


## Challenges in Ensuring Data Quality in Research Data Repositories – A Theoretical Examination


Besiki Stvilia<sup>1</sup>, Dong Joon Lee<sup>2</sup>, Yuanying Pang<sup>1</sup>, Fatih Gunaydin<sup>1</sup>


<sup>1</sup>School of Information, Florida State University


<sup>2</sup>Mays Business School, Texas A&M University

### Author Note

Besiki Stvilia, School of Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306-2100. E-mail: [bstvilia@fsu.edu](mailto:bstvilia@fsu.edu). ORCID iD  <https://orcid.org/0000-0002-2428-6627>.

Dong Joon Lee, Mays Business School, Texas A&M University, 4217 TAMU, College Station, TX 77843-4217. E-mail: [djlee@tamu.edu](mailto:djlee@tamu.edu). ORCID iD  <https://orcid.org/0000-0001-8994-163X>.

Yuanying Pang, School of Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306-2100. E-mail: [yp22c@fsu.edu](mailto:yp22c@fsu.edu). ORCID iD  <https://orcid.org/0009-0008-4262-1186>.

Fatih Gunaydin, School of Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306-2100. E-mail: [fg19d@fsu.edu](mailto:fg19d@fsu.edu). ORCID iD  <https://orcid.org/0000-0002-1956-7109>.

### Abstract

This qualitative study explores contradictions in data quality assurance (DQA) practices within research data repositories (RDRs), interpreted through the lens of activity theory and data quality literature. It analyzes data from 32 interviews with curators and repository managers representing 32 repositories across 30 universities in the United States.

The findings highlight several challenges faced by RDR staff, including differences in understanding DQA among stakeholders, the need for adequate resources, domain-specific knowledge, research expertise, and standardized data quality metrics. The study also identifies contradictions arising from misalignments in motivation between RDR staff and depositors, and friction between the DQA practices of research labs and the standardized DQA models promoted by RDRs. Additionally, it reveals contradictions between the RDRs' infrastructures and curation models and the evolving needs of stakeholders. The paper proposes resolution strategies for each identified contradiction.

## 1. Introduction

Data quality is an ethical issue. Data and information quality affects the quality of our decisions and activity outcomes and, ultimately, affects our lives and dignity [1]. Hence, data quality assurance (DQA) is critical to any data management workflow. DQA activities encompass a broad spectrum, including quality assessment and enhancement tasks carried out by data providers and repository personnel, data cleansing students undertake as part of coursework or DQA hackathons, and research reproducibility challenges<sup>1</sup>. The activities also comprise the assessment of dataset quality for the purposes of training AI models and/or informing decisions in policy and business domains [2,3,4].

There have been models, standards, laws, and best practices for guiding the design, implementation, and evaluation of DQA activities and workflows. The industry has employed several general quality assurance standards and methodologies (e.g., ISO 2500, ISO 8000, ISO 9000). Similarly, there exists a substantial amount of literature on and several models of data curation that encompass DQA activities (e.g., [5-8]). Data curation communities of practice are currently experiencing a revived focus on data quality and making datasets FAIR, which entails ensuring their findability, accessibility, interoperability, and reusability [9]. These communities actively create and distribute highly valuable procedures and scripts for data cleaning, normalization, linking, and disambiguation (e.g., [10]). However, efforts to implement DQA related components of FAIR framework operationalizations have been predominantly fragmented and situational, needing a solid foundation in the existing literature on

---

<sup>1</sup> <https://paperswithcode.com/rc2021>

metadata and data quality to improve their generalizability. Additionally, there is a need for more empirical studies that investigate and interpret DQA practices within research data repositories (RDRs) from the perspective of the data quality literature and how DQA literature can be used to guide and evaluate those practices.

DQA activities, like any other activity, are dynamic and are shaped by their evolving components and relationships, including the participant's needs, the challenges they face in the activity, and the solutions they seek. When tensions and misalignment with the activity's current structure and relationships affect the activity's outcome, those can be conceptualized as activity contradictions. In order to design novel and innovative forms of activities and services, it is crucial not only to identify the current problems and contradictions and how they have been resolved but also to interpret them through a theoretical lens such as activity theory. The latter can help better understand the ontological roots of those problems, design solutions for existing problems, and predict, prevent, and resolve future contradictions with similar theoretical structures [11]. There have been prior examinations of contradictions in research data curation and research information management (e.g., [11-13]). However, there is still a lack of *theory-guided empirical examination and interpretation of the challenges and problems of DQA practices in RDRs*.

## 2. Problem statement and research questions

Research data curation, including research DQA, is a complex sociotechnical process comprising multiple activities, actors and stakeholder groups, technologies, policies, standards, and research cultures. Identifying and understanding misalignments and challenges in the current DQA practices of RDRs and the strategies and solutions used to resolve those problems are critical for improving those practices.

Furthermore, by combining *theoretical reasoning* with *empirical data collection and analysis*, one can build a model that is both theoretically *informed and explained*, and *empirically validated*. This hybrid approach helps to create knowledge that is theoretically rigorous, practically applicable, and robust to new use cases and contexts [14]. Such theory-guided analysis of DQA problems and challenges can help design new, innovative forms of DQA activities that are better aligned with the concerns and emerging needs of the stakeholder groups. It

also helps develop best practices guides for and training research data curators in research DQA.

There is a need for a systematic, theory-based analysis, understanding, and interpretation of DQA challenges, issues, and solutions in RDRs. Such analysis can aid in a better understanding of the theoretical foundations of those often complex problems, designing solutions for existing issues as well as in anticipating, preventing, and resolving future DQA contradictions with similar theoretical structures.

Guided by activity theory, our study aims to address this need by examining the following research question:

What are the challenges and problems of DQA in RDRs, and what are some of the strategies for resolving those problems?

### 3. Related work

Quality is defined as "fitness for use" [15]. Various studies have explored the conceptualization of research data quality and researchers' perceptions and priorities regarding it (e.g., [16-20]). The understanding of what constitutes quality and useful data can differ within the same discipline, across different disciplines, and even within the same process [6]. A DQA process involves activities related to conceptualizing, measuring, and improving data quality [21]. Along with privacy and access, data quality is of significant ethical importance in data use and information system design. In the era of big data and data driven science, the saying "garbage in, garbage out" becomes even more relevant. Data quality has a direct impact on the quality of research outcomes, teaching, business decisions, and government policies, ultimately affecting human lives [1,22,23].

Universities are investing heavily in building reliable and secure infrastructures to manage digital research datasets created and used by their faculty and students. This investment is driven by the faculty's need to preserve and share their research data [24,25], mandates from state and federal funding agencies to openly share data for public benefit, research, and teaching purposes, as well as to enhance the reproducibility and replicability of research [26-28]. National and state laws also require ensuring data quality [29,30]. Additionally, some universities are interested in tracking and measuring the impact of these datasets, including for evaluating faculty for promotion and tenure [31]. However, a major obstacle to

data sharing and reuse is concerns about data quality. Data owners may worry about the quality and documentation of their data and its potential misuse or misinterpretation by others [19,32]. Conversely, users need data that is useful, valid, reliable, and accurately represents the phenomena they are studying or teaching, rather than just having access to large quantities of data [33,34]. Data creators often compile datasets for specific purposes, and without proper documentation, understanding these original purposes becomes challenging, hindering data reuse [20,35].

The study also draws on the digital data curation literature (e.g., [6,8]) to provide further context. While there are common infrastructure elements in digital data curation across various fields, the specific research tasks, data types, technologies, and approaches to managing, sharing, and assessing data and metadata may differ (e.g., [10,36,37]). Research data curation studies examined the contradictions of data curation activities in institutional repositories (IRs). These included but were not limited to a contradiction between dataset scale and existing IR software/storage capabilities, a lack of best practices for adopting tools, and a contradiction between available resources and activity objectives [7]. This literature also examined researchers' data management practices, including hurdles they encountered when working with RDR curators. For instance, earthquake science researchers found repositories' curation policies time-consuming and hindering their research activities [13]. While the prior studies provide valuable insights into the challenges and issues encountered by researchers and RDR staff during data curation activities in general, *there is a dearth of theoretical examination of contradictions in research DQA activities at RDRs.*

#### 4. Method

This paper presents findings from a part of a larger exploratory study. The study utilized datasets that included 122 approved applications for the CoreTrustSeal certification of trustworthy data repositories, interviews with 32 curators and repository managers, and 109 data curation-related documents from their repository websites. Data collection occurred from April 2022 to February 2023, encompassing a total of 146 unique RDRs.

The scope and focus of this particular paper are the barriers and challenges of DQA in RDRs and how these contradictions could be explained and resolved. It

reports on the analysis of 32 interview participants' answers to the related questions. The interviews were conducted between December 2022 and February 2023. The authors employed multiple methods to identify and recruit participants. The initial source and sampling frame for selecting interview participants was a list of 122 data repositories that were certified under the CoreTrustSeal Trustworthy Data Repositories Requirements as of March 2022. From this source, 30 U.S.-based repositories were identified. Additionally, a manual search was conducted across the web domains of 146 universities classified as R1 (doctoral universities with very high research activity) and 133 universities classified as R2 (doctoral universities with high research activity) by the Carnegie Classification of Institutions of Higher Education to identify additional RDRs. The following inclusion criteria were applied when selecting repositories for this study: repositories had to be U.S.-based, and the submission and/or curation of datasets had to be mediated by a curator or repository manager. Additionally, the repository's website needed to clearly specify who served as the manager and/or curator of the data collections. A total of 97 repositories and 138 associated managers or curators met these criteria. We contacted these 138 potential participants by email, and 32 agreed to participate in a Zoom interview.

Selecting interview participants from a single country (i.e., the U.S.) ensured consistency in data management regulations and mandates, and their impact on DQA practices in RDRs. This approach simplified the study's logistics by reducing the risk of misinterpretation or miscommunication due to language barriers and cross-country differences in human subject protection practices. It also provided a foundation for future cross-country comparisons.

The interview data used by this study represented 32 repositories and 30 universities in the US. 29 of these universities were R1 universities, and one was an R2 university. Some universities hosted or operated multiple repositories. Some other universities did not host their digital data collections in local repositories. Instead, they used external repository platforms (e.g., Dryad and Dataverse). If such a university provided significant data curation support to its researchers, we still counted its digital data collection(s) on the external data curation platform as an instance of an RDR. Out of the 32 repositories, 27 were generalist or domain-agnostic, while the remaining 5 were domain-specific. These domain-specific repositories focused on the social sciences (3), biology (1), and applied science and engineering (1). 59%(19) interview participants were female

and 41%(13) were male. 72%(23) of interviewees had a Master's degree and 28%(9) had a Ph.D. degree. Also, as was expected, the largest share of participants reported library and information science as the discipline of their highest degree (15). The disciplines reported by the participants included psychology, political science, computer science, biology, English, history, anthropology, social work, ecology, journalism, and geography.

The study was guided by a theoretical framework that comprised activity theory [11] and the information quality evaluation framework [21]. Activity theory is a psychological theory of a purposeful activity structure that was originally developed by Lev Vygotsky and his students and was later expanded by Yrjö Engeström. It consists of several conceptual models that can be applied to analyze, explain, and/or predict relationships in complex, real-world activity systems. Its application allows the identification of problems and opportunities for new interventions. The core model of activity theory defines the fundamental structure of activity by emphasizing the relationship between the subject (the individual or group engaged in the activity) and the object (the objective or purpose driving the activity). This subject-object relationship is further organized hierarchically into goal-directed actions, which are mediated by tools, and the organization and community through rules, conventions, and division of labor [11]. In this study, we applied activity theory to examine the underlying structure of DQA activities and problem types in RDRs. Activity theory's contradictions typology categorizes contradictions within an activity system into four different levels (see Figure 1). First-level contradictions pertain to issues within individual elements of an activity. Second-level contradictions are defined as the tensions or problems that occur between the components of an activity. Third-level contradictions indicate tensions between the current and the desired or emerging forms of the activity. These contradictions emerge when there is a need for a revised activity structure, more advanced activity objective or outcome.

Conversely, fourth-level contradictions involve problems that occur between different activity systems and affect the achievement of their shared outcome [23].

The typology helps not only explain, predict, and categorize those problems in an activity but also suggests possible resolutions through their theoretical analysis using the assigned contradiction categories' structures [23]. Misalignments or tensions in an organization's process structure lead to a perceived problem that the organization tries to identify, articulate, and resolve [38]. Activity theory can



help with that. We used the information quality evaluation framework to analyze how data quality is evaluated in those repositories. The framework includes a taxonomy of information quality dimensions and typologies of information quality problems and their impact on activities.

We used thematic content analysis to analyze the content of interview transcripts. The units of analysis were a repository, an activity, and a contradiction. We used the theoretical framework's concepts and research questions to develop a priori codes and then analyzed the content of the data for both the predefined and emerging codes iteratively. We then inductively mapped and merged the thematic codes found in the data into general categories that matched the research questions and high-level concepts of the guiding theoretical framework [14]. Two coders coded the data. Each coder coded half of the data. Following the completion of coding, the coders met to review their coding. They identified and discussed the cases where they had differing opinions, resolved the differences, and made updates to the related code assignments.

## **5. Findings and Discussion**

RDRs are complex sociotechnical systems and may comprise multiple activities shaped by the needs of various stakeholders. These may include researchers, curators, university administrators, scholarly associations and communities of practice, and governments. Our study identified three DQA activities in RDRs: data and metadata quality evaluation, intervention, and communication. The most frequently disclosed DQA activities performed by RDRs were the evaluation and intervention followed by the communication activity.

The first activity of a research DQA process is the evaluation of data and metadata quality. Participants spoke about evaluating datasets for missing and invalid values, meeting specific community standards and best practices, and ensuring the legal and ethical use of the data. They emphasized the importance of evaluating datasets for the completeness of documentation to enable users to understand and effectively use the research data. They revealed that the evaluation was often an iterative and collaborative process, where curators and/or data repository teams engaged in back-and-forth interactions with depositors to enhance data quality. An intervention activity follows the quality evaluation process to tackle the data and metadata quality issues identified during the evaluation. According to participants, the interventions generally involved



minimal changes to the underlying data, such as fixing obvious errors, formatting, and structural improvements within the dataset, or suggesting improvements without directly modifying the data. Participants revealed that, in most cases, researchers were responsible for fixing any identified data quality issues. Intervention activities also included educating researchers about DQA. Our analysis showed that data curators and librarians conducted outreach and workshops to teach researchers how to improve the quality of their data and associated metadata and enhance the usability and downstream impact of the data. Both data quality and intervention activities involved communication and collaboration between the curators and the data providers. There might be back-and-forth exchanges, requests for clarification, and coordination to ensure the necessary changes were made.

A detailed analysis of DQA activities in RDRs and their structures is presented elsewhere [39]. The focus of this paper is limited to identifying mismatches and misalignments in the activities' components, structures, and contexts that lead to barriers and problems affecting their success and interpreting them using the study's theoretical framework. Activity theory theorizes changes and innovations in an activity as an effect of the presence of contradictions and tensions in the activity and efforts to resolve them [11]. This section examines the challenges and problems research data curators and repository managers reported facing in their DQA work and classifies them using activity theory's contradictions typology (see Figure 1). In addition, each category of DQA challenges is accompanied by a *discussion* of the strategies for resolution, as reported by the study participants or suggested by the theoretical framework and relevant literature (see Table 1).

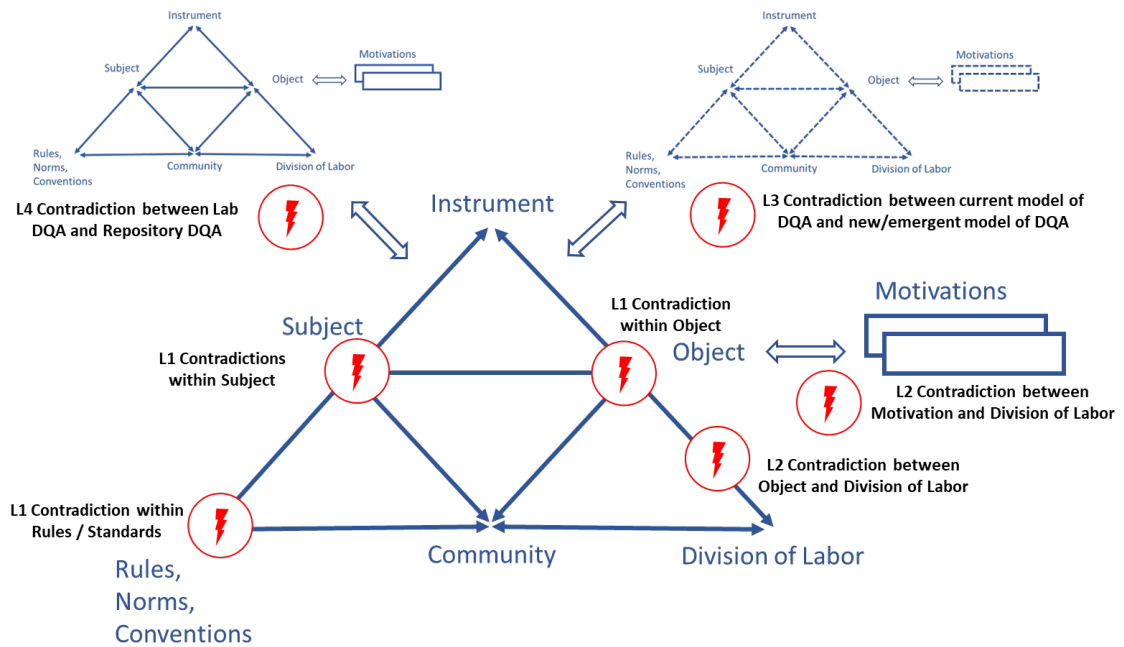


Figure 1. The levels of the DQA contradictions per activity theory's contradictions typology.

## Challenges of defining and measuring quality

A key issue noted by several participants was the challenge of defining data quality. Participants who worked for domain agnostic RDRs discussed the difficulty of determining what constituted quality and how to assess it accurately and consistently across different datasets from different domains. This challenge can be categorized as a First-level contradiction (Table 1, row 1).

I think the biggest one is how you measure quality and what we mean by quality. So, how do you capture that? How do you measure that in an equitable way across a lot of different, wildly different data sets? (i12).

Furthermore, the providers and curators of a DQA process may have different understandings of its objective – what good quality data and metadata are, which is a Second-level contradiction according to activity theory. One participant noted that they evaluated the data specifically for its usability and did not evaluate the research that produced it. The participant argued that the concept of data quality was challenging to communicate effectively and ensure that people grasp its intended meaning in a particular context "I think quality is a hard word to make sure people understand" (i16; Table 1, row 6).

Determining what data quality means (i.e., identifying the quality criteria or characteristics a stakeholder community(s) perceives quality through or cares about) is the first step in a data quality evaluation process. The next step is to be able to measure it consistently across different datasets and providers (i.e., developing quality metrics). For that, one needs stable, community approved reference sources [21]. Interview participants spoke about the challenges they faced in evaluating the quality of data and metadata in some research disciplines or areas that lacked clear, widely adopted standards for sharing and documenting research data. The absence of such standards made it challenging to develop consistent metrics for evaluating the quality of datasets and their accompanying metadata (Table 1, row 2).

### **Problem resolution strategies and discussion**

Thus, the first problem theme identified by the content analysis was the challenge of defining the objective of an RDR's DQA process – that is, defining what good quality data and metadata meant. According to activity theory, an activity's object is both the driver (i.e., motive) and the direction of the activity. It is shaped by an interplay of the motivations and priorities of the organization's stakeholder groups, accompanied by tensions among those priorities [11,42]. Hence, the theory predicts such a challenge for generalist repositories that might have multiple stakeholder communities.

One approach to resolving this problem would be adopting and using multiple, domain-specific models of DQA, including data quality definitions. The approach, however, would require the RDR staff to possess knowledge of those domains and communities. Alternatively, the RDR could identify a common denominator level of DQA for the domains represented by the datasets it curates. This approach has a limitation, however. If the domains are too disparate, such a common dominator level of DQA might not exist or not be adequate for meeting some of the stakeholders' needs for quality. Indeed, a few participants expressed disappointment that they could not achieve their DQA objectives due to a lack of subject expertise or resources and had to settle for satisficing or partial DQA targets instead.

Another related problem subtheme identified by the study was the differences in understanding a DQA process between the curator and dataset provider. Although most of the RDRs, especially domain agnostic RDRs, did not evaluate the research process that created a dataset, they might still evaluate some of the

intrinsic quality dimensions of the dataset, such as completeness (e.g., identifying missing values). Many also ensured that the dataset was usable and accompanied by adequate documentation and support material, such as its software code, to allow the end-users to reproduce or replicate the study and thus evaluate its scientific quality. The data quality literature also predicts this possible misalignment between data providers' and curators' understandings of data quality. Data providers may prioritize the scientific quality of their research. They may view publications as primary products of their research and the data itself as a byproduct. RDRs and curators, on the other hand, may view datasets as the main information products [42]. Consequently, curators may focus on ensuring datasets' product level quality characteristics, such as findability, accessibility, usability, and interoperability [9,39]. Potential mitigation of this misalignment could involve explaining the scope of the RDR's DQA process to the data providers and highlighting how treating datasets as primary products of their research could improve the visibility and impact of their research.

Interviewees suggested that engaging relevant scholarly societies to establish and advocate for clear, specific standards for documenting and managing research data could be a potential solution to the lack of community-approved reference sources for consistent assessment of data and metadata quality (see Table 1). Another resolution for the problem employed by some participants was to assemble local operational metadata profiles for documenting datasets from disciplines that did not have a widely adopted metadata vocabulary. The profile could be determined based on metadata term occurrence statistics in the domain's datasets and metadata descriptions. This is a traditional method of metadata schema or vocabulary construction [43,44]. The success of this approach depends on the repository having access to an adequate number of datasets and metadata records from the domain. Many, especially newly established repositories, may not have that access. Trustworthy Repositories Audit and Certification (TRAC) Criteria that are widely used for auditing and certifying RDRs require an RDR to specify "minimum metadata requirements to enable the designated community(ies) to discover and identify material of interest," implying the RDR has the knowledge of the community(ies)' metadata needs [45]. However, the TRAC model does not address the contingency when there are no community approved standards for knowledge organization [45]. It is important that the discipline's scholarly societies, communities of practice of data curators, and publishers are involved in the creation and promotion of domain-specific data and

metadata standards. As the literature shows, the social and political aspects of a metadata standard design and adoption are as crucial as its quality and representational soundness [46].

Data curation, encompassing DQA, is inherently a social process. A DQA process' objective must strike a balance among various motivations and interpretations of data quality [41]. Ultimately, DQA objects (i.e., objectives) should emerge from negotiation and compromise among the stakeholders of RDRs. In defining the focus of their DQA efforts, RDRs should take into account both the data quality models outlined in the literature (e.g., [21,42]) and the inherently negotiated, social nature of their DQA objectives. Kaptelinin's general characteristics of a successful activity object/objective can be a good starting point for the latter. These include a balance that ensures a proper *representation* of varying motives, *inspiration* where the object is not only feasible but also attractive and energizing, *stability* to prevent frequent changes, and *flexibility* to allow necessary changes and avoid obsolescence [47].

### **Insufficient resources and expertise**

Another group of challenges was associated with curators not having enough resources to provide the desired level of DQA of submitted datasets. If the repository does not have a large enough staff, curators cannot spend sufficient time on the dataset to ensure its quality (Table 1, row 3).

In terms of manpower sustainability, if we were to receive more deposits, we'd have to cut back the time spent on each dataset. Despite our quality principles and purposes, our resources are limited (i14).

An additional resource-related challenge was limited access to essential software and technology infrastructure for DQA. Curators without the necessary software to open specific data files faced uncertainty regarding data validity.

Participants also highlighted a lack of expertise as a primary barrier in research DQA. Curators and managers at generalist RDRs acknowledged challenges working with data from unfamiliar fields. They emphasized the need for domain expertise, research experience, statistical proficiency, and familiarity with new data formats (Table 1, row 4). One interviewee noted:

For newer data types like images, determining data quality measures and making data more usable is a major challenge (i8).

Some participants identified the insufficient data management expertise among data providers as a challenge. Since research DQA involves collaboration between the researcher and the curator, the absence of DQA competencies on the provider's part can negatively impact the effectiveness and cost of that collaboration (Table 1, row 4).

### **Problem resolution strategies and discussion**

The lack of resources can be mitigated by reducing the cost of DQA. The analysis showed different strategies RDRs used to reduce the cost of DQA (see Table 1). Some RDRs enhanced the quality of datasets on demand by applying quality enhancement actions to the datasets requested by end-users. The literature has also highlighted the importance of prioritizing data curation targets. For instance, Gene Ontology curators have prioritized their assessment of new entries in the literature to manage their workforce shortage [48]. In addition, RDRs used automated scripts to identify the most prevalent problems in datasets and generate some of the dataset's metadata automatically. The availability of Generative AI-based tools can further reduce the cost of data quality evaluation and enhancement by improving the accuracy and completeness of automated data quality profiling and metadata generation [49].

Likewise, participants disclosed that they evaluated dataset metadata on criticality and then prioritized the critical metadata elements when evaluating submitted datasets and making intervention requests to researchers. Furthermore, participants stated that they relied on communities of practice (e.g., the Data Curation Network (DCN)<sup>2</sup>) and used their members' collective knowledge and resources, such as documentation templates, to overcome the lack of necessary DQA expertise or resources within their respective RDRs. Finally, educating the administration about the importance of research DQA and collaborating with other research support units on campus, such as the office of research or high-performance computing, can make DQA effective and efficient and strengthen the RDR's appeal for more resources [50,51]. It is important to note that while the TRAC guidelines specify the types of processes and evidence required for RDRs

---

<sup>2</sup> <https://datacurationnetwork.org/>

to demonstrate financial sustainability, they do not offer specific guidance on how to address financial sustainability challenges [45].

Participants also shared strategies they used to address subject expertise gaps in DQA (see Table 1, row 4). They stressed the importance of comprehensive documentation accompanying data submissions. Some RDRs accepted only peer-reviewed datasets, those with technical papers, or datasets from trusted sources. Additionally, some organized their own expert peer reviews. They also highlighted attending professional development events and leveraging expertise from communities of practice, such as DCN, as effective approaches. One of the TRAC requirements is for RDRs to offer robust professional development opportunities to their staff to develop essential competencies [45]. Indeed, enhancing skills and expertise is one of the major motivations for data curators who participate in a community of practice [52]. Another strategy to address the issue of limited domain knowledge locally in a specific area is for RDRs to consider partnering with relevant scholarly societies, research communities, and publishers [13].

### **Misalignments in motivations**

Participants highlighted a significant challenge in motivating researchers to engage in DQA. This issue often stemmed from misaligned motivations between data curators and data curators. Researchers sometimes perceived DQA as burdensome, leading to hesitancy in participating in DQA and hindering data quality improvement.

It's a concern I have that the time sometimes required to work with us to improve the data is actually the main thing that prevents researchers from working with us to improve their data (i30).

Participants raised concerns about the difficulty of balancing the demand for quick data publication and the necessity of maintaining their data quality standards. Some researchers prioritized releasing their data quickly and obtaining a Digital Object Identifier (DOI), which could compromise the RDR's DQA process (Table 1, row 5).

Another challenge the study found was the misalignment of DQA practices during different phases of a dataset's curation lifecycle. Before data is deposited for



curation, it must be generated. A lab or research project's DQA practices play a crucial role in determining the quality of the data it generates and deposits to an RDR. Participants noted that this misalignment could make the curation of the data more difficult and expensive (Table 1, row 8).

Participants spoke about their difficulties in persuading researchers to change their existing data management practices. Participants disclosed that, in general, junior researchers tended to appreciate the advice and guidance curators provided. In contrast, older researchers following established practices for many years saw the additional DQA requirements as an annoyance. Participants mentioned a challenge when researchers often did not understand why curators approached them.

The issue of quality assurance is highly problematic. Because it puts us into some kind of position of authoritative judgment where I think the community we try to help doesn't necessarily see us (i23).

### **Problem resolution strategies and discussion**

Participants adopted several strategies to enhance researchers' motivations to participate in DQA and reduce their reluctance due to the perceived costs of DQA (Table 1, row 5). A common strategy was to alleviate the burden on researchers by performing preparatory tasks on their behalf. For instance, participants mentioned using metadata templates pre-populated with information from data submissions, providing researchers with a useful starting point. This approach often captured researchers' attention and led to more positive outcomes, as they were more willing to complete missing information and build upon the provided material. As one participant explained: "An approach I might take is if somebody submits a dataset, and there's no information about the methodology, I will track down the article if there is one. I will read the methodology, and I'll read the article, and I will make a suggested text to describe the methodology because they're much more likely to correct an error in the text that I've written than they are to write up the methodology themselves" (i30).

Additionally, reducing the cost of DQA for researchers involved curators providing them with guidance in the form of sample metadata and hints. These resources helped researchers understand how to describe their datasets and navigate repository requirements and expectations effectively. Participants also

emphasized the importance of not overwhelming researchers with numerous DQA requests at once. Instead, their strategy was to elicit an initial response while maintaining researcher engagement. One participant noted, "We try to prioritize and not ask for too many things at once. If we send an email with three or four questions, we're likely to get a response. But if we send a wall of text with 10 different questions, often, you just don't hear back" (i32). These strategies are similar to some practical approaches to work articulation, coordination, and quality control found in the literature on peer-curation communities (e.g., [53,54]).

Furthermore, participants employed persuasive communication strategies to enhance researchers' motivations for improving dataset quality. They emphasized the benefits of publishing high-quality data, mainly how it could increase the value of associated publications for citation by other researchers. This approach is aligned with research information management system (RIMS) managers' use of persuasive communication to boost both intrinsic and extrinsic motivations for contributing to research information curation [12].

Some repositories employed a combination of incentives and consequences to encourage researcher participation in DQA. Incentives included providing small grants to researchers to enhance dataset quality and associated metadata. Conversely, repositories sometimes withheld dataset publication and a DOI until researchers cooperated with DQA requirements. Finally, participants proposed that universities could enhance researcher motivation for DQA participation by implementing specific extrinsic incentives at the policy level, such as counting the creation and publication of high-quality datasets toward researchers' promotion and tenure, as suggested by the literature [31].

According to activity theory, a misalignment between the providers' and the RDRs' DQA practices can be classified as a fourth-level contradiction (see Table 1, row 8). Participants shared strategies to address or lessen the issue. The RDR and its staff could proactively engage in curating research data from a laboratory or a research project, beginning in the planning phase, as has also been suggested by the literature [13]. Moreover, the RDR could offer its data management and DQA infrastructure and services (e.g., assistance with writing data management plans; DMP Tool<sup>3</sup>) to the laboratory and utilize that infrastructure to align the

---

<sup>3</sup> <https://dmptool.org/>

laboratory's data curation processes with those of the RDR.

### **Tensions between the current DQA practices and infrastructure and new regulations and mandates**

Participants identified challenges related to the changing landscape of data curation brought about by new regulations, such as those mandating open access to federally funded research data and ensuring its quality [28,40]. Participants were apprehensive about the entry of new groups of researchers into the data curation ecosystem triggered by the regulations. Using new and evolving data formats and adhering to distinct data management practices, these groups of providers might require DQA services that differ from those traditionally offered by RDRs (Table 1, row 7).

Participants also noted having inadequate human and technical infrastructures for handling the increased data submissions and quality assurance requirements caused by the new regulations. They expressed concerns about whether the current data management and DQA technologies were scalable and robust enough to keep up with the expected increase in the number and size of deposited datasets.

We get more and more policies that require data sharing, which is great. At the same time, I have a little bit of a sense of dread because I wonder how much the cart is ahead of the horse. Hopefully, technology will catch up because that's going to be so difficult (i5).

### **Problem resolution strategies and discussion**

New innovative technologies allow organizations and society to engage in new forms of activities [47,55]. New technologies also have had dramatic and often unintended consequences on the ways in which information systems have been conceived, designed, implemented, and managed. Archivists and curators have long been aware of new technologies and data types' impact on the structure and roles of digital curation and archiving activities [56]. Generative AI advances, and new data and knowledge types (e.g., big data and LLMs) impact how RDRs are designed or should be designed and/or operated [57]. For instance, how can RDRs preserve, curate, and ensure the quality of large machine learning foundation

models and associated datasets that can cost hundreds of millions of dollars to create, contain billions of parameters and trillions of tokens, and are trained on gigabytes of data?<sup>4</sup> [58]. Individual universities or academia cannot solve those challenges alone. The government, industry, and research communities need to support and assist RDRs in developing effective evaluation frameworks, models, workflows, technologies, and funding mechanisms to address these issues (see Table 1, row 7). These could be accomplished by research funding agencies organizing “future directions” workshops and Delphi studies to define new adequate curation and DQA models for the emerging types of data and metadata as well as help establish associated shared organizational and computational infrastructure components (e.g., [35,59-63]).

Participants also disclosed several strategies they used to mitigate these contradictions (see Table 1, row 7). The RDRs of some smaller universities or universities with less established research data management infrastructure joined consortia led by larger universities to benefit from their shared infrastructure. Other RDRs cooperated with high-performance computing units on campus to gain access to scalable data storage and computing infrastructure and technology expertise. The literature also reports a similar approach where an RDR manager utilized a system with linked scalable storage for storing large data files [7].

Table 1. DQA contradictions and their resolutions. Note: *The prefix Lx is used to indicate the level of a particular contradiction according to activity theory.*

#	Problems	Resolutions
1	L1 (Object): RDR’s DQA Object is not clear. A generalist RDR receives different types of datasets from different domains. That makes it challenging to formulate DQA action goals and define what successful DQA means for RDR.	Determine and use the least common denominator level of curation and DQA Adopt multiple data type and/or domain specific definitions of data quality and DQA success from domain specific repositories
2	L1 (Rules, Standards) The community has no established/widely adopted data file and metadata standards. It makes it challenging to determine what reference baselines to use in DQA	Assemble the active/operational metadata schema for the community Engage relevant scholarly societies to establish and advocate for clear, specific standards for documenting and managing research data.

<sup>4</sup> <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/>

3	L1 (Subject): RDR lacks resources to provide adequate DQA	Optimize/reduce the cost of DQA activities Automate some of the DQA tasks Participate in a Community of Practice (e.g., DCN) Educate the administration and partner with other research support units on campus to get more resources Outsource some of the DQA to a consortium (e.g., Dataverse, Dryad)
4	L1 (Subject): RDR does not have the necessary domain, research, or statistical expertise, or knowledge of new data formats to provide the desired level of DQA RDR provider does not have the knowledge of data management and DQA principles	Stricten dataset acceptance screening criteria (e.g., accept data from predetermined trustworthy sources only; require comprehensive documentation for datasets) Organize DQ peer review locally. Participate in a community of practice (e.g., DCN) Partner with relevant scholarly societies and research communities Outsource DQA to Dryad or similar non-profit consortia Educated providers on data management and DQA
5	L2 (Division of Labor - Motivation): Providers and curators' motivations are misaligned, impeding their cooperation in the division of labor. The cause can be the researchers' perceived cost of participation in DQA. They need a DOI for the dataset without delay and with the least cost.	Curators engage researchers and reduce their DQA costs by doing some preparatory work beforehand and not asking too many questions at a time. Provide researchers with sample metadata and documentation hints to facilitate their sensemaking. Use persuasive communication to highlight the value of publishing high-quality datasets for enhancing researchers' impact. Refuse to give the DOI until the researcher cooperates on the DQA of the dataset. Count producing high-quality datasets toward P&T.
6	L2 (Division of Labor – Object): The curator and the researcher have different understandings of the scope of the Object of DQA.	Educate the researcher about DQA in general and the scope of the RDR's DQA process Ensure that the dataset is accompanied by adequate documentation and support material, such as code, to allow the end-users to reproduce or replicate the study and thus evaluate the science behind it
7	L3 (Current model of DQA – Emergent model of DQA): There is a realization that the current DQA practices and RDR's infrastructures are not adequate for the emerging data types and new groups of researchers' data curation needs. However, it is not clear what the new DQA model should be.	The government, industry, and scholarly societies help determine the new model of research DQA by organizing "future directions" workshops and providing adequate funding and shared research data management infrastructure to RDRs
8	L4 (Lab DQA – Repository DQA): The lab's DQA practice is not aligned with the standardized DQA model and practices promoted by the RDR.	The RDR inserts itself in the data curation and DQA process of the lab from the beginning, from the planning stage. The RDR to use the shared data curation infrastructure and technology to harmonize the research data curation practice of the lab with one of the RDR The RDR uses persuasive communication to engage researchers and convince them to switch to the standardized DQA model promoted by the RDR

## 7. Conclusion

This study examined the types of problems and tensions and their resolution strategies within the DQA practices of RDRs and interpreted and discussed them through the lens of activity theory and the literature. The analysis identified differences in the understanding of DQA among different stakeholders, the need for adequate resources, domain-specific knowledge, research expertise, and standardized metrics as challenges faced by RDR staff in their DQA work. The study also found tensions caused by misalignment in motivation between RDR

staff and depositors. It revealed frictions between the DQA practices of research labs and the standardized model of DQA promoted by RDRs. Additionally, there were tensions between RDRs' infrastructures and curation model and the emerging needs of stakeholders. These needs were shaped by new government regulations, which mandated open access to federally funded research data and new data types. Furthermore, the paper outlined resolution strategies for each identified contradiction, drawing from insights shared by participants and recommendations found in the literature (Table 1). They emphasized the need for standardized data quality practices, collaborative efforts with communities of practices and scholarly societies, and the adoption of automation to optimize DQA processes. Education and persuasive communication were stressed as crucial for engaging researchers and administrative bodies. Additionally, the strategies emphasized the importance of supporting researchers with adequate resources, recognizing high-quality data contributions, and defining future directions and funding for research DQA.

The future related study will examine researchers' perspectives on the contradictions identified by this study, as well as their emerging needs for DQA services from RDRs in light of the new government regulations and the introduction of new types of research data by new AI-based technologies.

## **8. Acknowledgment**

This research is supported by a National Leadership Grant from the Institute of Museum and Library Services (IMLS) of the U.S. Government (No: LG-252346-OLS-22). This article reflects the findings and conclusions of the authors and does not necessarily reflect the views of IMLS.

## **References**

1. Mason RO. Four ethical issues of the information age. *MIS Q.* 1986;10(1):5-12.
2. Gururangan S, Card D, Drier SK, et al. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. 2022. arXiv preprint arXiv:2201.10474

3. Scheuerman MK, Hanna A, Denton E. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc ACM Hum-Comput Interact.* 2021;5(CSCW2):1-37.
4. Stvilia B, Gibradze L. Seeking and sharing datasets in an online community of data enthusiasts. *Libr Inf Sci Res.* 2022;44(3):101160.
5. Ball A. Review of data management lifecycle models. University of Bath, IDMRC; 2012.
6. Higgins S. The DCC curation lifecycle model. *Int J Dig Curation.* 2008;3(1):134-140. <http://ijdc.net/index.php/ijdc/article/view/69>
7. Lee DJ, Stvilia B. Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE.* 2017;12(3):e0173987..
8. Lord P, Macdonald A. E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision. Bristol, UK: The JISC Committee for the Support of Research; 2003.  
<http://www.jisc.ac.uk/media/documents/programmes/preservation/esciencereportfinal.pdf>
9. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3(1):1-9.
10. The DataONE Webinar Series. Assuring the quality of your data: A natural history collection community perspective. The DataONE Webinar Series. <https://www.dataone.org/webinars/assuring-quality-your-data-natural-history-collection-community-perspective/> (2020, accessed 21 February 2024).
11. Kaptelinin V, Nardi B. Activity theory in HCI: Fundamentals and reflections. *Synth Lect Hum Cent Inform.* 2012;5(1):1–105.
12. Stvilia B, Lee DJ, Han N. "Striking out on your own" - A study of research information management problems on university campuses. *J Assoc Inf Sci Technol.* 2021;72(8):963-978.
13. Wu S, Worrall A, Stvilia B. Exploring data practices of the earthquake engineering community. *iConference 2016 Proceedings.* 2016 Mar 15.
14. Bailey KD. *Typologies and taxonomies: An introduction to classification techniques.* Sage; 1994 Jun 13.
15. Juran J. *Juran on quality by design.* New York: The Free Press; 1992.



16. De Roure. Replacing the paper: The twelve Rs of the e-Research. Nature Blog. 2010 Nov 27.
17. Faniel IM, Frank RD, Yakel E. Context from the data reuser's point of view. J Doc. 2019;75(6):1274–1297.
18. Gutmann M, Schürer K, Donakowski D, Beedham H. The selection, appraisal, and retention of social science data. Data Sci J. 2004;3(0):209–221.
19. Stvilia B, Hinnant C, Wu S, Worrall A, Lee DJ, Burnett K, Burnett G, Kazmer MM, Marty PF. Research project tasks, data, and perceptions of data quality in a condensed matter physics community. J Assoc Inf Sci Technol. 2015;66(2):246-263.
20. York J. Seeking equilibrium in data reuse: A study of knowledge satisficing. PhD Thesis, University of Michigan, US, 2022. DOI: 10.7302/6170
21. Stvilia B, Gasser L, Twidale MB, Smith LC. A framework for information quality Assessment. J Am Soc Inf Sci Technol. 2007;58(12):1720-1733.
22. Stodden V. Re-use and reproducibility: Opportunities and challenges. Open Repositories. <http://or2013.net/sites/or2013.net/files/OR2013-July92013-STODDEN.pdf> (2023, accessed 25 April 2024).
23. Zhou L, Divakarla M, Liu X. An overview of the Joint Polar Satellite System (JPSS) science data product calibration and validation. Remote Sens. 2016;8(2):139.
24. National Academies of Sciences, Engineering, and Medicine (NASEM). Advancing Open Science Practices: Stakeholder Perspectives on Incentives and Disincentives: Proceedings of a Workshop—in Brief. <https://www.nationalacademies.org/our-work/roundtable-on-aligning-incentives-for-open-science> (2020, accessed 17 April 2024)
25. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLoS One. 2020;15(3):e0229003.
26. National Academies of Sciences, Engineering, and Medicine (NASEM). Reproducibility and replicability in science. National Academies Press; 2019.
27. National Academies of Sciences, Engineering, and Medicine (NASEM). Reproducibility and replicability in science. National Academies Press; 2019.
28. Nelson A. 2022. OSTP Memo: Ensuring free, immediate, and equitable access to federally funded research. <https://www.whitehouse.gov/wp->

- content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf (2022, accessed 17 March 2024).
29. Barrett C. Are the EU GDPR and the California CCPA becoming the de facto global standards for data privacy and protection?. *Scitech Lawyer*. 2019;15(3):24-29.
  30. U.S. Congress. An Act to Protect Investors by Improving the Accuracy and Reliability of Corporate Disclosures Made Pursuant to the Securities Laws, and for Other Purposes. The Sarbanes-Oxley Act. 2002. 107th Cong., H.R. 3763.
  31. Lyon L. The informatics transform: Re-engineering libraries for the data decade. *Int J Dig Curation*. 2012;7(1):126–138.
  32. Cragin MH, Palmer CL, Carlson JR, Witt M. Data sharing, small science and institutional repositories. *Philos Trans R Soc A*. 2010;368(1926):4023-4038.
  33. Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf Commun Soc*. 2012;15(5):662-679.
  34. Ng A. AI does not have to be too complicated or expensive for your business. *Harvard Business Review*. <https://hbr.org/2021/07/ai-doesnt-have-to-be-too-complicated-or-expensive-for-your-business> (2021, accessed 12 April 2024).
  35. Swarup S, Braverman V, Arora R, Caragea D, Cragin M, Dy J, ... & Yang C. Challenges and opportunities in big data research: Outcomes from the second annual joint pi meeting of the NSF big data research program and the NSF big data regional innovation hubs and spokes programs 2018. *NSF Workshop Reports*. June 2018.
  36. Borgman CL, Wallis JC, Enyedy N. Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*. 2007;7:17-30.
  37. Chen S, Chen B. Practices, challenges, and prospects of Big Data curation: A case study in geoscience. *Int J Data Curation*. 2020;14:275–291.
  38. Simon HA. *The sciences of the artificial*. MIT Press; 1996.
  39. Stvilia B, Lee DJ. Data quality assurance in research data repositories: a theory-guided exploration and model. *Journal of Documentation*. 2024;80(4):793-812.

40. National Science and Technology Council (NSTC). Desirable characteristics of data repositories for federally funded research.  
<https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf> (2022 May, accessed 19 February 2024).
41. Nardi BA. Objects of desire: Power and passion in collaborative activity. *Mind Cult Act.* 2005;12(1):37–51.
42. Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. *J Manage Inf Syst.* 1996;12(4):5-33.
43. Soergel D. Indexing languages and thesauri: Construction and maintenance. Los Angeles, CA: Wiley; 1974.
44. Stvilia B, Gasser L. Value-based metadata quality assessment. *Libr Inf Sci Res.* 2008;30(1):67-74.
45. TRAC Metrics (TRAC). Center for Research Libraries. TRAC Metrics.  
<https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac> . (2024, accessed 20 May 2024).
46. Starr J. Information Politics: The Story of an Emerging Metadata Standard. *First Monday.* 2003;8(7).  
<https://firstmonday.org/ojs/index.php/fm/article/view/1065/985>
47. Kaptelinin V. The object of activity: Making sense of the sense-maker. *Mind Cult Act.* 2005;12(1):4-18.
48. Wu S. Exploring the data work organization of the Gene Ontology. PhD Thesis, Florida State University, US, 2014.  
[http://purl.flvc.org/fsu/fd/FSU\\_migr\\_etd-9267](http://purl.flvc.org/fsu/fd/FSU_migr_etd-9267)
49. Khan R, Gupta N, Sinhababu A, Chakravarty R. Impact of conversational and generative AI systems on libraries: A use case large language model (LLM). *Science & technology libraries.* 2023; Sep 11:1-5.
50. Arora R, Esteva M, Trelogan J. Leveraging high performance computing for managing large and evolving data collections. 2014.
51. Eppler MJ. Managing information quality: Increasing the value of information in knowledge-intensive products and processes. Springer Science & Business Media; 2006.
52. Han NE. Building a Community of Practice of Research Data Curators-A Qualitative Study. The Florida State University; 2022.  
[https://purl.lib.fsu.edu/diginole/2022\\_Han\\_fsu\\_0071E\\_17128](https://purl.lib.fsu.edu/diginole/2022_Han_fsu_0071E_17128)

53. Bearman D, Trant J, Chun S, et al. Social terminology enhancement through vernacular engagement: Exploring collaborative annotation to encourage interaction with museum collections. *D-Lib Magazine*. 2005;11(9):26.
54. Nov O. What motivates Wikipedians? *Commun ACM*. 2007;50(11):60-64.
55. Orlikowski WJ. Using technology and constituting structures: A practice lens for studying technology in organizations. *Organ Sci*. 2000;11(4):404-428.
56. Bearman D. An indefensible bastion: archives as repositories in the electronic age. *Archival management of electronic records*. 1991;13:14-24.
57. Yoon A, Kim J, Donaldson DR. Big data curation framework: Curation actions and challenges. *J Inform Sci*. 2022; DOI: 01655515221133528.
58. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. 2021. arXiv preprint arXiv:2108.07258.
59. Fabric Portal. <https://portal.fabric-testbed.net/> (2024, accessed 29 May 2024).
60. Greenberg J. Big Metadata, Smart Metadata, and Metadata Capital: Toward Greater Synergy Between Data Science and Metadata. *J Data Inf Sci*. 2017;2(3):19-36. <https://doi.org/10.1515/jdis-2017-0012>.
61. NSF Research Infrastructure Office (NSF RIO). 2024 Research Infrastructure Workshop. <https://researchinfrastructureoutreach.com/2024-research-infrastructure-workshop/> (2024, accessed 14 February 2024)
62. Office of Science and Technology Policy (OSTP). 2023. Report to the U.S. Congress on Financing Mechanisms for Open Access Publishing of Federally Funded Research. Washington, DC, USA.
63. Pasek JE. Historical development and key issues of data management plan requirements for National Science Foundation grants: a review. *Issues Sci Technol Librariansh*. 2017;87:1-1.