without penalty.

Research studies are designed to obtain new knowledge. It is possible that this new knowledge will help people in the future. You may not receive any direct benefit from participating in research studies. There also may be risks associated with participating in research studies.

Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about participating in this research study.

You will be given a copy of this consent form. You should ask the researchers above, or staff members who may assist them, any questions you have about this study at any time.

What is the purpose of this study?

The purpose of this research study is to understand the manuscript and data review process of data journals from the perspectives of editors, peer-reviewers, and authors in order to develop and evidence-based model for data policy implementation that yields the greatest access to quality research data.

Who is sponsoring this study?
This research is funded by the Robert Wood Johnson Foundation (RWJF)) (the Sponsor). This means that the research team is being paid by the sponsor for doing the study. In addition, Todd Vision, a co-investigator on this study, participates in unpaid activities which are not part of this study for Dryad, an entity involved with this study. These activities may include consulting, service on committees or boards, giving speeches, or writing reports.

If you would like more information, please ask the researchers listed in the first page of this form.

Are there reasons you should not be participate in this study?

You are being asked to participate in this study because you have been identified as an individual who serves as an editor or peer-reviewer of a data journal and/or an author who has had an article published in a data journal. You should not participate in this study if you have never served as an editor or peer-reviewer of a data journal, or have not had an article published in a data journal.

How many people will participate in this study?

If you participate in this study, you will be one of approximately 15 people taking part in this research study.

What will be the duration of your participation in this study?

Your active participation in this study take place over an approximately thirty-minute period, or the length of time required to complete an interview. will also be contacted via email to verify your interview responses to give you an opportunity to clarify or correct any misinterpretations. Your total time commitment to this study will be approximately one hour.

What will happen if you take part in the study?

- You will participate in an interview via web conferencing software that will take approximately 60 minutes.
- At any point during the interview(s) you may refuse to answer any question or end your participation in the study at any time.
- With your permission, interviews will be recorded for transcription purposes only. Once transcription is complete, audio recordings will be deleted.
- A list of primary findings from the interview(s) will be sent to you via email for verification. If any of the findings do not accurately reflect your statements, you will be contacted via telephone to correct inaccuracies. This telephone correspondence will take approximately 30 minutes.

- Your name and/or any personal identifiers will not appear in any reports or papers released; pseudonyms will be used to reference individual study participants.
- De-identified interview transcripts and any other related study data may be used in secondary analyses as part of future studies.

What are the possible benefits from participating in this study?

Research is designed to benefit society by gaining new knowledge. You will not benefit personally from being in this research study.

What are the possible risks or discomforts that may result from participating in this study?

There may be uncommon or previously unknown risks. You should inform the researcher of any problems that arise.

How will information about you be protected?

- The principal investigator, Thu-Mai Christian, and designated project team members will be the only people with access to files containing personally identifiable information. Any information you provide will be kept strictly confidential; you will not be individually identified in any reports or publications produced from this study.
- Your contact information will be kept separate from the data, and only for purposes of follow-up contact and delivery of primary findings. All study files will be password-protected and stored on a secure centralized server hosted by University of North Carolina Information Technology Services.
- Direct quotations from interview transcripts may be extracted and used in reports or papers; however, pseudonyms or generalized references such as "a researcher" will replace proper nouns in all study documents.

- Interviews will be recorded using the investigator's workstation computer, which is password protected. Audio recordings of interviews will be deleted immediately upon the completion of transcription, which will take place as soon as possible after the actual interview has concluded. At any point, you may request that the audio recorder be turned off.
- De-identified data from this study may be shared with the research community at large to advance science. We will remove or code any personal information that could identify you before files are shared with other researchers to ensure that, by current scientific standards and known methods, no one will be able to identify you from the information we share.

Although every effort will be made to keep research records private, there may be times when federal or state law requires the disclosure of such records, including personal information. This is very unlikely, but if disclosure is ever required, the University of North Carolina at Chapel Hill will take steps allowable by law to protect the privacy of personal information. In some cases, your information in this research study could be reviewed by representatives of the University, research sponsors, or government agencies for purposes such as quality control or safety.

What if you want to stop before your part in the study is complete?

You can withdraw from the study at any time, without penalty. The investigators also have the right to end your participation in the study at any time. This could be because you have had an unexpected reaction or have failed to follow instructions, or because the entire study has been stopped.

Will you receive anything for being in the study?

You will not receive anything for being in this study.

Will it cost you anything to be in this study?

Aside from your time, there is no cost to being in this study.

What if you are a UNC employee?

Participation in this research is not part of your duties at the University, and your refusal to participate will not affect your employment. You will not be offered or receive special considerations related to your job for participating in this study.

What if you have questions about this study?

You have the right to ask, and have answered, any questions you may have about this research. If you have questions about the study, complaints, concerns, or if a research-related injury occurs, you should contact the researchers listed on the first page of this form.

What if you have questions about your rights as a research participant?
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject, or if you would like to obtain information or offer input, you may contact the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.

Participant's Agreement

I have read the information provided above. I have asked all the questions I have at this time. I voluntarily agree to participate in this research study.

Signature of Research Participant	Date
Printed Name of Research Participant	

Appendix G. Interview Guide Instrument

Increasing the Value of Open Access through Open Data Publication Policies

INTERVIEW GUIDE: EDITORS

INTRODUCTION

Thank you for taking the time to speak with me about your experience as editor of [JOURNAL]. We are particularly interested in journals that have a data review mechanism in place. Because [JOURNAL] publishes peer-reviewed data papers, we hope that you can offer some insight into the data review process and the criteria with which data are evaluated.

POLICY CLARIFICATION

1) **Please tell me about your journal's data review policy and how it differs from that of journals that publish traditional articles that describe research findings.**
   a) What are your thoughts about the data policy?
2) **How has the review policy and workflow evolved over time?**
   a) How was the content and language of the current policy determined?

POLICY IMPLEMENTATION WORKFLOW

3) **I would like to get a picture of what the data review workflow is like for your journal. Could you describe it for me?**
4) **What have you found to be the most useful in implementing the data review workflow?**
   a) In what ways has this contributed to the success of the policy?
5) **What, if anything, would you like to have changed about the current manuscript and data review workflow?**
   a) [IF MENTIONED] What do you think about those challenges?
   b) [IF MENTIONED] What would have to be done to overcome or eliminate these challenges?

DATA REVIEW STANDARDS

6) **What are the standards by which data are evaluated during peer-review?**
   a) How have these standards evolved over time?
   b) How might they evolve in the future?
7) **What are common issues that prompt a revise and resubmit? rejection?**
   a) How often are submissions rejected due to these data issues?

COMMENTS

8) **Based on your experience as an editor of a data journal with a data policy, what else do you think we should consider as we develop a model for data policy implementation for other journals?**

Appendix H. Preliminary Codes and Sensitizing Concepts

Initial Codes	**Sensitizing Concepts**
policy	incentives
workflow	timeframes/timelines
success	tradeoffs
challenge	waste
standards/best practices	data security
data issue	risk
repository	sensitive data/privacy
revise resubmit	disagreement/conflict
rejection	mandatory
tools/service/provision	large datasets/file size/storage capacity
disagreement/conflict	legal restrictions
replication	digital persistence/fragility
quality	guides/examples
compliance	checklists
embargoes	instructions
recommendation	monetary cost
	labor cost
Initial Process Codes	communication breakdowns/correspondence
decision-making	ease/convenience
formatting	embargo/hold
explaining	replicability/reproducibility
supporting/helping	professional judgment

Appendix I. Code System

Code System	Definition	Frequency
Code System		
Stakeholders		
Stakeholders > editorial staff	Editor-in-chief, editorial staff, and subject editors of data journals.	63
Stakeholders > repository	External organizations that host datasets. This includes general purpose and disciplinary repositories and data journals' relationships to these organizations and their workflows.	36
Stakeholders > reviewers	Volunteer reviewers or referees of submitted manuscripts usually with relevant discipline-specific expertise.	29
Stakeholders > data curators	Data workers managing the technical dimensions of datasets or internal consistency of results. These may either be in-house or through a third-party.	28
Stakeholders > scientific community	The broader scientific academic community.	22
Stakeholders > author	The person(s) submitting a manuscript to the journal.	96
Stakeholders > publisher staff	The publisher's editorial or administrative team involved in operating or overseeing data journals.	13
Stakeholders > data end users	Specific or hypothetical users of published data sets.	15
Data publication		
Data publication > benefits of sharing data	Balancing the costs of reviewing the quality of datasets with possible benefits	21
Data publication > novelty/difficulty of data publication	Reflections on data publication as a new model of data sharing and new format of scholarly publication.	25
Data publication > scholarly incentives	Motivators or rewards for all stakeholders to share data for publication and review: accountability, feeling good or successful etc.	20
Data publication > comparison to others	Editors making observations about the differences between their journal and other journals, as well as justifications, and arguments for the perceived advantages of their journal's	26

	approach and how they distinguish themselves. Contrasts are made with respect to other scholarly journals broadly as well as others that engage in some element of data review.	
Data publication > recommendations/ effectiveness	Editors recommendations about implementing data review policies and practices and their perception of the high-level effectiveness of their journal's approach at the time of the interview.	26
RQ 1: Data Journal Policy		
RQ 1: Policy > policy development	Who developed the policy, how long ago, what format it takes, whether it is a single policy or a set of policies?	23
RQ 1: Policy > policy development > external/community standards	Which community standards do journals leverage to review data? This may be embedded within or adapted into the language of polices, guidelines, templates, and other documentation for a specific journal or publisher. These include taxonomies, guidelines, and principles that are usually domain specific or from the academic publishing community.	23
RQ 1: Policy > clarity/communicating expectations	How are the requirements of the data policy or policies communicated to stakeholders?	30
RQ 1: Policy > data availability	At which point during the paper submission and review workflow the dataset is made available for editing and review, if at all. For example, should the dataset must be deposited and citable before submission, or is the dataset is required after the paper is accepted?	29
RQ 1: Policy > data availability > rights and licenses	Data rights and licenses including embargoes, copyright, regulatory restrictions, proprietary data	17
RQ 1: Policy > data availability > sensitive data	Sensitive human data including ethical concerns	5
RQ 2: Review Standards		
RQ 2: Review Standards > scope of data publication	How editor's describe what a data paper is, sometimes in contrast to other more traditional scholarly output formats. This includes the minimum criteria to constitute a data set worthy of review for a given journal.	33

	RQ 2: Review Standards > scope of data publication > flexibility/strictness/depth	Flexibility and the level of depth in implementing standard rules or practices.	26
RQ 2: Review Standards > quality		Data quality: any mentions of the datasets that data journals review with respect to their perceived scientific and analytic value.	12
RQ 2: Review Standards > usefulness/usability		Occurrences of terms like reusability, utility, and usefulness. This also includes references to reuse as well as editors discussing whether data may be potentially useful in the future for further analysis.	19
	RQ 2: Review Standards > usefulness/usability > documentation/interpretability	Is the dataset comprehensive and complete? This includes references to data context, annotations, details, metadata, fields, titles, units, interpretability, understandability.	31
	RQ 2: Review Standards > usefulness/usability > accessible language	How accessible should the dataset be?	4
RQ 2: Review Standards > formatting/interoperability		Any mentions of dataset formatting (database format, file format, file size) including specific mentions of formats, languages, and technical standards (e.g., Excel, NetCDF, CSV, PDF, GIS, SQL). When data is made available for review to a journal, do they have a standard in place for the resolution, rawness, or degree of processing of the data?	31
RQ 2: Review Standards > findability/identifiers		References to availability and findability, links, persistent identifiers, bidirectional linking, DOI, dead links, citation, supporting information.	16
RQ 2: Review Standards > veracity/soundness/rigor		Is the data true or accurate? Not noise? Scientific? Sound? Reasonable? Plausible? Direct occurrences of terms including rigor, meaningfulness, understandability. Do the contents of a dataset match analytic output? Is the dataset measuring something real? Signal not noise.	18

RQ 2: Review Standards > novelty/originality	How shiny, novel, original is the data? Does it have a high impact or newness factor?	8
RQ 2: Review Standards > verificability/replicability/reproducibility	Mentions of computation replicability or verificability. This constitutes a methodological approach to detail-oriented data review in which datasets and related materials are cross-checked by a second party to see that reported results or values are consistent with output of codes, calculations, analyses etc.	26
RQ 3: Workflow	Review workflow: how data journal editors talk about the process of editing, processing, and managing dataset publication. This includes editorial review, data curatorial review, and peer-review.	
RQ 3: Workflow > time, money, expertise constraints	Constraints in the workflow including rate-limiting steps/pain points/bottlenecks	40
RQ 3: Workflow > managing file formats	File sizes and formats that prove difficult to accommodate or plan for.	12
RQ 3: Workflow > workflow management software	Editorial and manuscript management software or other software used to process, transfer, annotate, and share datasets	28
RQ 3: Workflow > editorial role		
RQ 3: Workflow > editorial role > facilitation/workflow metaphors	Process-oriented language that editors use to describe the process of managing data review from submission to publication.	75
RQ 3: Workflow > editorial role > advocacy	Editor acting as a representative for another stakeholder's interest in the review workflow. Considering end-users in process.	19
RQ 3: Workflow > editorial role > marketing (product-	Product-centered, commercial, advertising language	19

	centered language)		
	RQ 3: Workflow > editorial role > managing edge cases/exceptions	Exceptions to the standard review protocol or processes and due to some characteristics of the data or their circumstances. How editors manage them.	15
	RQ 3: Workflow > editorial role > reputational management	Credibility in the work of editors including considerations of how their journal appears within the landscape of data sharing, scholarly communications, and disciplinary communities.	19
	RQ 3: Workflow > editorial role > trust/delegation of authority	Editors relying on the skills, time, and expertise of others to accomplish their goals.	25