This is a preprint of an article accepted for publication in the Journal of Documentation: Stvilia, B. & Lee, D.J. (2024). Data quality assurance in research data repositories: A theory-guided exploration and model. *Journal of Documentation*. https://doi.org/10.1108/JD-09-2023-0177

Data Quality Assurance in Research Data Repositories: A Theory-Guided Exploration and Model

Besiki Stvilia[1], Dong Joon Lee[2],

[1]School of Information, Florida State University

[2]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. ORCID iD https://orcid.org/0000-0002-2428-6627.

Dong Joon Lee, Mays Business School, Texas A&M University, 4113 TAMU, College Station, TX 77843-4113. E-mail: djlee@tamu.edu. ORCID iD https://orcid.org/0000-0001-8994-163X.

Abstract

This study addresses the need for a theory-guided, rich, descriptive account of research data repositories' (RDRs) understanding of data quality and the structures of their data quality assurance (DQA) activities. Its findings can help develop operational DQA models and best practice guides and identify opportunities for innovation in the DQA activities.

The study analyzed 122 data repositories' applications for the Core Trustworthy Data Repositories, interview transcripts of 32 curators and repository managers, and data curation-related webpages of their repository websites. The combined dataset represented 146 unique RDRs. The study was guided by a theoretical framework comprising activity theory and an information quality evaluation framework.

The study provided a theory-based examination of the DQA practices of research data repositories summarized as a conceptual model. We identified three DQA activities: evaluation, intervention, and communication and their structures, including activity motivations, roles played, and mediating tools and rules and standards. When defining data quality, study participants went beyond the traditional definition of data quality and referenced seven facets of ethical and effective information systems in addition to data quality. Furthermore, the participants and RDRs referenced thirteen dimensions in their DQA models. The study revealed that DQA activities were prioritized by data value, level of quality, available expertise, cost, and funding incentives.

The study's findings can inform the design and construction of digital research data curation infrastructure components on university campuses that aim to provide access not just to big data but trustworthy data. Communities of practice focused on repositories and archives could consider adding FAIR operationalizations, extensions, and metrics focused on data quality. The availability of such metrics and associated measurements can help reusers determine whether they can trust and reuse a particular dataset. The findings of this study can help to develop such data quality assessment metrics and intervention strategies in a sound and systematic way.

To the best of our knowledge, this paper is the first data quality theory guided examination of DQA practices in RDRs.

1. Introduction

The ethical implications of data quality are undeniable (Mason, 1986). The quality of data and information directly influences the effectiveness of our decisions, the results of our activities, and, ultimately, our lives, personal respect, and reputation. Therefore, ensuring data quality (DQA) is an essential element of all data management processes. DQA tasks can vary widely, including quality assessments and enhancement efforts undertaken by data providers and personnel, data cleansing by students for class projects or during DQA hackathons, evaluating the quality of data sets used in training AI models, or policy and business decision making (Gururangan et al., 2022; Scheuerman et al., 2021). Several general quality assurance standards and strategies, like ISO 8000, ISO 9000, and ISO 19157, are commonly used in the industry. Similarly, the literature offers numerous studies and models on data curation (e.g., Ball, 2012; Burton and Treloar, 2009; Higgins, 2008; Lee and Stvilia, 2017, Lord and Macdonald, 2003). There's a revived focus on data quality, and the goal of making data sets FAIR, meaning findable, accessible, interoperable, and reusable, within data curation practitioner communities. These groups create and disseminate important methods and scripts for data cleaning, normalization, linking, and disambiguation (Wilkinson et al., 2016; The DataOne Webinar Series, 2020a, 2020b). However, the efforts to operationalize the FAIR framework have largely been lacking in firm roots in the metadata and information quality literature, restricting their broad applicability to DQA process design. There's a dearth of studies that examine and interpret DQA practices in research data repositories through the lens of the data quality literature.

2. Research questions

The understanding of what defines high-quality, useful data, or when such data becomes useful and usable, can differ even within the same process, field, and across different fields (Higgins, 2008; Stvilia et al., 2015). Prior to creating data and metadata quality evaluation standards, measures, and interventions, research data repositories and their stakeholders must establish and agree upon their definition of data quality (DQ), what "fitness for use" or "fitness for reuse" (Juran, 1992) means to them and what the best practices are for ensuring it. Similarly, users of these datasets need to clearly comprehend the repository's DQ model and understand the DQ virtues the repository evaluates and ensures its datasets for to determine if these virtues and DQA actions align with their own DQ needs and preferences. For example, some users might prefer raw, dirty data to hone their data cleaning and organization skills (Stvilia and Gibradze, 2022). Although previous conceptual models of research DQ and studies on researcher priorities and perceptions of DQ exist (e.g., Huang et al., 2012; Stvilia et al., 2015), there's a notable absence of recent analysis of research data repositories' DQA practices that is rooted in the information and DQ literature. Providing A detailed and descriptive explanation of how RDRs perceive data quality, along with a conceptual model summarizing the structure of their DQA work, can aid in the development of context specific operational DQA models and guides for RDRs. Additionally, it can highlight areas for potential innovation in RDRs' DQA practices. This paper presents part of a larger exploratory research study that seeks to fill this gap. Specifically, the paper discusses the following research questions:

1. How do research data repositories define data quality?
2. How do RDRs ensure data quality?

3. Related work

Juran (1992) defines quality as "fitness for use." There have been multiple conceptual models of research data quality and studies of researcher perceptions and preferences for data quality (e.g., Gutmann, et al., 2004; Huang et al., 2012; Stvilia et al., 2015). What is considered quality and useful data, and when such data becomes useful can vary, even within the same procedure, field, and across various procedures within those fields (Higgins, 2008). A DQA process includes activities pertaining to the conceptualization, measurement, and intervention of data quality (Stvilia et al., 2007). Data quality, alongside privacy and access, holds significant ethical implications in data use. In the era of big data, generative AI, and an overwhelming quantity of research data and publications, the saying "garbage in, garbage out" remains as relevant as ever. The quality of data directly influences the quality of research outcomes, teaching, business decisions, policies, and ultimately, human lives (Mason, 1986).

Universities are making considerable investments in constructing reliable and secure infrastructures for managing digital research datasets produced and used by their faculty and students. These efforts are motivated by faculty members' need to preserve and share their research data (NASEM, 2020; Tenopir et al., 2020), requirements from state and federal funding bodies for open data sharing for the advantage of taxpayers, research and teaching purposes, and to boost research reproducibility and replicability (NASEM, 2019, 2020; Nelson, 2022; NSTC, 2022). National and state laws also mandate data quality assurance and the protection of individual privacy (Barrett, 2019; U.S. Congress, 2002). Additionally, some universities are interested in tracking and evaluating the impact of these datasets, including for faculty promotion and tenure assessment

(Lyon, 2012). Concerns about data quality present a significant barrier to data sharing and reuse. Data owners may worry about the quality and documentation of their data, along with its potential misuse or misinterpretation by others (Stvilia et al., 2015). Conversely, users need useful, valid, and reliable data that accurately represent the phenomena they are studying or interested in, rather than merely having access to a plethora of data (Boyd & Crawford, 2012; Ng, 2021). Furthermore, data creators usually gather or compile datasets for specific purposes or uses, and without appropriate documentation, understanding these purposes becomes difficult, hampering data reuse (Swarup et al., 2018).

This study leverages the digital data curation literature (Higgins, 2008; Lee & Stvilia, 2017; Lord & Macdonald, 2003) for additional context. While digital data curation shares common infrastructure elements across different disciplines, the specific research tasks, types of data, and methods for managing, sharing, and assessing data may vary (Borgman et al., 2007; Chen & Chen, 2020; Stvilia et al., 2015). Moreover, DQA work requires access to suitable infrastructure for data quality evaluation, monitoring, and intervention. The industry has developed general data cleaning tools, such as Open Refine. Data curation consortia and communities of practice (e.g., DataOne, Data Curation Network (DCN)), and individual data repositories also develop their own data cleaning and normalization modules (DataOne Webinar Series, 2020a).

The FAIR (Findable, Accessible, Interoperable, and Reusable) conceptual framework has gained widespread popularity in the community of practice of data repository managers and curators. Its facets and criteria define minimum sets of metadata and repository system requirements that support data-related actions: finding, accessing, interoperating, and reusing (Dunning & De Smaele, 2017). There have been examinations of how fairly repositories meet those requirements (e.g., Dunning & De Smaele, 2017). There have also been attempts to further refine and extend the FAIR criteria for evaluating individual aspects of a data repository's quality, such as service quality (Koers et al., 2020), including developing very valuable specific metrics and their usage scenarios (Devaraju & Herterich, 2020). The current FAIR operationalizations focus on aspects of metadata, system, and service quality for a data repository's success (DeLone & McLean, 1992). One of the main aspects of the information system success model, data quality, however, is seen as “complementary...implicit to FAIR” (Koers et al., 2020, p. 9). While ensuring the quality of descriptive metadata and system services can help with data quality, they cannot replace it. Assessing and communicating information about a dataset’s quality (e.g., accuracy, completeness, reliability) can be very helpful for enabling the dataset’s reuse. For instance, simple metrics communicating whether the dataset has missing values or cases, whether it completely represents the study sample (e.g., some participants may decline to share their data with open access), or whether the associated study and the dataset have been peer reviewed, can be helpful to a researcher’s decision-making on whether to reuse the data or not. In order for a researcher to trust and reuse data, the quality of the data is the most important factor (Yoon & Lee, 2019).

A related concept to data quality is data value, which is shaped by its informativeness - the questions it can answer, the concepts and relations it illustrates, and how novel and sought-after these questions and concepts are (Stvilia and Gasser, 2008; Stvilia et al., 2015; Stvilia and Gibradze, 2022). The quantity or scale of data can be a predictor of its value. For example, large consumer data usually equates to higher value and increased market share for the company that possesses the data (Sun et al., 2018). Another value-related concept is cost. The cost of creating data often plays a role in assessing its value (Stvilia & Gasser, 2008). The higher value associated with big data also incurs a higher cost of curating it. Moreover, the value of preserving and curating an existing dataset can be measured in relation to the cost of recreating it from scratch when required (e.g., DNA sequencing data or simulated data; Zilinski et al., 2016). Data quality and value can help identify and prioritize DQA targets (Bowker, 2006; Stvilia et al., 2007; Stvilia et al., 2015; Stvilia, 2021). Although repository managers can't eliminate all quality issues in their data repositories, they can focus on addressing the data quality issues that are critical or of significant concern. The challenge is determining which data quality issues are important. There's limited research on how institutional data repositories assess and set their DQA priorities.

4. Design

This paper reports on a part of a larger, exploratory study. The datasets used by the study included approved applications of 122 data repositories for the CoreTrustSeal Board for the Core Trustworthy Data Repositories, interviews with 32 curators and repository managers, and data curation-related webpages of their repository websites (109 documents). The data was collected from April 2022 to February 2023. The combined dataset represented 146 unique RDRs. The combined dataset represented 146 unique RDRs. Domain-agnostic repositories accounted for nearly one-third of the total number, while domain-specific RDRs were dominated by linguistics, social sciences, and earth science disciplines (see Figure 1).

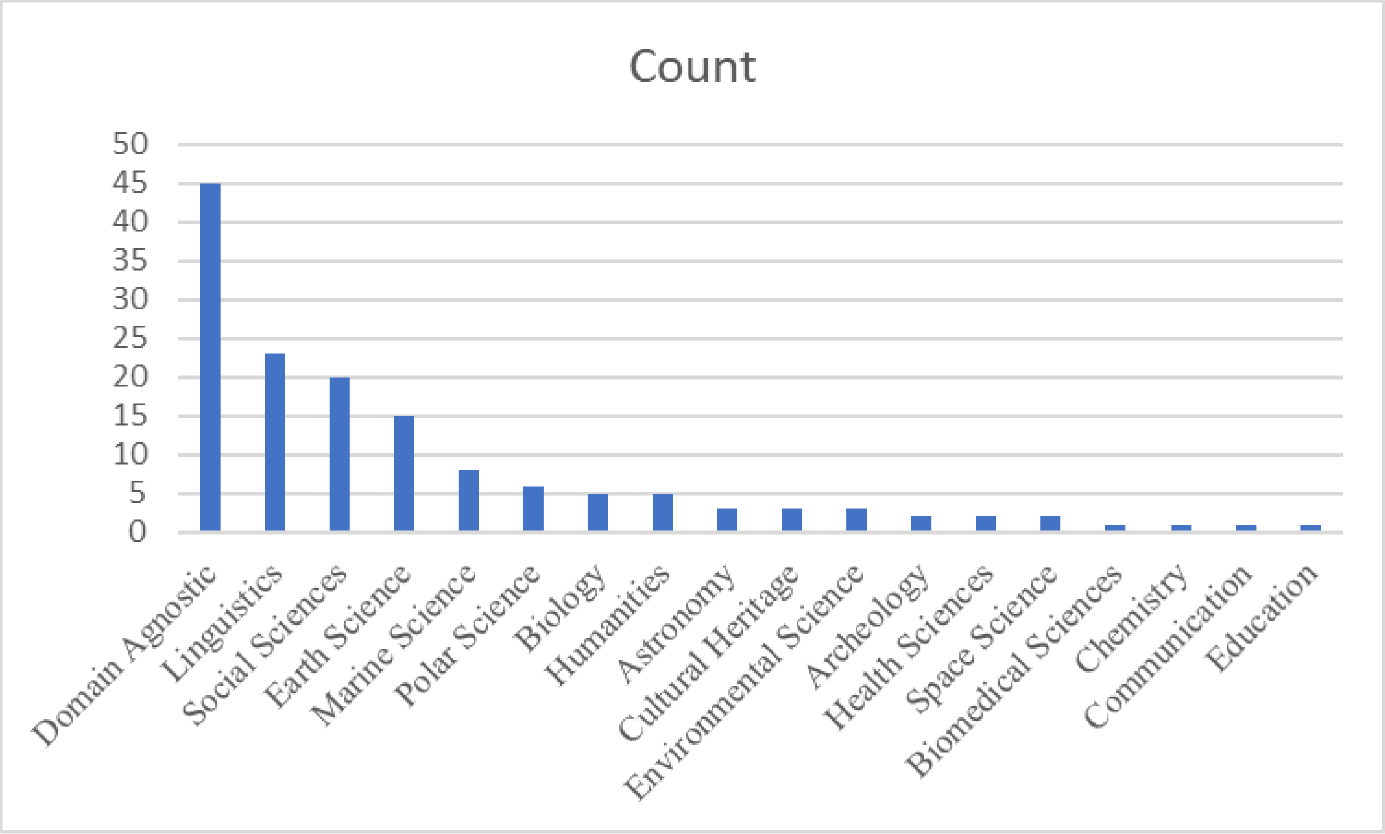

Figure 1. Disciplinary spread of RDRs represented by the sample.

The CoreTrustSeal applications dataset predominantly consisted of subject/domain-specific repositories (100), with an additional 22 being categorized as domain-agnostic repositories. Out of the 122 repositories surveyed, 30 were based in the United States, with the remainder distributed across various other nations.

In contrast, the interview data covered 32 repositories and 30 universities within the US. Among these, 29 universities were classified as R1 research institutions, with one as an R2, based on the Carnegie Classification. Some universities managed multiple repositories, while others utilized external cloud platforms like Dryad or Dataverse for their digital data collections. Even so, if a university provided substantial data curation support, its collections on external platforms were considered as an instance of an RDR. Of the 32 repositories, 27 were generalist or not specific to any domain, and 5 were specialized. Regarding the demographics of the interview participants, 59% were female and 41% were male. The majority of interviewees, 72%, held a Master's degree, while 28% had a Ph.D. The most common field of the highest degree among participants was library and information science (15 participants), though other disciplines such as psychology, political science, computer science, biology, English, history, anthropology, social work, ecology, journalism, and geography were also represented.

Thematic content analysis was employed to analyze the data, with repository used as the main unit of analysis. This study was guided by a theoretical framework comprising activity theory (Kaptelinin & Nardi, 2012) and an information quality evaluation framework (Stvilia et al., 2007). Activity theory offers conceptual models to understand DQA activity structures, focusing on goal-driven actions mediated by tools and the organization and its stakeholder communities. The information quality evaluation framework aided in assessing how data quality is perceived in these repositories, providing a classification of information quality dimensions and types of information quality issues and their effects on activities.

To carry out the analysis, we formulated a priori codes derived from the theoretical framework and the research questions. We then iteratively examined the dataset content, looking for both the predetermined and emerging codes. Through an inductive process, we mapped and merged the thematic codes from the datasets into 15 overarching categories, aligning with the research questions and the high-level concepts from the study's theoretical framework (Table 1; Bailey, 1994). Two authors coded each dataset, with each coder handling half of the dataset. Upon completing their coding work, they met to compare and discuss their coding results, particularly focusing on areas of disagreement. This discussion allowed them to resolve any differences and update the corresponding code assignments accordingly.

Table 1. Multiple data sources integration matrix. This paper reports only on the themes that are highlighted with a gray background.

Major theoretical concepts used to define 15 categories of themes	**Activity Theory**	**Information Quality Framework, and the Literature**	**Data Source**		**Examples of the associated code(s) and/or subcodes**
			Documents	**Interview**	
Activity	x	x	x	x	Activities; Conceptualization; Evaluation; Intervention; Communication
Subject	x	x	x	x	Subject
Object/Objective	x	x	x	x	Objective; Data Quality Definition
Motivations	x	x		x	Motivations
Data types		x	x	x	Data Types
Metadata types		x	x	x	Metadata Types
Actions	x	x	x	x	Actions; Define DQ; Evaluate DQ; Evaluate MQ; Coordinate DQA; Educate DQA; Assist DMP; Evaluate Provider; Modify Data; Remove Data, Document DQA actions, ...
Tools	x	x	x	x	Tools; DQ Measurement Tools; DQ Intervention Tools; DQ Communication Tools
Community	x	x	x	x	Domain Specific RDR; General RDR; Community of Practice
Norms and Rules	x	x	x	x	Norms; Rules; Reference Bases; DQA Models; Standards; Metadata Vocabularies; Data Formats; Ontologies; …
Division of Labor, Roles	x	x	x	x	Provider; Curator; User; Distributed DQA Workflow; Roles …

DQ Problem Types and Dimensions		x	x	x	DQ Problem Types; DQ Dimensions; Completeness; Consistency; Simplicity; Accuracy; Relevancy …
DQA Strategies; Prioritize		x	x	x	DQA Strategies; Prioritize
Challenges, Contradictions	x		x	x	Challenges; Contradictions; First Level Contradiction; Second Level Contradiction; Third Level Contradiction; Fourth Level Contradiction
Skills		x	x	x	Skills; Soft skills; Data management skills; Domain expertise; Technical skills; Research skills

5. Findings and discussion

Dataset quality assurance can be conceptualized as a system of activities that *evaluate*, *intervene*, and *communicate* data and metadata *quality* (Stvilia et al., 2007; Stvilia, 2021). Each of these activities can be conceptualized and analyzed using the general model of activity structure from activity theory as a theoretical lens. The model conceptualizes activity as a relation between *subject* and *object* that is mediated by *tools* and the *community* through its *rules* and norms and the community justified *division of labor* and *roles* (see Figure 2).

The content analysis identified instances of all DQA activities (i.e., evaluate, intervene, and communicate). DQA in repositories is collaborative work. The analysis identified multiple roles involved in DQA. The roles can be grouped into three categories identified for information and data quality ecosystems by previous studies (Stvilia et al., 2007; Stvilia, 2021): data creators/providers (e.g., field scientist, instrument mentor, site operator), DQA agents (e.g., data curator, librarian, approver, DQA analyst, dataset specialist, documentalist, scientific editor, domain expert), and data users (see Figure 3). The following subsections describe each activity and its structure.

5.1 Evaluation

Any DQA intervention should be preceded by data and metadata quality assessment. Quality assessment is necessary to identify quality problems and determine where to intervene and how. One cannot evaluate the quality of data, however, without having a clear understanding or *conceptualization* of what data quality means in a particular context.

To understand how RDR managers and data curators perceived data quality, the study asked interview participants to *define data quality* in their own words. The definitions were then split into sentence-based themes, and duplicate/similar statements were combined. The resulting statements were categorized into eight groups (as shown in Table 2). The categories demonstrated that the participants had a broader understanding of data quality, encompassing not only its traditional definition but also factors tied to information system quality. In particular, the categories covered essential components of an ethical and successful information system, such as the quality of the data creation process, adherence to ethical and legal standards, data provenance and authenticity, documentation quality, and system quality (Mason, 1986; DeLone, et al., 2003).

Table 2. The categories of participant-defined data quality characteristics.

1. Data Documentation Quality	**5. Data Quality**
Data quality is driven by the FAIR principles (Findable, Accessible, Interoperable, Reusable)	Data quality varies depending on the discipline and fitness for specific purposes
Variables and methodology are well-documented, and data limitations are noted	Data completeness to the best possible level
Data are well-documented via narrative and administrative metadata	Well-organized data with a clear organization strategy
Documentation allows secondary users to understand the data	Data checked for common data errors and is error-free
Clear and well-defined organization strategy, and metadata completeness	Trustworthiness of data
Metadata is complete and provides enough detail for data use	Trust in the reliability and accuracy of the data
Metadata supports data understanding without needing to contact the original owner	**7. Data Provenance and Authenticity**
2. Ethical and Legal Compliance	Clear and well-defined provenance of data
Data access complies with ethical requirements (e.g., informed consent)	Listing of original sources and code used to create the data
Data access complies with legal requirements (e.g., copyright)	Inclusion of license information for data reuse
3. Data Collection/Creation Process Quality	Data remains close to its original form and is not compressed or corrupted
High-quality data collected through a sound methodology	**8. System Quality**
Data collected through reliable and valid processes	Adequate procedures for data access, transport, and usage are in place
4. Data Compatibility and Accessibility	
Data is available in compatible formats whenever possible	
Data is accessible and available for use	
Practical measures of usability and accessibility are considered	
Data accessibility is ensured through appropriate software, understanding of data structure, and descriptive metadata	

The literature asserts that data quality is multidimensional (Stvilia et al., 2007). *Data quality dimensions* are data virtues or characteristics used to define and/or communicate the concept of quality. Indeed, our analysis of the documentary data showed that, in the aggregate, repositories referenced 11 dimensions when reporting on their data quality assurance practices: *Accuracy, Completeness, Simplicity, Consistency, Currency, Precision, Lack of Redundancy, Relevancy, Reliability, Reputation,* and *Soundness* (see Figure 3). As a group, the subject repositories referenced all 11 dimensions, while the institutional repositories group referenced only four dimensions: *Accuracy, Completeness, Simplicity,* and *Relevancy*. The most frequently referenced dimensions were *Completeness, Relevancy, Accuracy, Simplicity,* and *Consistency*. Repositories checked datasets for completeness of data values and metadata elements. They also examined whether datasets were relevant to their curation scopes or measured phenomena. The third most frequently referenced dimension was *Accuracy*. Repositories checked whether datasets contained plausible values for specific variables. Repositories assessed datasets for their simplicity or complexity, as well as the availability of adequate and relevant metadata and documentation to make them easy to understand and accessible. Additionally, they verified the accuracy of the data values and ensured that they were consistent with the scales and metadata used by datasets.

In addition to examining the documentary data, the study also asked interview participants to select data quality dimensions they used when evaluating the quality of datasets in their repositories. For this question, the study utilized the set of quality dimensions identified by Stvilia et al. in 2015. The findings of the interview data analysis agreed with the results of the documentary analysis. *Completeness, Consistency, Simplicity, Accuracy,* and *Relevance* were still the most frequently referenced dimensions, though in a slightly different order (see Figure 2).

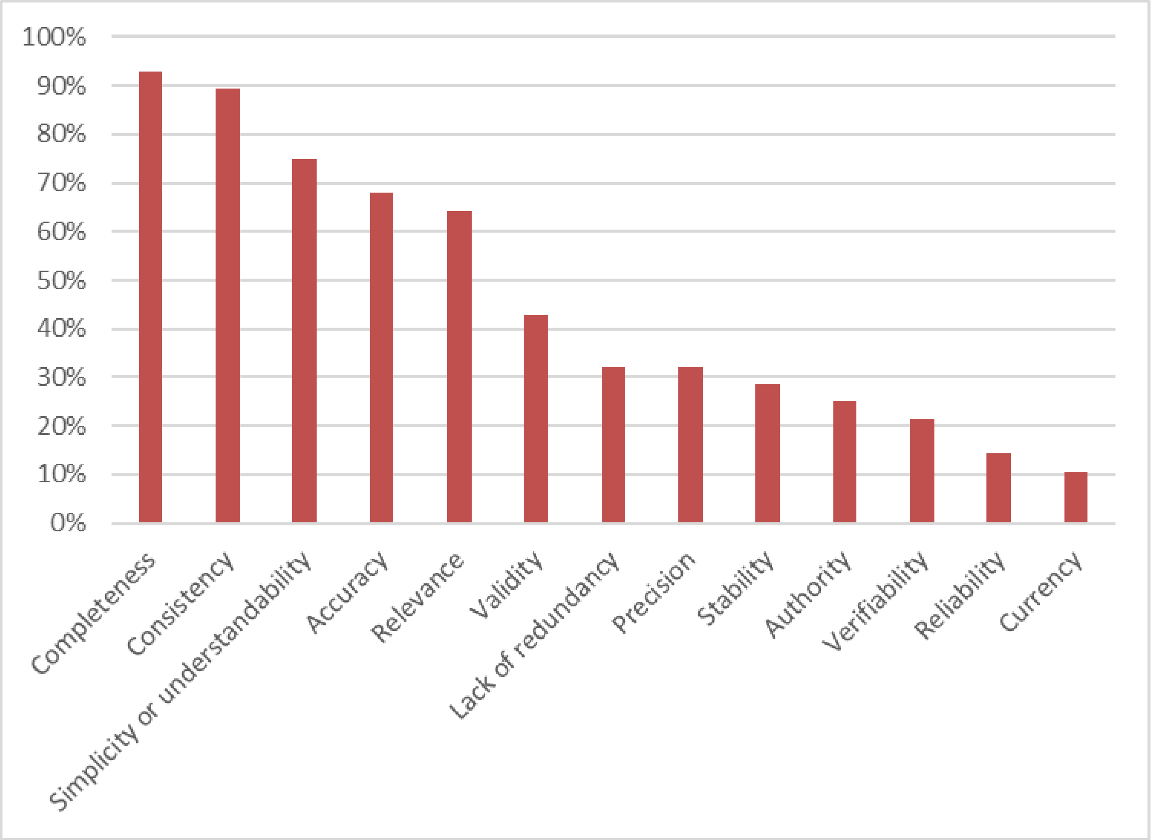

Figure 2: Distribution of interview participants' use of data quality dimensions reported as a percentage of the total (based on the responses of 28 participants who completed this question).

Apart from the dimensions of data quality, repositories' certification proposals also referenced assessments of data authenticity, integrity, and privacy. This finding was further supported by the interview data and participants' descriptions of data quality, as shown in Table 2. The first two can be categorized as data security dimensions, as they referred to efforts made to protect data from unauthorized changes and modifications. The latter is a privacy dimension. Even though data security, privacy, and quality are interrelated concepts that are essential in information assurance workflows, it is crucial to distinguish between them (Mason, 1986). This is because there are trade-offs that need to be considered when implementing actions to ensure data privacy, quality, and security. For instance, to ensure data privacy, providers or curators may need to anonymize a dataset by removing identifiers and/or injecting noise or irrelevant data. These, however, may degrade the dataset's quality (i.e., accuracy, make it statistically invalid) and render it unusable for research or policymaking purposes (Stvilia and Gibradze, 2022).

Data and information quality problems and evaluations can be grouped into three categories: *intrinsic, relational,* and *reputational* (Stvilia et al., 2007). Intrinsic DQ can be assessed by measuring internal/intrinsic characteristics of data in relation to some general reference standards or sources of a particular culture (e.g., language dictionary and grammar specifications). Representational DQ is assessed relative to its context of use using context-specific standards and benchmarks (e.g., activity or community-specific data model or profile; business rules). Reputational DQ is measured based on the reputational position of data in a cultural or community structure. Reputational DQ is an indirect evaluation of quality grounded in the credibility of the origin of data and its mediation (i.e., expertise and trustworthiness, Choi and Stvilia, 2015). The analysis found references to all three types of DQ problems and *evaluations*. RDRs checked datasets for missing or invalid values. They used community data quality models, metadata profiles, and ontologies to evaluate relational data quality. Finally, they used providers' reputation to predict the quality of their datasets.

Metadata is defined as structural data that enables specific functionality(s) on datasets. To use/reuse a dataset in an activity, it needs to be documented with high quality metadata (Consultative Committee for Space Data Systems, 2012). This applies to DQA activities as well. To interpret data and evaluate its intrinsic, relational, and reputational quality, data needs to be accompanied by appropriate metadata describing its context of creation, manipulation, and use (Stvilia et al., 2007).

> Information on the actual quality of the source data is seldom provided, so this generally has to be inferred from the metadata provided (d106).

Interview participants, too, spoke of the importance of evaluating datasets for the completeness of documentation to enable users to understand and effectively use the research data. Adopting a user-centered evaluation was also highlighted to ensure that the data was user-friendly and met the needs of potential users. Furthermore, when speaking about metadata quality, participants often referenced the FAIR framework to make datasets discoverable, understandable, and reusable.

Data quality is contextual. Different research disciplines and areas may use different research methodologies as well as models and reference sources for evaluating the quality of data. Hence, data repositories that curate data from different disciplines may need to *use multiple subject specific DQA models and references* (e.g., data quality standards). Likewise, *DQA actions* and *tools* (e.g., metadata templates) can be data type specific. Qualitative data may require different quality control actions than time series data. There can be derived datasets that are produced through processing and aggregation of raw datasets collected from sensors. The derived datasets may require different DQA processes than the original data, including different DQ evaluation models.

> Given the relatively wide range of data published through the ORNL DAAC, there is no single approach or applicable standard for representing data quality (d1).

The analysis revealed that data quality evaluation in RDRs was often an iterative and collaborative process, where curators and/or data repository teams engaged in back-and-forth interactions with depositors to evaluate the submitted data and metadata to meet the repository's quality standards.

> This iterative and collaborative approach is used to evaluate the quality of data and metadata and to assess the data and metadata's adherence to the relevant schema employed by the RDA (d100).

In addition, most of the RDRs enabled end-users to provide data quality feedback or suggest improvements. The analysis identified *division of labor* and multiple *roles* contributing to data quality evaluation in research data curation ecosystems. Data quality assurance agents comprised RDR managers, data curators, and data librarians who reviewed submitted datasets and associated metadata and made quality intervention recommendations. At some RDRs, data curators also collaborated with subject librarians and subject experts to evaluate datasets that require domain-specific expertise.

Furthermore, some domain-specific RDRs used a peer-review process and specialized roles focused on quality assurance, such as scientific editors, for data quality evaluation. The latter were responsible for the final check of data and metadata, ensuring that community standards were respected. In addition, data and scholarly communication librarians were often involved in helping people find and use data, which might require evaluating the quality of data sources and providing guidance on data usage. Thus, a DQA workflow system must support collaboration among the different DQA roles.

The complexity of the division of labor in DQA is determined by the nature and complexity of the data curation work in a repository. Resource-rich repositories can treat datasets as products and use a complex data curation workflow. It may comprise a clearly defined division of labor with peer-review mechanisms and strong user services components. Hence, its DQA workflow may involve the evaluation of all its components and tools, including the user guides and reference sources.

> SAuS is used to ensure that the User Guide is reviewed by at least one ORNL DAAC staff member not involved in developing the User Guide (d1).

Repositories that serve as data aggregators of an established network of data providers may use distributed DQ governance that is aligned with the distributed nature of their data curation workflow. Furthermore, large research networks have more DQA resources and expertise they can tap into for evaluating data quality. They may establish a formal body, such as an advisory committee, to coordinate the network's data quality assurance work.

> The Quality Assurance Advisory Committee, QAAC exists to improve the overall quality of seismic and other data collected and managed by IRIS by promoting these principles; coordinating the development and use of tools to measure data quality, and enhancing quality control feedback to network operators, thus encouraging high-quality, seismic datasets for the broad community (d 34).

When asked about *motivations*, interview participants disclosed that they evaluated research data quality motivated by the desire *to add value, promote its reusability; maintain the reputation and credibility, and ensure the efficiency of their RDRs; support research integrity, and meet the mandates and expectations of funders and the government* (see Figure 3).

5.2 Intervention

A data quality evaluation activity is usually followed by an intervention activity to tackle the data and metadata quality issues identified during the evaluation (Stvilia et al., 2007). Participants revealed that, in most cases, researchers were responsible for fixing any identified data quality issues. The data quality intervention approach typically involved contacting the researcher and requesting them to make the necessary changes. RDR managers and data curators usually focused on augmenting metadata, including descriptive records, documentation, and labels surrounding the data, migrating datasets into more accessible or open file formats. They also offered guidance, suggestions, and templates to help researchers complete missing information or enhance the metadata. Furthermore, interview participants also disclosed assisting researchers with making changes to datasets and preparing them for publications, particularly when the tasks were time-consuming or required specialized data management expertise.

Product quality can be enhanced by improving the quality of the *process* that generates the product and/or *reworking or scrapping* the product (see Figure 3; Stvilia et al., 2007). The analysis showed that actions like a rejection of a submitted dataset or removal of a curated dataset due to quality problems can be grouped under the scrap category. The rework category may include intervention actions such as standardizing the data schemes and values of a dataset, adding missing data or metadata, and adding associated objects such as software and user guides to make the data more understandable and reusable.

Repositories may have less control over a data production process. Still, one way to improve the quality of the data creation process is to educate data producers/providers. Our analysis showed that data curators and librarians conducted outreach and workshops to educate individual researchers, and project teams, especially graduate students and postdocs, about the principles and best practices in data management and sharing. They taught researchers how to identify potential quality issues or errors and improve the quality of their data and associated metadata, how to share their data in an accessible and preservable format, and how to enhance the usability and downstream impact of the data. They taught researchers about the significance of metadata and including appropriate naming conventions; how to meet metadata requirements, and adhere to standards for data citation, discovery, and identification. Data curators and librarians also educated depositors on the acceptable terms of reuse and how to obtain appropriate informed consent for data sharing, how to protect the confidentiality of study subject information, and how to include licensing information to establish the conditions under which the data can be used. Finally, data librarians and curators used data curation and DQA questionnaires and guidelines as educational resources to help depositors understand the requirements and expectations for data and metadata quality.

In addition to educating researchers about data management best practices and standards, some repositories affected the quality of dataset production by consulting research teams on data management planning. They helped researchers to write and/or evaluate data management plans required by funding agencies.

> Researchers interested in including their resources in the IDS repository are invited to develop a data management plan. In coordination with the staff of the CLARIN centre, which is offered as a free service already in the early stages of their projects (d85).

Furthermore, repositories can affect the quality of a dataset production process indirectly by controlling the quality or credibility of their providers. Some RDRs selected and/or accepted datasets based on the provider's reputation.

> Actual data currently is only accepted when data depositors can be trusted (d85).

Data librarians and curators do have more control over the curation of a dataset after the dataset is ingested in a repository. Both data and metadata quality change over time. They need continuous maintenance to keep data reusable. These might include updating and converting datasets' schemes, content, and metadata to align and comply with the community's standards, rules, and laws (e.g., GDPR; see Figure 2). Furthermore, a quality bug/error could be discovered after a dataset is ingested in a repository. If the PI(s) or the original data provider(s) were not available to correct or resolve the data quality problem, the curator might need to update the data or its metadata.

The most drastic DQA action is the removal of a dataset from the repository. That can be caused by the dataset not meeting the repository's quality requirements, the provider declining to make necessary corrections, or the dataset violating existing law.

> The Repository has a history of making corrections and improvements to data and metadata based on such feedback including taking down content that was felt to impinge on the rights of 3rd parties (d3).

As with any other change, data quality interventions need to be properly documented to support provenance-based activities. These can be done by updating associated DQA metadata, including data quality reports.

> The ARM infrastructure conducts an extensive data reprocessing program that is informed by the data quality assessment process. Reprocessing is performed to fix known data issues and has been used extensively throughout the lifetime of ARM. Reprocessing requires the modification or elimination of previous DQRs [i.e., data quality reports] and the subsequent reissuing of data to all who may have downloaded the data from the data archive (d56).

To make a dataset reusable, curators need to ensure the quality of its metadata. They may request and/or add metadata to a dataset to enhance its quality at both the collection and content entity levels. For instance, molecular biology communities and their repositories curate knowledge about specific molecules such as proteins. That knowledge is represented as sequence data and associated annotations. Curators need to ensure that the data and metadata are consistent with each other and accurately represent the current knowledge about a molecule species.

> Curators assimilate all the information from various sources, reconcile any conflicting results and compile the data into a concise but comprehensive report, which provides a complete overview of the information available about a particular protein (d54).

5.3 Communication

Communication is an essential part of a DQA process. There is an iterative feedback loop between the data evaluation and intervention activities (see Figure 3). DQ information can also be generated and communicated throughout a dataset's lifecycle, including its creation, sharing, and use activities. Data quality levels need to be communicated to users so that they can make informed decisions on whether they can use the data for a particular task. The analysis found that data quality communication could involve multiple roles from data curation ecosystems in RDRs. Depositors might document data quality-related problems in user guides and readme files when they submitted their datasets to repositories. Curators, on the other hand, communicated information on the quality problems they identified and the quality assurance actions taken through data quality reports and metadata. They communicated with data providers, metadata specialists, and IT departments, exchanging information and addressing data and metadata quality problems in datasets and DQA workflows.

Data curators recognized that researchers possessed the most accurate information about their own data and relied on them to provide missing details, validate changes, and ensure the accuracy and integrity of the data.

> If a data manager notices any other quality issues, inconsistencies, or errors in the data, the depositor is contacted for further information regarding metadata and data or for a new version of the dataset. If the issue cannot be resolved, this is also documented publicly in the descriptive metadata (d38).

In addition, data curators, data librarians, and scholarly communication librarians communicated with providers and users to facilitate data use and enhance their DQA literacy. They also set key DQA standards, best practices, and guidelines, and communicated them to data providers, users, and other stakeholders. Repository managers and digital publishing and copyright librarians served as a public face for data services, answered questions, and communicated key standards related to data licensing, sharing permission, and other aspects of ethical sharing and uses of data. Furthermore, data librarians might provide general data management support to the campus, which involved communication with researchers and other stakeholders about DQA activities.

Most of the subject RDRs provided users with means for communicating data quality feedback to the curator(s) and/or the provider(s). Scientific communities and funding agencies increasingly focus on the reproducibility and replicability of research findings. Making research transparent, including providing open access to research data and collecting and sharing end-user feedback on the quality of data are essential for ensuring the quality and rigor of research findings. Data openness and transparency enable higher use, and more quality evaluation and intervention, translating into higher quality data (Orr, 1998).

> This policy of maximal openness allows for any party to assess the scientific and scholarly quality of data as much as possible, which is common practice in the area of language resources (d64).

Interview participants expressed a desire for standardized data quality communication statements. Most of the repositories enabled users to communicate data quality feedback directly using tools such as a system provided contact form. A few of them utilized a ticket system or a quality rating schema. Some RDRs also communicated user feedback about datasets indirectly by collecting and sharing datasets' altmetrics.

> Data users can raise issues with archived data using the PANGAEA contact which is connected with the ticket system. In addition, data sets can be rated via social networks including altmetrics (d84).

In addition, if data is distributed with a little delay after its generation, it increases the chance of users downloading data that contains errors and inconsistencies. It is essential that the quality of data is evaluated promptly and quality evaluation results are proactively pushed/communicated to the users who already accessed the data. That way the users can learn about quality problems and download reprocessed, error-free copies of the data, and/or use the data while informed about its quality.

> All the generated data quality information is provided to users in different phases of data discovery, downloading, and use. Users who ordered the data in the past will be notified if data quality information is available for the data they received in the past (t56)

In a research network, data quality feedback and evaluations can also be communicated among network members to help harmonize the network's DQA standards and their actual implementations.

> Feedback from these quality checks is regularly communicated back to the Meertens Institute and allows us to

> align our internal quality standards with the ones formulated by the CLARIN community (d117).

Some participants pointed out that although they did not directly communicate about the quality of datasets curated by their RDR with end-users, the RDR attaining an external certification, such as the CoreTrustSeal certification, would indicate the RDR's adherence to recommended practices and standards of data quality assurance. Thus, it could signal to end-users that their RDR's datasets met a certain level of quality.

5.4 Policies, tools, and reference sources

Repositories used a variety of policies to guide data curation and ensure the quality, accessibility, and ethical compliance of the deposited data.

As with any activity, repositories' data and metadata quality evaluation activities are shaped and mediated both by their local standards and norms, as well as by the norms, standards, and tools of stakeholder communities (see Figure 3).

> In the cases of contributions from the International Tree-Ring DataBank and the International Multi-Proxy Paleo-Fire Database, these communities have given explicit direction for additional quality control checks to be done. These checks are performed using tools and quality metrics that these communities have provided (d61).

Repositories referenced inhouse and community software they used to visualize and identify quality problems in datasets. The most frequently referenced type of software, however, was one that not only identified quality problems but also measured quality by applying a set of metrics. Some repositories also shared data quality analysis and visualization tools with users so that users could analyze the quality of datasets and determine the datasets' suitability to their tasks.

The DQA communication activity, on the other hand, may use a metadata vocabulary, a quality rating scheme with labels or badges that explicate the repository's DQA model and the quality of individual datasets to end-users.

> In order to help the Designated Community evaluate the quality of datasets, we provide, where relevant DQA badges indicating whether expert reviewers have reviewed the dataset, is associated with a peer reviewed published article, (forthcoming) has been reviewed and approved to appear on an institutional partner showcase, has been reviewed and curated (and updated according to feedback by the Data Producer) by a relevant subject matter (d123).

Another frequently referenced tool that repositories used to communicate DQA information to end users was user guides. In addition, to describing the semantic structure or the schema of a dataset, the user guide could also convey information on any quality assessment or intervention procedures applied to the dataset, and/or any known quality problems or limitations. User guides supplemented with data lineage tools can be used to generate and communicate data provenance information.

When describing DQA intervention activities, repositories referenced software they used to clean data and/or add value to it through automated annotations and tagging. Repositories also used software to convert data into more accessible and efficient file formats.

Institutional RDRs are often integrated with universities' research information and data management infrastructure. They may use that infrastructure for storing, analyzing, and providing access to data. They may also use specific systems or services, such as research information management systems, to enhance the quality of datasets by documenting and linking them to related research information.

> Where a journal publication occurs after approval, and if the depositor fails to send us the details, curators often make use of the University's CRIS and the associated web portal, to find and link journal publications to our datasets (d10).

When discussing their DQA activities, repositories mentioned different kinds of reference sources they used to evaluate and/or enhance their data and metadata quality. Most of those reference sources were descriptive metadata schemas and vocabularies. They ranged from general descriptive metadata schemas such as MARC and MODS to domain specific metadata schemas, profiles, and standards such as DDI, CESSDA Metadata Model or ISO 19115. Likewise, repositories used general and subject specific controlled vocabularies such as DDC and Space Physics Archive Search and Extract Ontology (SPASE) to standardize data values or annotate datasets. Finally, repositories used best practices guides and recommended data exchange formats such as JSON to align their DQA and data curation practices with the communities' norms and expectations.

5.5 Optimizing – Cost and Value

Participants disclosed that they used the importance or value of a dataset to their stakeholder communities to *prioritize* their DQA activities (see Figure 3). The value of a dataset can be shaped by different factors such as its scale/size, the number of variables included, and the level of demand measured by the frequency of its use (Stvilia and Gibradze, 2022). For instance, repositories may curate datasets that are small and study specific. Their reuse scope can be narrow and limited to the evaluation or the replication of the original study. Large scale, representative datasets of specific populations, government, or industry activities, on the other hand, can be reused in new studies as standalone or aggregated with related data. Hence, their value and importance could be higher.

> Datasets of particular importance and important study collections falling into the Data Archive's core areas of collection are processed (cumulated, harmonized, standardized), documented, and enhanced in much greater depth - not only on study level, but on the level of individual questions and variables (d121)

Another factor that shapes DQA priorities is the current quality of datasets. A participant revealed that they prioritized popular datasets with quality problems for DQA interventions which is in agreement with the information quality evaluation theory (Stvilia et al., 2007).

> For example, if something comes to us and it's very well documented. There's not as much we would add to it, and maybe we think it's not gonna be that popular because of the topic area. In that case, the project may choose level one, which is like the base level of curation. So, at that level, we're not, we're gonna do some set of curation tests, but not that many. Versus maybe if we get a very complex study and we know it's gonna be really popular, but it's not in the best shape (p 23).

Other factors that may affect DQA priorities are the provider's priorities, such as specific deadlines of events that use a particular dataset (e.g., a workshop), as well as the level of the provider's engagement in the curation of the dataset and the amount of resources the provider or their funders are willing to invest in the dataset's quality assurance. In general, participants mentioned that they tried accommodating researchers' deadlines and needs, and the amount of time spent on DQA might depend on their familiarity with the file format and research area.

> I would say one factor affecting how much time I would spend on data quality issues is my familiarity and experience with a given file format. Um, or maybe even that experience with the research area (p 26).

In addition to repositories being guided by their data collection policies to determine the priorities and curation levels for datasets, research communities can have specific policies defining what data should be curated and at what level. Also, providers and users may have different needs and uses of a dataset. Hence, they may also have different priorities for different characteristics of the dataset's quality. Some might study data quality problems. Hence, they might need datasets with quality problems.

> Moreover, different researchers have different goals in the creation of transcriptions. Everyone wants to get the words right, but some care about breath groups and intonation. Others want to code errors. CHAT allows for all of these possibilities (d18).

A DQA activity can also be shaped by its cost. The manual DQA of big time-series streaming data is often not feasible. Instead, automated statistical quality control checks can be applied that are monitored and evaluated by human quality inspectors. Furthermore, because of the scale and dynamic nature of streaming data, it is essential that data quality problems for this type of data are detected and corrected early to limit the spread and impact of inaccurate data.

> The first component is a "rapid evaluation and response" piece involving data inspection and assessment. It is designed to identify gross and some more subtle issues within the data streams as fast as possible and relay that information to site operators and the instrument mentors so that the (potential) problem-resolution process can begin. The goal of this component is to minimize the amount of data that is affected by the problem (t56)

Certain types of data, such as numerical or coded data, are better suited for automated quality checks, such as verifying values and ensuring consistency, compared to other types of data. Furthermore, data and metadata quality evaluation on some dimensions can be more amenable to automation than on others. For instance, technical quality assessment or the assessment of data authenticity and integrity could be automated (e.g., using checksum algorithms). Assessing the completeness of data, however, might require manual evaluation by domain experts and, hence be more costly.

Activities are driven by specific needs and related motivations. Hence, the completion of an activity depends on the strength of the associated motivations (Nardi, 2005; Stvilia et al., 2019). In addition to conducting data management training workshops to enhance researchers' data management skills, some repositories provided additional extrinsic motivations for researchers to engage in DQA and offset its cost. In particular, they provided small grants for documenting datasets and making them sharable. The repositories also collaborated with publishers to help researchers get credit for shared datasets by publishing related data papers in peer-reviewed journals.

Thus, repositories considered a dataset's *value*, its *current state of quality*, their available *subject expertise* and knowledge of the dataset's *format*, the *cost* of quality assurance, and the available *funding* from the depositor or funding agency when prioritizing datasets for quality assurance (see Figure 3). By considering these aspects, repositories tried to effectively allocate resources and prioritize their data curation activities to meet the needs of researchers and users. RDRs aimed to strike a balance between addressing data quality issues and meeting researchers' needs in a timely manner.

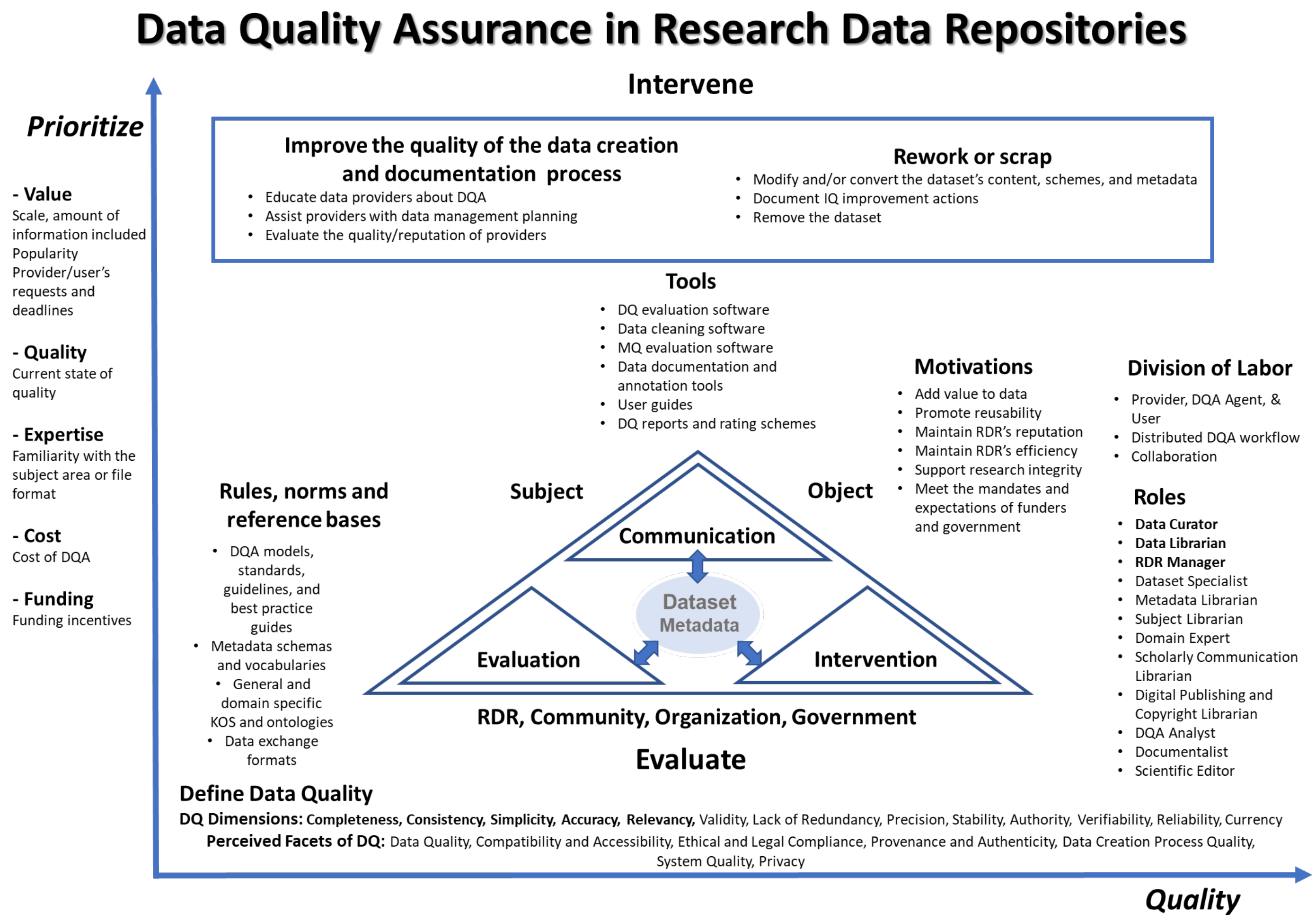
Data Quality Assurance in Research Data Repositories
Intervene
Prioritize
- Value
Scale, amount of information included
Popularity
Provider/user's requests and deadlines
- Quality
Current state of quality
- Expertise
Familiarity with the subject area or file format
- Cost
Cost of DQA
- Funding
Funding incentives
Improve the quality of the data creation and documentation process
• Educate data providers about DQA
• Assist providers with data management planning
• Evaluate the quality/reputation of providers
Rework or scrap
• Modify and/or convert the dataset's content, schemes, and metadata
• Document IQ improvement actions
• Remove the dataset
Tools
• DQ evaluation software
• Data cleaning software
• MQ evaluation software
• Data documentation and annotation tools
• User guides
• DQ reports and rating schemes
Motivations
• Add value to data
• Promote reusability
• Maintain RDR's reputation
• Maintain RDR's efficiency
• Support research integrity
• Meet the mandates and expectations of funders and government
Division of Labor
• Provider, DQA Agent, & User
• Distributed DQA workflow
• Collaboration
Roles
• Data Curator
• Data Librarian
• RDR Manager
• Dataset Specialist
• Metadata Librarian
• Subject Librarian
• Domain Expert
• Scholarly Communication Librarian
• Digital Publishing and Copyright Librarian
• DQA Analyst
• Documentalist
• Scientific Editor
Rules, norms and reference bases
• DQA models, standards, guidelines, and best practice guides
• Metadata schemas and vocabularies
• General and domain specific KOS and ontologies
• Data exchange formats
Subject
Object
Communication
Dataset Metadata
Evaluation
Intervention
RDR, Community, Organization, Government
Evaluate
Define Data Quality
DQ Dimensions: Completeness, Consistency, Simplicity, Accuracy, Relevancy, Validity, Lack of Redundancy, Precision, Stability, Authority, Verifiability, Reliability, Currency
Perceived Facets of DQ: Data Quality, Compatibility and Accessibility, Ethical and Legal Compliance, Provenance and Authenticity, Data Creation Process Quality, System Quality, Privacy
Quality

Figure 3. A DQA model for research data repositories. *Notes: The most frequently referenced DQ dimensions are in bold.*

6. Conclusion

This study provided a theory-based examination of the DQA practices of research data repositories summarized as a conceptual model. We identified three DQA activities: evaluation, intervention, and communication and their structures, including activity motivations, roles played, and mediating tools and rules and standards. When defining data quality, study participants went beyond the traditional definition of data quality and referenced seven facets of ethical and effective information systems in addition to data quality. Furthermore, the participants and RDRs referenced thirteen dimensions in their DQA models. The most commonly mentioned dimensions were Completeness, Consistency, Simplicity, Accuracy, and Relevancy. These dimensions and facets help us better understand how RDRs define data quality. The study found that evaluating DQ was an iterative, contextual, collaborative process that, in some cases, involved distributed teams. In their data quality interventions, RDRs primarily focused on reworking datasets and their metadata. However, they also took steps to enhance the process of data creation and documentation. These included educating depositors about DQA, helping them with data management planning, and evaluating their trustworthiness and reputation. In addition, the study revealed that DQA activities were prioritized by data value, level of quality, available expertise, cost, and funding incentives.

The study's findings can inform data curators' and repository managers' DQA practices, including the design of their services, data stewardship, and policies. The study's findings can also inform the design and construction of digital research data curation infrastructure components on university campuses that aim to provide access not just to big data but trustworthy data - data that could be used with confidence in research, teaching, policymaking, and developing services for consumers and society in general. For instance, as we argued above, communities of practice focused on repositories and archives could consider adding FAIR operationalizations, extensions, and metrics focused on data quality. The availability of such metrics and associated measurements can help reusers determine whether they can trust and reuse a particular dataset. The findings of this study can help to develop such data quality assessment metrics and intervention strategies in a sound and systematic way. The study's outcomes can also inform the data curation and data science training and education curricula of LIS schools and communities of practice.

A future related study will examine how end-users perceive and evaluate the quality and credibility of data provided by RDRs.

7. Acknowledgment

We would like to express our gratitude to the participants of our study. We extend our appreciation to Leila Gibradze for her invaluable assistance with the analysis of data.

This research is supported by a National Leadership Grant from the Institute of Museum and Library Services (IMLS) of the U.S. Government (grant number 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. Bailey, K.D., 1994. Typologies and taxonomies: An introduction to classification techniques, Vol. 102. Sage.
2. Ball, A. 2012. Review of data management lifecycle models. University of Bath, IDMRC.
3. Barrett, C. 2019. 'Are the EU GDPR and the California CCPA becoming the de facto global standards for data privacy and protection?', Scitech Lawyer, 15(3), pp. 24-29.
4. Borgman, C.L., Wallis, J.C. & Enyedy, N. 2007. 'Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries', International Journal on Digital Libraries, 7, pp. 17-30.
5. Bowker, G.C. 2005. Memory practices in the sciences. Vol. 205. Cambridge, MA: Mit Press.
6. Boyd, D. & Crawford, K. 2012. 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon', Information, communication & society, 15(5), pp. 662-679.
7. Burton, A. & Treloar, A. 2009. 'Designing for discovery and re-use: The 'ANDS Data Sharing Verbs' approach to service decomposition', International Journal of Digital Curation, 4(3), pp. 44-56. Available at: http://ijdc.net/index.php/ijdc/article/view/133/.
8. Chen, S. & Chen, B. 2020. 'Practices, challenges, and prospects of Big Data curation: A case study in geoscience', International Journal of Data Curation, 14, pp. 275-291.
9. Choi, W. & Stvilia, B. 2015. 'Web credibility assessment: conceptualization, operationalization, variability, and models', Journal of the Association for Information Science and Technology, 66(12), pp. 2399-2414.
10. Consultative Committee for Space Data Systems. 2012. Reference model for an open archival information system (OAIS). CCSDS 650.0-M-2. Consultative Committee for Space Data Systems. Available at: https://public.ccsds.org/pubs/650x0m2.pdf.
11. DeLone, W.H. & McLean, E.R. 2003. 'The DeLone and McLean model of information systems success: A ten-year update', Journal of Management Information Systems, 19(4), pp. 9-30.
12. Devaraju, A. and Herterich, P. 2020. D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories. Zenodo. doi: 10.5281/zenodo.3678716.
13. Dunning, A., De Smaele, M. and Böhmer, J. 2017. Are the FAIR data principles fair?. International Journal of digital curation, 12(2), pp.177-195.
14. Gutmann, M., Schürer, K., Donakowski, D. & Beedham, H. 2004. 'The selection, appraisal, and retention of social science data', Data Science Journal, 3(0), pp. 209-221.
15. Higgins, S. 2008. 'The DCC curation lifecycle model', International Journal of Digital Curation, 3(1), pp. 134-140. Available at: http://ijdc.net/index.php/ijdc/article/view/69.

16. Huang, H., Stvilia, B., Jörgensen, C. & Bass, H., 2012. 'Prioritization of data quality dimensions and skills requirements in genome annotation work', Journal of the American Society for Information Science and Technology, 63, pp. 195-207. doi:10.1002/asi.21652.

17. Huberman, M. and Miles, M.B., 2002. The qualitative researcher's companion. Sage.

18. Juran, J., 1992. Juran on quality by design. New York: The Free Press.

19. Kaptelinin, V. and Nardi, B., 2012. 'Activity theory in HCI: Fundamentals and reflections'. Synthesis Lectures on Human Centered Informatics, 5(1), pp.1–105.

20. Koers, H. Gruenpeter, M., Herterich, P., Hooft, R., Jones, S., Parland-von Essen, J., and Staiger, C., 2020. Assessment report on 'FAIRness of services'. Zenodo. doi: 10.5281/zenodo.3688762.

21. Lee, D.J. and Stvilia, B., 2017. 'Practices of research data curation in institutional repositories: A qualitative view from repository staff'. PLoS ONE, 12(3), e0173987.

22. Lord, P. and Macdonald, A., 2003. E-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision. Bristol: The JISC Committee for the Support of Research. Available at: https://digitalpreservation.gov/news/2004/e-ScienceReportFinal.pdf.

23. Lyon, L., 2012. 'The informatics transform: Re-engineering libraries for the data decade'. International Journal of Digital Curation, 7(1), pp.126–138.

24. Mason, R.O., 1986. 'Four ethical issues of the information age'. MIS quarterly, 10(1), pp.5-12.

25. Nardi, B.A., 2005. 'Objects of desire: Power and passion in collaborative activity'. Mind, Culture, and Activity, 12(1), pp.37–51.

26. National Academies of Sciences, Engineering, and Medicine (NASEM), 2019. Reproducibility and replicability in science. National Academies Press.

27. National Academies of Sciences, Engineering, and Medicine (NASEM), 2020. Advancing Open Science Practices: Stakeholder Perspectives on Incentives and Disincentives: Proceedings of a Workshop–in Brief. Available at: https://nap.nationalacademies.org/catalog/25725/advancing-open-science-practices-stakeholder-perspectives-on-incentives-and-disincentives.

28. National Science and Technology Council (NSTC), 2022. 'Desirable characteristics of data repositories for federally funded research'. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

29. Nelson, A., 2022. 'OSTP Memo: Ensuring free, immediate, and equitable access to federally funded research'. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

30. Ng, A., 2021. 'AI doesn’t have to be too complicated or expensive for your business'. Harvard Business Review. Available at: https://hbr.org/2021/07/ai-doesnt-have-to-be-too-complicated-or-expensive-for-your-business

31. Orr, K., 1998. 'Data quality and systems theory'. Communications of the ACM, 41(2), pp.66–71.

32. Ryan, R.M. and Deci, E.L., 2000. 'Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being'. The American Psychologist, 55(1), pp.68–78.

33. Scheuerman, M.K., Hanna, A. and Denton, E., 2021. 'Do datasets have politics? Disciplinary values in computer vision dataset development'. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), pp.1-37.

34. Stvilia, B. and Gibradze, L., 2022. 'Seeking and sharing datasets in an online community of data enthusiasts'. Library & Information Science Research, 44(3), 101160.

35. Stvilia, B., 2021. 'An integrated framework for online news quality assurance'. First Monday, 26(7). Available at: https://firstmonday.org/ojs/index.php/fm/article/view/11062

36. Stvilia, B. and Gasser, L., 2008. 'Value-based metadata quality assessment'. Library & Information Science Research, 30(1), pp.67-74.

37. Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C., 2007. 'A framework for information quality Assessment'. Journal of the American Society for Information Science and Technology, 58(12), pp.1720-1733.

38. Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M. & Marty, P. F. 2015. 'Research project tasks, data, and perceptions of data quality in a condensed matter physics community', Journal of the Association for Information Science and Technology, 66(2), pp. 246-263.

39. Stvilia, B., Wu, S. & Lee, D. J. 2019. 'A framework for researcher participation in research information management systems', The Journal of Academic Librarianship, 45(3), pp. 195-202. Available at: https://doi.org/10.1016/j.acalib.2019.02.014

40. Sun, Z., Strang, K. & Li, R. 2018. 'Big data with ten big characteristics', Proceedings of the 2nd International Conference on Big Data Research, New York, NY: ACM, pp. 56–61.

41. Swarup, S., Braverman, V., Arora, R., Caragea, D., Cragin, M., Dy, J. ... & Yang, C. 2018. 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. Available at: https://par.nsf.gov/servlets/purl/10113364

42. The DataOne Webinar Series 2020a. Assuring the quality of your data: A natural history collection community perspective. Available at: https://www.dataone.org/webinars/assuring-quality-your-data-natural-history-collection-community-perspective/

43. The DataOne Webinar Series 2020b. FAIR'er data through Semantics in NSF's DataONE and Arctic Data Center. Available at: https://www.dataone.org/webinars/fairer-data-through-semantics-nsfs-dataone-and-arctic-data-center/

44. U.S. Congress 2002. The Sarbanes-Oxley Act, 107th Cong., H.R. 3763.

45. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A. ... & Mons, B. 2016. 'The FAIR Guiding Principles for scientific data management and stewardship', Scientific data, 3(1), pp. 1-9.

46. Yoon, A. and Lee, Y.Y., 2019. Factors of trust in data reuse. *Online Information Review*, *43*(7), pp.1245-1262.

47. Zilinski, L. D., Barton, A., Zhang, T., Pouchard, L. & Pascuzzi, P. 2016. 'Research data integration in the Purdue libraries', Bulletin of the Association for Information Science and Technology, 42(2), pp. 33–37.