This is a preprint of an article accepted for publication in the Journal of the Association for Information Science and Technology:
Stvilia, B., Pang, Y., Lee, D. J., & Gunaydin, F. (in press). Data quality assurance practices in research data repositories – A systematic literature review. An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology.* https://doi.org/10.1002/asi.24948

Data Quality Assurance Practices in Research Data Repositories – A Systematic Literature Review

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

Fatih Gunaydin[1]

[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.

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

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

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

Abstract

Data quality issues can significantly hinder research reproducibility, data sharing, and reuse. At the forefront of addressing data quality issues are research data repositories (RDRs). This study conducted a systematic analysis of data quality assurance (DQA) practices in RDRs, guided by activity theory and data quality literature, resulting in conceptualizing a data quality assurance model (DQAM) for RDRs. DQAM outlines a DQA process comprising evaluation, intervention, and communication activities and categorizes 17 quality dimensions into intrinsic and product-level data quality. It also details specific improvement actions for data products and identifies the essential roles, skills, standards, and tools for DQA in RDRs. By comparing DQAM with existing DQA models, the study highlights its potential to improve these models by adding a specific DQA activity structure. The theoretical implication of the study is a systematic conceptualization of DQA work in RDRs that is grounded in a comprehensive analysis of the literature and offers a refined conceptualization of DQA integration into broader frameworks of RDR evaluation. In practice, DQAM can inform the design and development of DQA workflows and tools. As a future research direction, the study suggests applying and evaluating DQAM across various domains to validate and refine this model further.

1. Introduction

Ethical implications of data quality are undeniable. The quality of data and information affects our decision-making, the outcomes of our activities, and, by extension, our wellbeing and reputation (Mason, 1986). As such, maintaining high-quality data is a critical component of any data management workflow. The contexts of data quality assurance (DQA) actions encompass a wide range, including but not limited to quality checks and improvements carried out by data curators and repository managers, data cleaning by researchers and students for academic projects or in DQA hackathons, assessing dataset quality for AI training or the credibility of AI applications, or in policy and business strategizing (Gururangan et al., 2022; Scheuerman et al., 2021; Schwabe et al., 2024).

Significant investments are being made by universities and national research laboratories to establish robust, secure systems for managing the digital research data of their faculty, researchers, and students. These initiatives are driven by the researchers' need for data preservation and dissemination (NASEM, 2020; Tenopir et al., 2020), along with mandates from governmental funding entities that require public sharing of data to benefit the public, support research and teaching, and enhance research reproducibility (NASEM, 2019, 2020; Nelson, 2022; NSTC, 2022). Additionally, compliance with federal and state regulations and laws necessitates ensuring data quality and safeguarding privacy (Barrett, 2019; U.S. Congress, 2002). Some institutions also aim to track and analyze the impact of data produced by their faculty for promotion and tenure decisions (Lyon, 2012). However, sharing and reusing data are often hindered by concerns about data quality. Data owners may worry about the adequacy of documentation of their data and its potential misuse and misinterpretation by users (Tenopir et al., 2015; Stvilia et al., 2015). Conversely, users seek data that is useful and accurately and completely represents the studied phenomena (Boyd & Crawford, 2012; Ng, 2021). Moreover, data is often collected for specific purposes, and without sufficient metadata and explanation, determining those original purposes can be challenging, limiting its further use (Swarup et al., 2018).

One of the main distribution channels of research data is research data repositories (RDRs) operated by research universities, national research labs, and research consortia and networks. In an RDR, DQA is usually a part of its data curation process. RDRs and associated communities of practices have developed conceptual models of data curation as well as models and tools for DQA and system trustworthiness evaluation (e.g., Ball et al., 2012; CoreTrustSeal Standards and Certification Board, 2022; Lacagnina et al., 2022; Peng et al., 2015). In addition, RDRs' DQA practices are often shaped by general quality standards such as ISO 8000, ISO 9000, and ISO 19157. The academic field, too, has contributed several frameworks and models for information and data quality assurance (e.g., Eppler, 2003; Stvilia et al., 2007; Wang & Strong, 1996).

Furthermore, there is a renewed emphasis on enhancing research data quality, aiming for data to be findable, accessible, interoperable, and reusable (FAIR; Wilkinson et al., 2016) in data management and curation communities. Studies have assessed how well data repositories adhere to these principles. In addition, there have been efforts to refine and expand the FAIR principles to assess specific elements of data repository quality, like the quality of services provided, introducing specific metrics and scenarios for their application (Dunning et al., 2017; Devaraju & Herterich, 2020; Koers et al., 2020). The emphasis of current FAIR implementations is on the quality of metadata, systems, and services. Data quality is often seen as implicitly included within the FAIR principles (ECDRI, 2018). While the quality of metadata and system services is essential, it cannot substitute for the intrinsic quality of the data itself, a key component of any information system success (DeLone & McLean, 2003; Koers et al., 2020). Providing detailed information about a dataset's quality characteristics, such as its accuracy, completeness, and reliability, is crucial for enabling its reuse (Peng et al., 2022; Zhou et al., 2016). Metrics that inform about missing values, the study sample's representativeness, and the study's and dataset's peer review status help researchers evaluate the dataset for potential reuse. Ultimately, the decision to trust and reuse a dataset heavily relies on its quality (Yoon & Lee, 2019).

Government research agencies and research funding bodies, such as the National Science Foundation (NSF), National Institutes of Health (NIH), and National Science and Technology Council (NSTC), emphasize high data quality in repositories. Their strategies for obtaining this objective include requiring research projects to develop data management plans and endorsing the assessment and certification of RDRs for their reliability and utility in data curation and sharing. Specifically, the NSF and NIH mandate that researchers submit data management plans to promote data integrity, reproducibility, and high quality metadata (European Commission, n.d.; NSF, 2024; NIH, 2024; NSTC, 2022). Additionally, community-based services like CoreTrustSeal Certification offer peer evaluations of the human and technical infrastructure of research data repositories, as well as their DQA practices, to foster continuous improvements (CTSC, 2022). The European Commission has also played an important role in shaping data practices in RDRs by publishing a report on the European strategy for data (European Commission, 2020). This strategy emphasizes Europe's commitment to openness, fairness, diversity, democracy, and confidence in data sharing and reuse. The strategy report delves into several critical data quality-related issues that need addressing. These include data interoperability, authenticity, infrastructure, technologies, governance, and skills. To tackle these challenges, the European Union (EU) has devised a comprehensive framework. This framework aims to strengthen Europe's capabilities in adhering to FAIR data principles. It involves developing robust data infrastructure for hosting, processing, and utilizing data, as well as investing in continuous resources and training (European Commission, 2020).

Thus, RDRs are at the forefront of research data quality activities. Surveying the current landscape of RDRs' DQA practices through a theoretical lens of the data quality literature can lead to reusable knowledge about the sociotechnical aspects of DQA work associated with this type of information system. Although there have been empirical studies of DQA practices in RDRs (e.g., Kindling & Strecker, 2022; Stvilia & Lee, 2024), a systematic, comprehensive literature review of RDRs' DQA practices is lacking. Our study addresses this gap.

2. Research Questions

Research data curation, which includes DQA, entails a multitude of activities, stakeholders, technologies, policies, and standards, making it a complex sociotechnical process. One primary inhibitor of research reproducibility and replicability, as well as research data sharing and reuse, is concern about data quality. Examining the current landscape of DQA in RDRs is crucial for understanding how RDRs define and ensure data quality and what skills RDR staff needs to complete DQA activities. This understanding can inform the design and development of new, effective DQA workflows and the evaluation of the existing ones in RDRs. It can also facilitate the development of DQA guidelines and training programs for RDR managers and data curators. Despite a considerable amount of literature on research data curation, there is a lack of systematic literature analysis specifically focused on DQA in RDRs. The recent introduction of laws and regulations mandating open access to publicly funded research, as well as ensuring the quality of its data, further underscores the importance of such examination.

This paper addresses this gap by examining the literature on DQA in RDRs for the following research questions:

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

3. Design

The study used a systematic literature review to address the above-defined research questions. To identify potential articles for our literature analysis, we searched the Web of Science (WOS). We refined our search query iteratively to ensure that it accurately represented our research questions and yielded as many relevant articles as possible. The finalized version of the search query was the following:

TS (Topic) = ("data quality" AND (data repositories OR database OR data archival systems OR archives)) OR AB (Abstract) = ("data quality" AND (data repositories OR database OR data archival systems OR archives))

To be included in the study, articles had to meet two additional criteria besides being relevant to the research question: they had to be peer-reviewed and published in English between 2013 and 2023. We decided to set the temporal scope for our literature review to last 10 years (i.e., 2013 to 2023). Selecting the most recent articles for a systematic literature review helps with maintaining the review's relevance and accuracy. Recent articles capture the latest findings, methodologies, and theoretical developments, ensuring that the review reflects the current, state of the art perspective in the subject area under the examination (Petticrew & Roberts, 2006). At the same time, we thought that the 10 year temporal scope of the review was sufficient to provide a thorough coverage of most characteristic DQA practices in RDRs.

The search of the WOS database produced 281 articles, as shown in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) task flow diagram (Page et al., 2021; Figure 1). These articles underwent a meticulous manual screening process by two authors to evaluate their relevance to the research questions and compliance with the inclusion criteria (see Figure 1). Based on abstracts, the initial screening resulted in the shortlisting of 104 articles for further evaluation. Subsequently, these articles underwent a rigorous full-text review conducted by the same two authors. Following this second round of screening, a total of 45 articles were selected for in-depth content analysis.

The study used NVivo software to conduct a thematic content analysis of the selected articles. The analysis was anchored in a theoretical framework based on activity theory (Kaptelinin & Nardi, 2012) and data quality literature. A detailed description of the theoretical framework is presented in the next section of the paper. The use of the theoretical framework allowed for a structured and systematic analysis of the literature and summarization and interpretation of the analysis findings according to the various codes identified. The first step of the content analysis involved developing a set of a priori codes grounded in the theoretical framework and research questions. Next, one author iteratively analyzed the content of the collected articles, searching for both the a priori and newly emerging themes (Bailey, 1994).

However, this phase of our literature analysis did not lead to a priori thematic and data saturation (Saunders et al., 2018). In particular, the coded content did not fully exemplify the theoretical concepts and relationships predicted by the study’s guiding theoretical framework. The concept of saturation in empirical data analysis is fluid, as new findings may emerge with additional data sources and/or cases. Hence, saturation determination ought to focus on identifying when additional data gathering exhibits 'diminishing return’ and no longer contributes significantly to the overarching narrative or theoretical framework (Bailey, 1994; Strauss & Corbin, 1990). We employed a "snowball" method to broaden our data sample. We utilized references from the articles we had already collected and analyzed to find more sources to include in our literature review. This method resulted in the addition of 21 articles (see Figure 1). Two authors analyzed half of this extended sample (i.e., 66 articles) using the coding scheme developed at the initial analysis phase as a starting point. To ensure coding reliability, the authors reexamined a random sample of six articles from the coded sample. They discussed the codings of the subsample, identified and resolved differences, and then updated their coding of the rest of the articles in the sample accordingly. After completing coding, the study utilized a bottom up, inductive approach to integrate thematic codes found within the sample into 13 categories. The categories were in line with the research questions and the broad concepts derived from the study's theoretical framework (see Table 1).

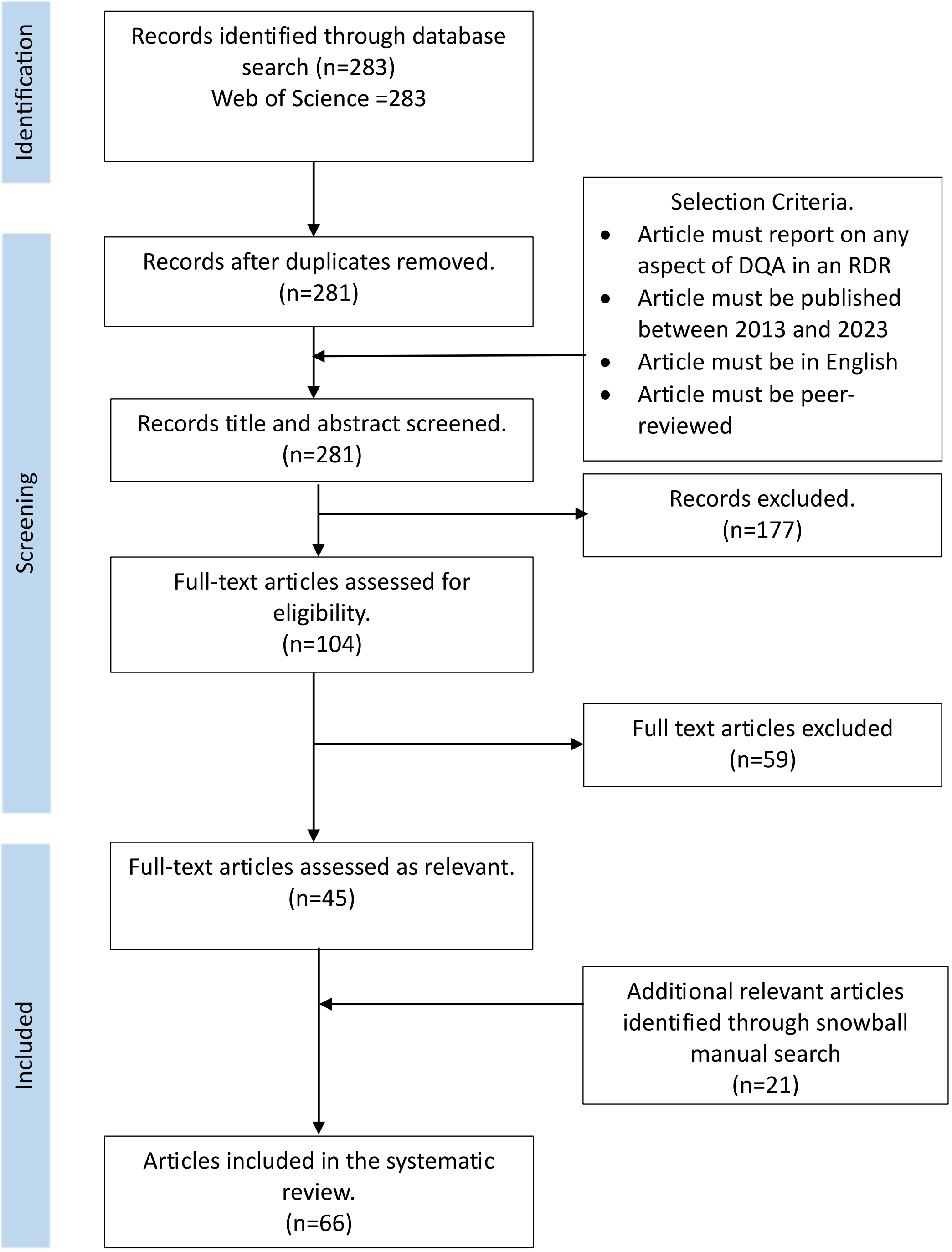

Identification
Screening
Included
Records identified through database search (n=283)
Web of Science =283
Records after duplicates removed.
(n=281)
Selection Criteria.
• Article must report on any aspect of DQA in an RDR
• Article must be published between 2013 and 2023
• Article must be in English
• Article must be peer-reviewed
Records title and abstract screened.
(n=281)
Records excluded.
(n=177)
Full-text articles assessed for eligibility.
(n=104)
Full text articles excluded
(n=59)
Full-text articles assessed as relevant.
(n=45)
Additional relevant articles identified through snowball manual search
(n=21)
Articles included in the systematic review.
(n=66)

Figure 1. The sample selection process.

Table 1. Major themes of the analysis.

Major theoretical concepts used to define 13 categories of themes	**Theories**	**Examples of the associated code(s) and/or subcodes**
Activity	Activity Theory	Activities, Conceptualization, Evaluation, Intervention, Communication
Subject	Activity Theory	Subject
Object/Objective	Activity Theory, DQ Literature	Objective, Data Quality Definition
Data types		Data Type
Metadata types		Metadata Type
Actions	Activity Theory, DQ Literature	Actions, Define DQ, Evaluate DQ, Evaluate MQ, Coordinate DQA, Educate DQA, Data Creation Process Improvement, Product Quality Control, Rework, Scrap, Improve Design of DQA Activities, Improve Quality of DQA Infrastructure, Improve Quality of Staff
Tools	Activity Theory, Infrastructure Theory	Tools, DQ Measurement Tools, DQ Intervention Tools, DQ Communication Tools
Community	Activity Theory	Community, RDR, Organization, Government
Norms and Rules	Activity Theory	Rules, Norms, Standards
Division of Labor, Roles	Activity Theory, DQ Literature	Division of Labor, Roles, Provider, Curator, User
DQ Dimensions	DQ Literature, D&M Model, Mason Model	Intrinsic Quality Dimensions, Product Level Quality Dimensions
DQA Strategies, Prioritize	DQ Literature	DQA Strategies, Prioritize, Prioritize by Current Quality, Prioritize by Available Expertise
Skills	Infrastructure Theory, DQ Literature	Skills

4. Theoretical background and conceptualization of DQA in RDRs

This literature analysis was informed by activity theory, supplemented with theoretical concepts and models drawn from the domains of information quality and information systems. The following section outlines the framework and its components. Activity theory presents a general model for understanding the concept of activity. It defines an activity as a relationship between the *subject* of the activity and the *objective* it aims to achieve, which is mediated by the *tools* used and the *community* or *organization* to which the subject belongs. The community or organization mediates the activity's subject through its rules and conventions while at the same time mediating the activity's objective through *division of labor* and associated *roles* (Kaptelinin & Nardi, 2012; Figure 2). Infrastructure theory further explains that the tool component of the activity model integrates human and technical infrastructures of the activity through the use of *standards*. Thus, for a tool to effectively mediate an activity in an organization, its technical and human components must be well-aligned with each other, including the *skills* of the activity's participants (Lee et al. 2006; Star & Ruhleder, 1994; Figure 2).

The data quality literature can provide further specificity for conceptualizing DQA activities in RDRs. In particular, it posits DQA as a process of data quality definition, evaluation, and intervention (Stvilia et al., 2007; Wang et al., 1998). A general definition of information quality originated from the manufacturing quality literature is "fitness for use" (Juran, 1992). This definition has been widely adopted and used for defining information or data quality, pointing to the contextual nature of quality evaluation (Lee et al., 2002; Stvilia et al., 2007; Wang & Strong, 1996). It is no surprise that many data and information quality evaluation frameworks reference the contexts of evaluation when conceptualizing data quality. These contexts can be broadly categorized as intrinsic and extrinsic. An example of referencing the intrinsic context of evaluation is the intrinsic data quality type proposed by Wang and Strong (1996), which assesses the quality of a data product standalone relative to a general, product specific standard or norm. The contextual data quality included in the data quality typologies by Wang and Strong (1996) and Stvilia et al. (2007) can serve as an example of grounding data quality definitions in the extrinsic context of evaluation. The contextual data quality type can be further divided into meeting subjective expectations and objective task or activity specific requirements for quality (Eppler, 2003). Furthermore, the quality of data can be evaluated directly by evaluating its content and the process of its creation or indirectly based on the reputation of its source and/or the record of its mediation (i.e., reputational quality; Stvilia et al., 2007).

Data quality is a multidimensional concept (Wang & Strong, 1996). That is, data quality can be perceived and conceptualized through a set of virtues or components commonly referred to as data dimensions. There have been multiple attempts to develop systematic sets of dimensions in the data and information quality literature (e.g., Eppler, 2003; Lee et al., 2002; Wang & Strong, 1996). A recent review of data quality evaluation frameworks can be found in Cichy and Rass (2019). Many dimension-based conceptualizations of data and information quality have been tailored towards the domain of data creation and use or the type of data. For instance, for research data, Stvilia et al. (2015) identified 13 dimensions prioritized by condensed matter physicists in their perception of data quality. The dimensions ranged from accuracy and reliability to currency and simplicity. A study of molecular biology researchers, on the other hand, identified 16 dimensions as relevant to their understanding of data quality, ranging from accessibility and accuracy to unbiased and understandability (Huang et al., 2012).

The data quality literature also distinguishes between *intrinsic quality* characteristics or dimensions and quality dimensions emphasized at the *product* level of data, such as accessibility or ease of understanding (Wang &

Strong, 1996; Figure 2). Indeed, the total data quality management (TDQM) approach proposed by Wang et al. (1998) stipulates treating and managing datasets as information or data products. This approach includes understanding the needs of reusers for data products, managing data products' production process and their lifecycles, and assigning someone to manage them as a product manager. TDQM stipulates that if data is treated as a byproduct instead of a product, then ensuring its product level quality might not be prioritized (Wang et al., 1998).

The interpretations of the structures of data quality definitions and their operationalizations can be further informed by DeLone & McLean's information systems success model (D&M model; DeLone & McLean, 2003) and Mason's ethical dimensions of information technology (Mason, 1986). The D&M model offers a structure to evaluate the success of information systems through six constructs, including three quality constructs: *system quality*, *information quality,* and *service quality*. For RDRs, this model suggests that RDR managers and curators, as well as other stakeholders, may connect an RDR's DQA specific objective (i.e., serve high quality data products) to all three quality related predictors of the RDR's overall success. That is, according to the D&M model, the perception of RDR's success is shaped by how usable, reliable, fast, and ethical the RDR is to its stakeholders and how well it ensures the quality of data it manages and the quality of its services. The latter includes the desirable characteristics of the service and support provided by the RDR's staff, such as technical *expertise*, *reliability, responsiveness,* and *empathy* (Petters et al., 2013).

The system and data quality constructs of the D&M model can be expanded with the Mason model, which focuses on the desirable ethical characteristics of successful information systems. It highlights the importance of considering privacy, quality, ownership, and accessibility in managing information systems. In the context of DQA, this model can help with inferencing the need for RDRs to ensure the intrinsic *quality* of data while respecting the *privacy* of data subjects, manage data as a product with proper *intellectual property rights* and usage terms, and provide equitable *access* to data (see Figure 2).

In a DQA workflow, assessing data quality is typically followed by an intervention activity aimed at addressing or alleviating the identified issues with the quality of data and metadata (Stvilia et al., 2007; Wang et al., 1998). In product manufacturing, quality can be improved by improving the production process or enhancing the quality control of produced products. The latter is accomplished by *reworking* or *scrapping* already-made products (Cook, 1997). Hence, data quality interventions in RDRs can also be divided into two categories: *data creation process improvement* and *product quality control* interventions.

In addition, the intervention category may also include the activities of improving an RDR's DQA process itself, making it more effective and efficient (Ballou et al., 1998). These activities may affect all DQA activities and comprise *improving the design* of DQA activities, *enhancing the quality of the RDR's DQA infrastructure* used, including the quality of its human infrastructure - *the quality of human resources* performing DQA activities (Star & Ruhleder, 1994; Figure 2). For example, the literature on online communities shows that the English Wikipedia community constantly monitors and aligns the design of its information quality assurance activities with its evolving information quality assurance challenges and problems. These include evaluating and selecting members for serving in quality assurance roles and designing and deploying new information quality assurance tools, such as bots, to make the community's quality assurance actions more effective and scalable (Stvilia et al., 2008).

Furthermore, a DQA process is usually distributed and involves multiple roles and agents. Hence, a *communication activity* is essential to any data and information quality assurance work and accompanies all other DQA activities. It may include discussing information quality problems, communicating quality evaluation information, and discussing and coordinating quality evaluation and intervention actions (Stvilia et al., 2008; Figure 2).

Finally, it is important to remember that resources for managing data and ensuring its quality are limited. RDRs may need to *prioritize their DQA activities*. RDRs may need to decide which datasets to evaluate and intervene in and when and to what extent. To determine the priority level of a dataset for DQA, data curators may consider the organization's collection development policy, the priorities of stakeholder communities, the current quality of the dataset, and the expertise they have (Cosley et al., 2005; Stvilia et al., 2007, 2008; Figure 2).

Figure 2 summarizes the theoretical framework synthesized in this section. It specifies the structure of a DQA process, which comprises three main activities: evaluation, intervention, and communication. The diagram illustrates how these activities are influenced by the RDR, its parent organization, a community of practice, or the government through various tools, rules, conventions, policies, laws, and the division of labor with associated roles. Additionally, the framework includes the conceptualizations of the evaluation and intervention activity structures from DQA theory and highlights the DQA prioritization facets, such as the current level of a data product's quality and available expertise in the RDR. The framework also emphasizes the importance of skills and expertise in aligning the human and technical infrastructures used in a DQA process.

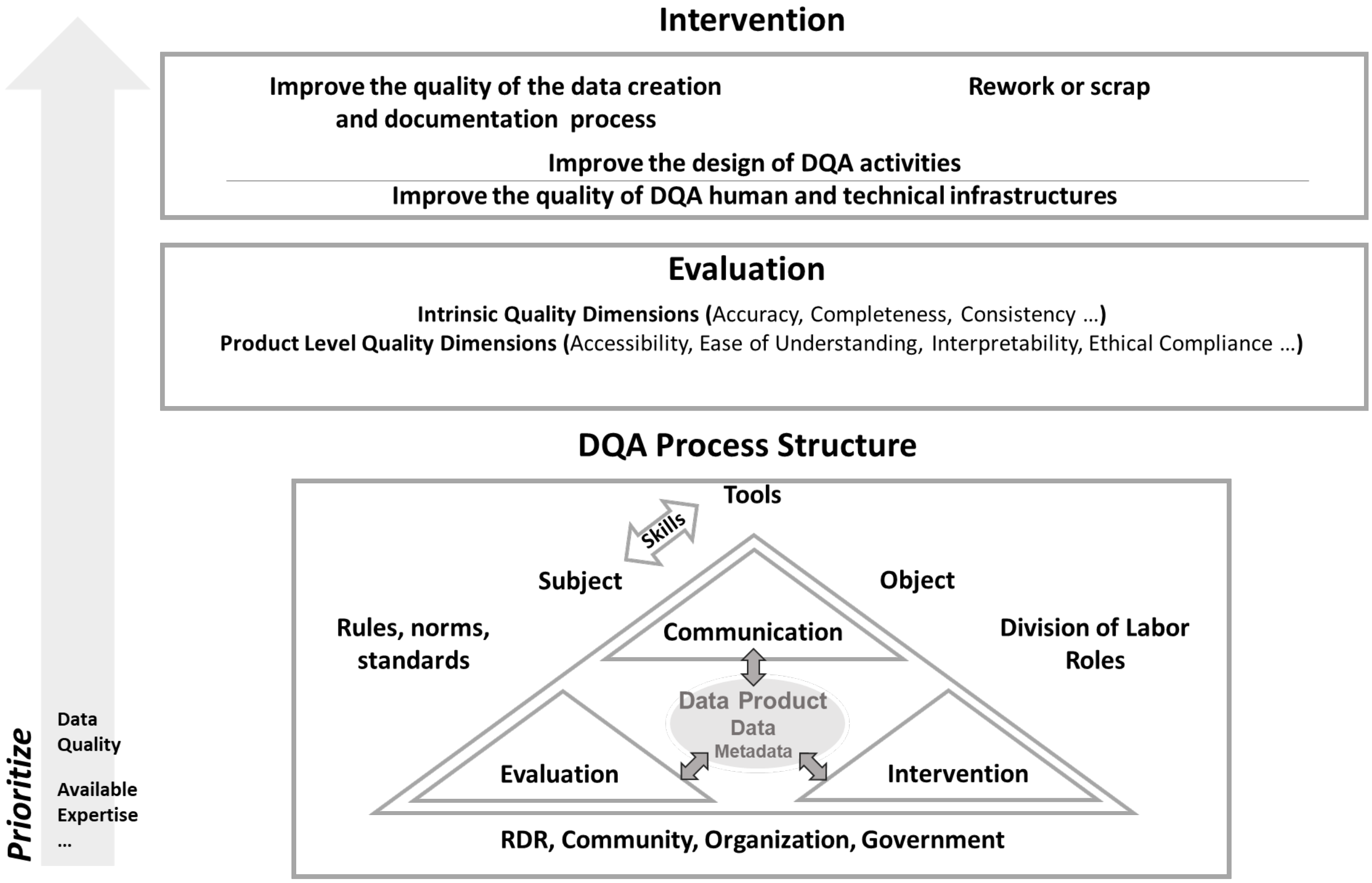

Intervention
Improve the quality of the data creation and documentation process
Rework or scrap
Improve the design of DQA activities
Improve the quality of DQA human and technical infrastructures
Evaluation
Intrinsic Quality Dimensions (Accuracy, Completeness, Consistency …)
Product Level Quality Dimensions (Accessibility, Ease of Understanding, Interpretability, Ethical Compliance …)
DQA Process Structure
Tools
Skills
Subject
Object
Rules, norms, standards
Communication
Division of Labor Roles
Data Product
Data
Metadata
Evaluation
Intervention
RDR, Community, Organization, Government
Prioritize
Data Quality
Available Expertise
…

Figure 2. The theoretical framework of the study.

5. Findings and Discussion

In the preceding section, we developed a theoretical framework for this study, as illustrated in Figure 2. This framework guided our literature analysis and facilitated the interpretation of our findings. In this section, we present the results of our analysis, structured according to the same framework, culminating in the synthesis of a data quality assurance model for RDRs.

5.1 Data Types, Roles, and Uses

Before discussing DQA activities in RDRs, it is crucial to explore both the types of data targeted by these activities and the roles involved. RDRs handle a variety of data types, including observational, experimental, simulation, and derived data. Common methods for data collection include questionnaires, systematic surveys, observations of clinical variables, experiments, and the use of secondary data (Chen et al., 2020; García-de-León-Chocano et al., 2015; Hirdes et al., 2013; Marinov et al., 2014).

DQA roles can be grouped into three main categories: data creators/providers (i.e., researchers who submit datasets), DQA agents (including roles like data curators and RDR managers, DQA analysts, and domain experts), and data users. Data uses can be defined as the examination or analysis of data to provide actionable knowledge in different activities such as research, design, services, etc. (Colquhoun et al., 2020; Hirdes et al., 2013; Marinov et al., 2014; Stvilia & Lee, 2024).

5.2 Evaluation

Our analysis revealed that RDRs may define quality differently at various levels of granularity. Some studies defined data quality as the degree to which data meet general use requirements (e.g., Gonzalez-Vidal, 2022; Lacagnina et al., 2022; Stamnas et al., 2016). Other definitions of data quality referenced ensuring clear identification, functionality, and long term accessibility of data (Peer et al., 2014) or achieving data maturity regarding searchability and accessibility of data and associated services (Peng, Gross, & Edmunds, 2022). Kindling and Strecker (2022) examined metadata quality and data quality criteria to identify data repositories' operational definitions of data quality. Their study participants prioritized appropriate data documentation and suitability to the scope of the repository as the most relevant dimensions of quality, while timeliness and novelty were the least frequently selected criteria. More recently, Stvilia and Lee (2024) explored how managers and curators of RDRs perceive data quality. They identified several facets referenced in these practitioners' definitions of data quality: data documentation quality, ethical and legal compliance, the quality of the data collection and creation process, data compatibility and accessibility, data quality, data provenance and authenticity, and system quality. Furthermore, Stvilia and Lee (2024) analyzed the quality dimensions RDRs used to assess data quality, revealing that completeness, consistency, simplicity, accuracy, and relevance were the dimensions most frequently utilized by the repositories.

DQ dimensions and associated metrics are used to characterize and communicate data quality problems. Studies reported data inaccuracies across various areas, including biology (e.g., Urbano et al., 2021), meteorology (e.g., Freeman et al., 2017; Longman et al., 2018), geography (e.g., Samuel-Rosa et al., 2019), health (e.g., Hirdes et al., 2013; Vihinen et al., 2016), among others. For example, in health-related datasets, inaccuracies in recording patients' weight and height have led to measurements that exceed normal thresholds (Hirdes et al., 2013). Similarly, geographic datasets have shown incorrect latitude and longitude entries that do not conform to national ranges (Samuel-Rosa et al., 2019). Such errors can significantly impact subsequent research. The sources of these inaccuracies vary; some are due to faulty tools or technician mistakes (Juárez et al., 2019; Kulmukhametov et al., 2021; Urbano et al., 2021), while others result from data transformation errors (Longman et al., 2018).

Data inconsistencies often arise when researchers seek to combine datasets from disparate time periods or organizations. For instance, variations in data formats and nomenclature arise due to the different goals pursued by each organization during data collection (Corrales et al., 2018; Grossberg et al., 2018; Juárez et al., 2019; Shekhovtsov & Eder, 2022; Urbano et al., 2021; Vignolo et al., 2021). Also, researchers may need to integrate historical and contemporary datasets in their studies. However, they may encounter challenges due to discrepancies in the definitions of some variables (Borries et al., 2013; Freeman et al., 2017). For instance, Freeman et al. (2017) reported variations in the concept of sea level pressure across different datasets, while Borries et al. (2013) indicated the lack of a universally accepted definition for life history. Furthermore, some medical data may exhibit logical inconsistencies, such as alterations in blood pressure that do not align with established pathological understanding (e.g., Hirdes et al., 2013).

Incompleteness pertains to the absence or omission of data or expected information inside a dataset. This issue can have a substantial impact on the use and dependability of data for the purposes of analysis, decision-making, and operational procedures. Data loss can occur due to a multitude of circumstances. One of the most frequent issues is the failure to get all the necessary data during a data gathering process (Araujo-Pradere et al., 2019; Heidebrecht et al., 2014). In addition, there can be missing variables due to legal and privacy requirements (Freeman et al., 2017; Windhager et al., 2019). Likewise, redundancy and duplication of data might raise doubts about its reliability and can negatively affect its reuse (Chen et al., 2020; Corrales et al., 2018; Zhang et al., 2019). Finally, a dataset missing or having incomplete metadata can degrade its usability. Heidebrecht et al. (2014) highlight that the absence of information regarding the reliability and suitability of data impacts its subsequent analysis and reuse.

To gain an understanding of how RDRs define and operationalize data quality, we crosswalked and aggregated dimensions reported in the literature as a part of data quality definitions and/or discussions of data quality problems. We removed redundancies and grouped the resultant set into two groups: intrinsic data quality and data product level quality dimensions. The latter includes system quality dimensions as well as some data quality dimensions such as usability and reputation (see Tables 2,3). To provide definitions for each of the dimensions in the list, we adapted definitions of information and data quality dimensions from Stvilia et al. (2007) and Wang and Strong (1996).

Our analysis of the literature revealed nine intrinsic quality dimensions that the reviewed papers employed while defining or referring to data quality in DQA activities in RDRs (see Table 2). The most frequently referenced dimensions were accuracy, completeness, consistency, and currency/timeliness. The least frequently used dimensions were precision and simplicity. The analysis also identified eight data product level dimensions that are emphasized when an evaluator takes the data product perspective (Table 3). As expected, the most frequently referenced dimension was accessibility, followed by usability and reputation. The least frequently mentioned dimensions were integrity and authenticity.

The intrinsic vs. product level quality perspectives on data quality reported in the data quality literature have been reflected in how different studies in this literature analysis defined data quality (see Tables 2,3). In particular, research study designs and participants often conflated conceptual components of data quality (i.e., dimensions) with data quality metrics used to operationalize or measure those dimensions. To compare data quality assessment models reported in the literature and interpret and use/reuse them in a systematic way, it is important that data quality dimensions are distinguished from their operationalizations (i.e., metrics; Stvilia &

Lee, 2024). Similarly, our analysis showed that study designs and participants often do not distinguish among data, metadata, system, and service quality characteristics when defining data quality (e.g., Kindling & Strecker, 2022; Peer et al., 2014; Stvilia & Lee, 2024).

Table 2. Intrinsic data quality dimensions.

Dimension (# of studies)	Definition	Studies referencing or discussing the dimension
Accuracy (26)	The degree to which data is a correct or valid representation of an object, relation, process, or event	Aerts et al., 2021; Freeman et al., 2017; García-de-León-Chocano et al., 2015; Freeman et al., 2017; Gualo et al., 2021; Gutmann et al., 2004; Hirdes et al., 2013; Juárez et al., 2019; Kindling & Strecker, 2022; Kulmukhametov et al., 2021; Lacagnina et al., 2022; Liao et al., 2021; Lin et al., 2020; Longman et al., 2018; Rajan et al., 2019; Reimer et al., 2016; Samuel-Rosa et al., 2019; Shekhovtsov & Eder, 2022; Singh et al., 2020; Smith et al., 2018; Stvilia & Lee, 2024; Tian et al., 2021; Urbano et al., 2021; Vihinen et al., 2016; Zhang et al., 2019; Zhou et al., 2016
Completeness (26)	The extent to which data is a complete representation of another object, relation, process, or event	Aerts et al., 2021; Araujo-Pradere et al., 2019; Estiri et al., 2018; Freeman et al., 2017; García-de-León-Chocano et al., 2015; Grossberg et al., 2018; Gualo et al., 2021; Gutmann et al., 2004; Heidebrecht et al., 2014; Johnston et al., 2018; Kapsner et al., 2021; Kindling & Strecker, 2022; Lacagnina et al., 2022; Peer et al., 2014; Rajan et al., 2019; Reimer et al., 2016; Sáez et al., 2016; Shekhovtsov & Eder, 2022; Singh et al., 2020; Smith et al., 2018; Stvilia & Lee, 2024; Thomer et al., 2022; Tian et al., 2021; Vihinen et al., 2016; Windhager et al., 2019; Zhang et al., 2019
Consistency (25)	The extent of consistency in using the same values or structure for representing an object, relation, process, or event	Aerts et al., 2021; Arkhangelskiy et al., 2020; Borries et al., 2013; Corrales et al., 2018; Freeman et al., 2017; García-de-León-Chocano et al., 2015; Grossberg et al., 2018; Gualo et al., 2021; Hirdes et al., 2013; Juárez et al., 2019; Kindling & Strecker, 2022; Kulmukhametov et al., 2021; Longman et al., 2018; Rajan et al., 2019; Reimer et al., 2016; Sáez et al., 2016; Samuel-Rosa et al., 2019; Shekhovtsov & Eder, 2022; Singh et al., 2020; Stvilia & Lee, 2024; Tian et al., 2021; Urbano et al., 2021; Vihinen et al., 2016; Vignolo et al., 2021; Zhang et al., 2019
Currency / Timeliness (14)	The age of data The extent to which the age of the data is appropriate for the task at hand	Aerts et al., 2021; Freeman et al., 2017; Gonzales-Vidal et al., 2022; Gualo et al., 2021; Heidebrecht et al., 2014; Kindling & Strecker, 2022; Rajan et al., 2019; Reimer et al., 2016; Shekhovtsov & Eder, 2022; Singh et al., 2020; Stvilia & Lee, 2024; Tian et al., 2021; Vihinen et al., 2016; Zhang et al., 2019
Lack of Redundancy (7)	The extent of redundancy in data representing an object, relation, process, or event	Chen et al., 2020; Corrales et al., 2018; Freeman et al., 2017; García-de-León-Chocano et al., 2015; Rajan et al., 2019; Stvilia & Lee, 2024; Zhang et al., 2019
Reliability (6)	The extent to which the correctness of data is verifiable or provable	Kindling & Strecker, 2022; Lacagnina et al., 2022; Lin et al., 2020; Gudmundsson et al., 2018; Smith et al., 2018; Stvilia & Lee, 2024
Relevancy (3)	The extent to which data is applicable to the task at hand	Arkhangelskiy et al., 2020; Kindling & Strecker, 2022; Stvilia & Lee, 2024
Precision (2)	The extent of precision/data of data in representing an object, relation, process, or event.	Stvilia & Lee, 2024; Zhou et al., 2016
Simplicity (1)	The extent of cognitive complexity of data	Stvilia & Lee, 2024

Table 3. Data product level quality dimensions.

Dimension (# of studies)	Definition	Studies discussing or referencing the dimension

Accessibility / Availability (9)	The extent to which data are available or easily and quickly retrievable	Arkhangelskiy et al., 2020; Chen & Chen, 2020; Kindling & Strecker, 2022; Larsen et al., 2016; Lin et al., 2020; Shekhovtsov & Eder, 2022; Stvilia & Lee, 2024; Thomer et al., 2022; Vihinen et al., 2016
Usability / Interpretability (6)	The extent of data being usable and interpretable	Arkhangelskiy et al., 2020; Chen & Chen, 2020; Johnston et al., 2018; Kindling & Strecker, 2022; Lacagnina et al., 2022; Smith et al., 2018
Ethical Compliance (5)	The extent to which data complies with the ethical principles of a particular community or culture, including the privacy and confidentiality of data	Gutmann et al., 2004; Kindling & Strecker, 2022; Peer et al., 2014; Stvilia & Lee, 2024; Vihinen et al., 2016
Legal Compliance (3)	The extent to which data complies with the laws of a particular country	Kindling & Strecker, 2022; Peer et al., 2014; Stvilia & Lee, 2024
Stability (3)	The extent of data remaining accurate and complete in time and space	Aerts et al., 2021; García-de-León-Chocano et al., 2015; Zhou et al., 2016
Integrity (3)	The extent to which the integrity of data is preserved	Liao et al., 2021; Lin et al., 2020; Stvilia & Lee, 2024
Reputation / Credibility (2)	The degree of reputation or credibility of data	Gualo et al., 2021; Stvilia & Lee, 2024
Authenticity (2)	The extent to which data is authentic and genuine	Lin et al., 2020; Stvilia & Lee, 2024

5.3 Intervention

Our analysis suggests that the theoretical framework's process and product intervention categories can be subdivided into more specific groups (see Figure 3). The quality of a dataset is primarily influenced by the scientific process used to generate it (Michell, 1990). Multiple studies in this review underscored the importance of establishing quality criteria (e.g., reliability, validity, accuracy) for data collection procedures to ensure data quality (Bayraktarov et al., 2019; Freeman et al., 2017; Heidebrecht et al., 2014; Hirdes et al., 2013). RDRs may enhance the quality of data creation and documentation processes by having RDR staff collaborate with research teams to provide data management planning consultations, which include adhering to standards and best practices for creating and documenting data products and associated items such as software and instruments (Corrales et al., 2018; Stvilia & Lee, 2024). Studies also show that adopting a standardized approach to creating, organizing, and documenting datasets can improve data completeness and comparability, thereby enhancing data quality. For example, McGrath et al. (2022) emphasized the importance of a single body coordinating the collection of clinical medicine data to ensure high data quality. Additionally, food composition data studies utilized the EuroFIR standard thesaurus to ensure consistent value recording for metadata recording and coding (Westenbrink et al., 2016).

Data product quality control may involve selecting and accepting datasets based on the provider's reputation, ensuring that only data from trusted depositors is curated in an RDR. This method, combined with the rejection of datasets due to poor quality or the depositor's refusal to make necessary adjustments, can be classified as scrap data product control (Figure 3). In the rework category of product control, RDRs may request that depositors implement specific quality interventions on submitted datasets, or curators may perform these actions themselves. This process often entails providing depositors with specific guidelines to enhance the data and metadata quality of their submissions. RDR managers and curators collaborate with researchers to discuss and implement necessary changes to their datasets. They may also provide depositors with templates and guides to address information gaps and enhance metadata quality, preparing datasets for publication (Austin et al., 2016; Stvilia & Lee, 2024; Trisovic et al., 2021; Urbano et al., 2021; Westenbrink et al., 2016).

Domain-specific RDRs or those of large research laboratories, more so than generalist repositories, may enhance the quality of data and metadata through direct interventions. These interventions typically require depositor review and approval, either before or after they occur (Stvilia & Lee, 2024). A common data product quality intervention begins with enriching a dataset's metadata with descriptive records and comprehensive documentation, including provenance records that contextualize the data and make it more understandable. RDRs may also convert datasets to more accessible or open file formats to increase their utility. Improvements to dataset usability and accessibility can involve updating, harmonizing, and converting their schemas, content, and metadata to align with community standards and legal requirements (Borries et al., 2013; García-de-León-Chocano et al., 2015; Owens et al., 2022; Stvilia & Lee, 2024; Urbano et al., 2021).

Improving data product quality may involve interventions in the intrinsic quality of a dataset, such as adding missing data, eliminating duplicates and inconsistencies, and correcting data errors identified through statistical analyses like frequency, average, range calculations, or grouping. Additionally, incorporating related items such as software and user guides and verifying their output and content against the data they describe significantly enhance data trustworthiness, comprehensibility, and reusability for end users (García-de-León-Chocano et al., 2015; Koshoffer et al., 2018; Peer et al., 2014; Stodden, 2013). Some studies have reported that RDRs use data visualization to resolve conflicts and assess data accuracy (Kulmukhametov et al., 2021; McFarland et al., 2013). Furthermore, RDRs have employed regression models to estimate missing values, replacing them with the median, mean, or mode (Corrales et al., 2018). Another study highlighted an RDR's use of triangulation with multiple data sources to correct inaccuracies (Samuel-Rosa et al., 2019). In cases where datasets do not meet quality benchmarks, they may need to be rejected or removed from the repository (Hirdes et al., 2013; Stvilia & Lee, 2024). Finally, some RDRs conduct periodic assessments and interventions to maintain dataset quality, ensuring their actionability and stability (Aerts et al., 2021).

Our analysis also indicated that researchers who regard data as mere byproducts of their projects may need to adjust this perspective upon depositing data into RDRs for sharing. For RDRs, these datasets are essential information products intended for user consumption. This shift implies that while depositors and RDR curators may focus on maintaining the intrinsic quality (e.g., accuracy, completeness, consistency) of the datasets to ensure the quality and success of the associated research projects, RDR curators and end-users may also prioritize the product-level quality characteristics of the same datasets, such as usability (Michener, 2015; Peng, Gross, & Edmunds, 2022; Stvilia & Lee, 2024).

Efficient and robust DQA workflows and tools are crucial for accurately assessing and improving a dataset's quality as needed. Thus, DQA may entail refining the design of DQA activities and ensuring adherence to quality standards and best practices. Typically, the design of an organization's DQA process begins by defining data quality for the organization and its stakeholders. Adopting an existing data quality definition and framework and implementing uniform policies and procedures are vital for the soundness of an RDR's DQA process (García-de-León-Chocano et al., 2015; Gualo et al., 2021; Lavery et al., 2022; Stvilia & Lee, 2024). Moreover, operationalizing process quality criteria into robust metrics and quality checks is essential to identify and rectify logic errors, outliers, missing and invalid values, and duplicates (Chen et al., 2020; Gudmundsson et al., 2018; De Rosa et al., 2023; Estiri et al., 2018; Freeman et al., 2017; Hall & Jensen, 2022; Heidebrecht et al., 2014; Hirdes et al., 2013; Juárez et al., 2019; Thomer et al., 2022; Urbano et al.). Additionally, creating, maintaining, and standardizing a quality assurance provenance record for datasets is crucial for assessing their reliability and usability by reusers (Juárez et al., 2019; Peng et al., 2022).Another key aspect of ensuring the quality of a DQA process is for RDRs to foster strong collaboration and coordination among all stakeholders, including data providers and users, to ensure a unified approach to data quality (Stvilia & Lee, 2024). An RDR's DQA process can be further enhanced by adding dedicated DQA staff who focus on assessing and managing data quality (Heidebrecht et al., 2014; Stvilia & Lee, 2024). Domain-specific RDRs often engage domain experts to enhance the robustness of their DQA workflows. Additionally, establishing user feedback mechanisms can facilitate user engagement in identifying data quality issues, thereby making the RDR's DQA process more efficient (García-de-León-Chocano et al., 2015; Kulmukhametov et al., 2021; Stvilia & Lee, 2024; Urbano et al., 2021).

In addition, enhancing the DQA competencies and skills of staff, along with the DQA components of an RDR's infrastructure, can significantly improve its DQA process. RDRs can achieve this through comprehensive staff training and development programs. By offering continuous education opportunities in DQA, RDR staff members can stay abreast of evolving industry standards and technological changes, enabling them to implement effective and efficient solutions for ensuring the quality of emerging data types (Heidebrecht et al., 2014). Simultaneously, investing in technology is crucial for improving the overall technical infrastructure quality, ensuring that the RDR remains at the forefront of the field and can meet its DQA needs and challenges for new data types and scales (Zhou et al., 2016). Also, an RDR seeking third-party certification and review of its data curation process can benefit from an external evaluation and validation of its DQA process design, potentially leading to process improvements (Stvilia & Lee, 2024).

Managing complex data products for long-term research projects may involve all the above-mentioned data quality intervention facets (see Figure 3). The duration of research projects can vary, with some being short-term and others being long-term, longitudinal studies that require a long-term data management approach embedded within the project and data production process (Kaplan et al., 2021). An embedded approach can lead to higher data management and DQA literacy among the project staff, stronger collaboration to achieve the project's data management objectives, and improved local technology and human infrastructures (e.g., expertise) for data management (Kaplan et al., 2021; Zhou et al., 2016).

The closure of a long-term research project and the transfer of its data collection to an off-site repository for preservation and/or wider sharing can be considered a decontextualizing activity that could lead to data quality problems (Stvilia et al., 2007). Similarly, data product stability can be affected when it relies on and aggregates data produced by multiple researchers or centers (García-de-León-Chocano et al., 2015). Spatial shifts in the context of data management and use may require additional investment to enhance the quality of the data product and its metadata, improving its access, stability, and interoperability with other relevant datasets in the destination repository and for the quality needs of new external users (Kaplan et al., 2021; Stvilia et al., 2007).

The quality of data products can also change over time. Temporal changes in data can be caused by changes in the underlying entities they represent or by changes in the data itself. These changes can be direct, by altering the data, or indirect, due to changes in the data creation process and its context, including mediating factors such as tools, rules, and policies (Stvilia et al., 2007). For instance, a change in a data collection protocol may affect the accuracy or completeness of collected data, as well as the stability and interoperability of the data for aggregation in longitudinal research or practical applications (Aerts et al., 2021).

Live or near real-time data products can be particularly prone to spatial data problems. Malfunctioning or degraded sensors, delays in data transmission, external interference affecting sensor readings, miscalibrated measurement instruments, and software errors can all affect the accuracy, completeness, and stability of a data product (Zhou et al., 2016). To mitigate DQA problems in long-term or real-time data products, data curators can implement robust data transmission protocols with error-checking and redundant systems to prevent data loss and ensure completeness. Regular calibration and maintenance of data collection instruments and sensors can help avoid inaccuracies. Automated validation checks, statistical data profiling, and deduplication processes can identify and eliminate duplicate data using unique identifiers and data monitoring. Standardizing data formats and protocols can resolve inconsistencies and enhance data stability and interoperability. Finally, applying noise identification and removal algorithms can enhance data accuracy (Zhou et al., 2016).

5.4 Communication

Communication plays a crucial role in the DQA process and may involve a continuous cycle of information sharing and feedback among data quality planning, evaluation, and intervention activities. Our analysis showed that a DQA process in an RDR is usually distributed and involves multiple roles and agents. Multiple rounds of quality assurance tasks between data providers and curators can generate some confusion regarding mappings between datasets, agents, DQA actions, resources, and data quality information (Randles et al., 2020). Effective communication between data providers, curators, and users is therefore essential during data quality assurance processes in RDRs (Stvilia & Lee, 2024). Information about data quality can be shared throughout a dataset's lifecycle, from creation to sharing and usage, involving various participants within the RDR's data curation ecosystem (Cho et al., 2015; Peng et al., 2022). When depositors submit datasets, they may include information about any data quality issues in user guides or readme files. Curators may then relay information about identified quality issues and the steps taken to address them through data quality reports and metadata, collaborating with data providers, metadata experts, and IT departments to solve problems related to data and metadata quality. Curators often rely on researchers for the most accurate data information, seeking their assistance to fill in missing information, approve modifications, and verify data accuracy and integrity (Stvilia & Lee, 2024).

RDR managers and curators also communicate pre-submission data and metadata quality requirements to researchers through guides, templates, and other communication methods. Sometimes, they embed themselves in research projects to consult on best practices for data cleaning, data provenance documentation, and preparation for publication (Peng et al., 2022; Stvilia & Lee, 2024).

Curators' communication with data providers and users aids the efficiency and effectiveness of the DQA process (Araujo-Pradere et al., 2019). Users, who may be researchers familiar with data creation and analysis methodologies, can assess the data quality (Faniel et al., 2019) and determine its suitability for their research (Faniel et al., 2016; Faniel & Jacobsen, 2010; Trisovic et al., 2021). It is essential for users to be informed about data quality so they can decide if it meets their data quality needs (McFarland et al., 2013; Vihinen et al., 2016). Users' ability to communicate with data providers to acquire extensive background information about the data enhances data quality (Araujo-Pradere et al., 2019; Faniel et al., 2016, 2019; Faniel & Jacobsen, 2010). Scientific communities and funding bodies increasingly emphasize the importance of reproducible and verifiable research. The push for transparency and openness in research, including the open sharing of research data and gathering feedback from end users on data quality, is crucial for maintaining the integrity and thoroughness of research outcomes. Openness and transparency in data handling lead to increased usage, more comprehensive quality assessments, and corrective actions, ultimately resulting in improved data quality (Orr, 1998; Peng et al., 2022; Stvilia & Lee, 2024). Some RDRs enable users to offer feedback on data quality directly through means like contact forms. Others use ticket systems or implement a system for rating the quality (Stvilia & Lee, 2024). Research communities served by more centralized, subject specific RDRs have developed metadata models to communicate information about the quality of a dataset throughout its lifecycle. For example, the Earth science community has proposed a metadata model that captures information about the quality of data used in the scientific process of creating, curating, and using a data product (Peng et al., 2022). RDRs and research communities have also developed models to communicate the levels of quality or maturity of their data products (Zhou et al., 2016).

Moreover, RDR managers, data curators, and scholarly communication librarians engage in dialogue with both providers and users to make data more accessible and to improve understanding of RDRs' DQA practices. They share critical information about essential DQA standards, practices, and guides with data providers and users (Stvilia & Lee, 2024). Some RDRs and research networks implement training programs for individuals utilizing data (McGrath et al., 2022) and define protocols for data usage to help determine the appropriateness and precision of the data used in research (Longman et al., 2018; Trisovic et al., 2021). They also provide training to RDR staff members to prevent data quality problems (Heidebrecht et al., 2014).

5.5 Standards and Tools

Repositories adopt diverse policies, standards, and best practices to mediate data curation processes and ensure the quality, accessibility, and ethical handling of data (see Figure 3).

Our literature analysis identified several international standards that have been used or proposed to be used for ensuring data quality in RDRs. The ISO 9000 family promotes adopting a quality management system focused on continual process improvement based on measurements and meeting requirements (Lacagnina et al., 2022;

Peng et al., 2015). ISO 19157 provides a framework for capturing data quality evaluation methods and results, classifying quality assessments as direct (i.e., inspecting dataset values) or indirect (i.e., using external knowledge like lineage; Lacagnina et al., 2022). The Open Archival Information System (OAIS) Reference Model, ISO 14721, provides a conceptualization and terminology for archival systems, a necessary context for DQA in RDRs (Peng et al., 2015). ISO 16363, often referred to as the Trustworthy Digital Repositories (TDR) checklist, is a well-established standard that provides audit metrics for certifying trustworthy digital repositories based on the OAIS model (Corrado, 2019; Peng et al., 2022; Yoon, 2014). Its companion ISO 16919 defines requirements for auditing bodies (Corrado, 2019; Yoon, 2014). Other relevant standards identified by the literature analysis include ISO 25012 and 25024, which define quality characteristics, an evaluation process, and associated measures for data products (Gualo et al. 2021), and ISO 19115 for providing standardized metadata to users (Owens et al. 2022). Repositories wishing to be certified typically undergo an external audit based on ISO 16363 (Corrado, 2019). Third-party quality evaluation bodies like CoreTrustSeal play a crucial role in assessing the quality of a RDR's DQA workflow and overall system quality. By setting a series of rigorous standards and requirements, these bodies provide an external benchmark for RDRs to meet or exceed, which not only ensures compliance with best practices but also instills trust in the users of these repositories (Stvilia & Lee, 2024).

Repositories employ various software tools for identifying and visualizing data quality issues and measuring data quality through metrics. Some also provide these tools to their users, enabling them to evaluate data quality and its relevance to their work. Communication about DQA might include the use of metadata vocabularies or quality rating systems that clarify the DQA approach of the repository and the quality of individual datasets to users (Peng et al., 2022; Stvilia & Lee, 2024; Zhou et al., 2016). To aid their designated communities in assessing dataset quality, repositories may issue DQA badges based on criteria like expert review, association with peer-reviewed publications, or inclusion in institutional showcases (Stvilia & Lee, 2024). Moreover, user guides are crucial for conveying information about the dataset's structure, quality assessments, interventions, and any known limitations. These guides, along with data lineage tools, facilitate the sharing of data provenance information (Juárez et al., 2019).

Repositories also utilize software for data product quality enhancement through annotations, tagging, and conversion into more user-friendly formats (Colquhoun et al., 2020; Kapsner et al., 2021; McFarland et al., 2013; Owens et al., 2022). RDRs are often part of university data management systems and leverage this infrastructure for data storage, analysis, and access, enhancing dataset quality through documentation and linkage to related research. RDRs and associated communities also develop and use quality evaluation, cleaning, validating, and enhancement software libraries for performing statistical analyses and quality checks on datasets (Estiri et al., 2018; Kapsner et al., 2021). In instances where journal publications are associated with datasets post-approval, curators might use university research information management systems (RIMS) and web portals to link these publications to datasets if depositors do not provide the necessary details (Stvilia & Lee, 2024).

5.6 Optimizing

When prioritizing datasets for quality assurance, repositories may take into account the *value* of the dataset, its *quality* level, the *expertise* they have in the subject the dataset belongs, the *cost* of ensuring its quality, and the *funding* provided by the depositor or funding bodies to ensure the dataset's quality. In doing so, repositories seek to allocate their resources judiciously and set priorities for their data curation tasks in a way that satisfies the requirements of depositors and users (Lacagnina et al., 2022; Stvilia & Lee, 2024; Figure 3). RDRs are under pressure to demonstrate the net value of their curation activities, including DQA activities, to secure ongoing funding or attract new funding. RDRs have found that general data use metrics, such as the number of downloads, are inexpensive and scalable, but they might not capture the more intangible value of DQA to the value of an activity that uses their data (Parr et al., 2019).

The significance or usefulness of a dataset to RDRs' stakeholders is a key factor in determining how RDRs prioritize their DQA efforts. The perceived value of a dataset could be influenced by various elements, such as its size, the number of variables, and usage frequency (Kindling & Strecker, 2022; Lacagnina et al., 2022; Stvilia & Lee, 2024). For example, small, specific study datasets might have a limited scope for reuse. In contrast, large datasets representing whole populations or sectors could hold more significant value due to their potential for broader application in new research. Hence, RDRs may subject those datasets to extensive processing, documentation, and enhancement efforts (Stvilia & Lee, 2024). The existing quality of datasets also guides DQA priorities, with more attention often given to popular datasets with identifiable quality issues. The level of curation and DQA applied may vary depending on the dataset's documentation quality, its expected popularity, its initial quality state, and DQA resources available (Kindling & Strecker, 2022; Stvilia & Lee, 2024; Zhou et al., 2016).

DQA priorities may also be influenced by the dataset provider's specific needs, such as deadlines for events utilizing the dataset, the provider's engagement level in the dataset curation process, and the resources available for investing in quality assurance from the provider or their funders. Repositories aim to accommodate researchers' timelines and needs, with the time dedicated to DQA potentially varying based on the curators' familiarity with the file format and the research domain (Stvilia & Lee, 2024).

Repositories' data collection policies, along with specific research community guidelines, help determine which datasets are curated and to what extent (Zhou et al., 2016). Furthermore, motivations and specific needs drive DQA activities, with the intensity of these motivations affecting the likelihood of their completion. Some repositories incentivized researchers by offering small grants to document and share datasets, encouraging engagement in DQA and helping cover some of the associated costs. Collaborations with publishers to credit researchers for shared datasets through data papers were another strategy RDRs used to enhance the researchers' motivations to engage in DQA (Stvilia & Lee, 2024).

5.7 Skills

In the evolving research data management ecosystem, shaped by new governmental regulations and the rise of big data and emerging data formats, there is a growing need to reassess the skills required for effective research

DQA. Our analysis found that DQA skills can be categorized into the following categories: knowledge of data management, technical abilities, research insight, soft skills, and specific domain expertise (see Figure 3).

Key data management skills include understanding data organization, quality assurance principles, tools usage, data preservation, handling large data volumes, and knowledge of data quality and metadata standards. The ability to meticulously identify and address errors and contradictions in data was deemed essential (Corrado, 2019; Kulmukhametov et al., 2021; Lacagnina et al., 2022; Peer et al., 2014).

Technical skills such as the ability to manage and assess large datasets, supported by proficiency in programming languages like R and Python, enable curators to automate some of the DQA tasks (Arkhangelskiy et al., 2020; Kulmukhametov et al., 2021; Lee & Stvilia, 2017; Peer et al., 2014).

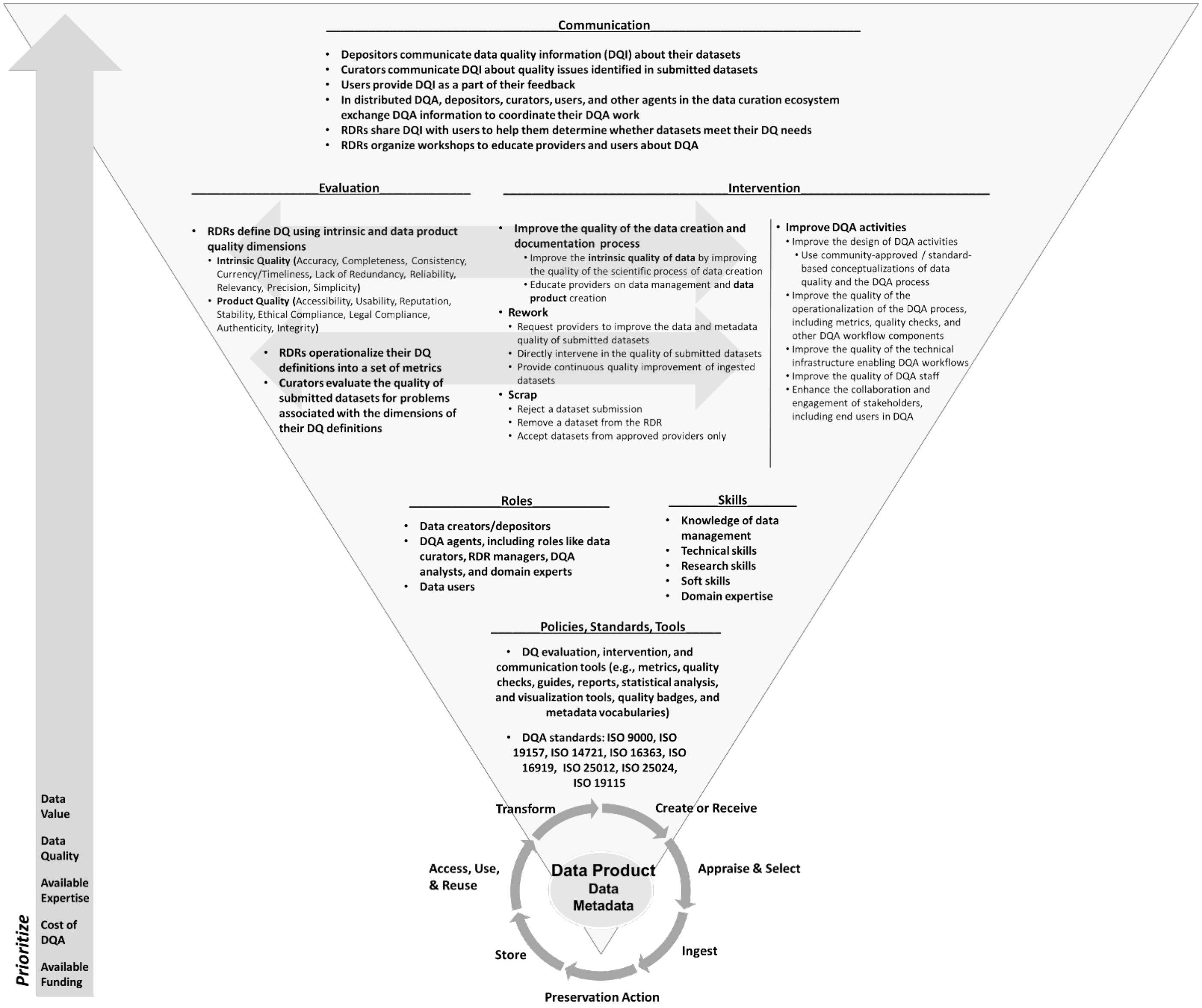
Knowledge of domain-specific data formats and persistent identifier systems is crucial for managing various data types. A background in specific scientific disciplines aids in posing pertinent questions and grasping the contextual nuances of data when assessing or communicating with depositors about the quality of their datasets (Kindling & Strecker, 2022; Lacagnina et al., 2022; Peer et al., 2014; Samuel-Rosa et al., 2019).

Studies also emphasized soft skills (Peng et al., 2022). Effective communication is vital for constructively articulating a DQA process. Furthermore, collaboration, patience, flexibility, and the ability to adapt to diverse research cultures are important. Curators must navigate different research traditions, ensuring that DQA tasks are pursued diligently. Leadership and management skills are necessary to maintain a smooth DQA workflow, ensure clarity in responsibilities, and foster accountability for any changes made (Peer et al., 2014; Lee & Stvilia, 2017).

Finally, the analysis determined that research skills are essential for research DQA. This includes knowledge of statistics, which is crucial for analyzing data, drawing valid conclusions about data quality problems, and mitigating those problems, as discussed in Section 5.3. These skills are not confined to any single domain. They are universally applicable across various fields of study, enhancing a data curator's ability to contribute effectively to the DQA of research data (Kindling & Strecker, 2022).

5.8 Comparison of Data Quality Assurance Model with CoreTrustSeal Trustworthy Repository Requirements and Data Stewardship Maturity Matrix

The first research question of this literature analysis sought to examine how RDRs define data quality. The analysis identified 17 dimensions used by RDRs when defining or referring to data quality (see Figure 3, Table 2,3). As predicted by this study's theoretical framework, RDRs referenced not just data quality dimensions when reporting on data quality but also components of system quality such as access, usability, and ethical and legal compliance (DeLone & McLean, 2003; Mason, 1986). We did not identify the use of service quality characteristics such as staff expertise, reliability, responsiveness, and empathy in the quality definitions (DeLone & McLean, 2003). However, our analysis revealed that some studies referenced staff expertise as a target of DQA intervention activities. Available staff expertise was also included as one of the facets of DQA optimization strategies (see Figure 3). The DQA model (DQAM) synthesized from the findings of this literature analysis divides the dimensions into two categories: intrinsic quality and product level data quality dimensions (see Figure 3). It is not surprising that different RDRs and their stakeholders have varying understandings of data quality. According to the data quality literature guiding this study, researchers may see their datasets as byproducts of their research projects when they submit them to an RDR. However, the curators, managers, and users of the RDR may view the same datasets as products that need to meet certain quality standards and priorities related to data reuse, as noted by Wang et al. (1998).



Communication

- Depositors communicate data quality information (DQI) about their datasets
- Curators communicate DQI about quality issues identified in submitted datasets
- Users provide DQI as a part of their feedback
- In distributed DQA, depositors, curators, users, and other agents in the data curation ecosystem exchange DQA information to coordinate their DQA work
- RDRs share DQI with users to help them determine whether datasets meet their DQ needs
- RDRs organize workshops to educate providers and users about DQA

Evaluation

- RDRs define DQ using intrinsic and data product quality dimensions
  - **Intrinsic Quality (**Accuracy, Completeness, Consistency, Currency/Timeliness, Lack of Redundancy, Reliability, Relevancy, Precision, Simplicity**)**
  - **Product Quality (**Accessibility, Usability, Reputation, Stability, Ethical Compliance, Legal Compliance, Authenticity, Integrity**)**
- RDRs operationalize their DQ definitions into a set of metrics
- Curators evaluate the quality of submitted datasets for problems associated with the dimensions of their DQ definitions

Intervention

- Improve the quality of the data creation and documentation process
  - Improve the **intrinsic quality of data** by improving the quality of the scientific process of data creation
  - Educate providers on data management and **data product** creation
- Rework
  - Request providers to improve the data and metadata quality of submitted datasets
  - Directly intervene in the quality of submitted datasets
  - Provide continuous quality improvement of ingested datasets
- Scrap
  - Reject a dataset submission
  - Remove a dataset from the RDR
  - Accept datasets from approved providers only

- Improve DQA activities
  - Improve the design of DQA activities
    - Use community-approved / standard-based conceptualizations of data quality and the DQA process
  - Improve the quality of the operationalization of the DQA process, including metrics, quality checks, and other DQA workflow components
  - Improve the quality of the technical infrastructure enabling DQA workflows
  - Improve the quality of DQA staff
  - Enhance the collaboration and engagement of stakeholders, including end users in DQA

Roles

- Data creators/depositors
- DQA agents, including roles like data curators, RDR managers, DQA analysts, and domain experts
- Data users

Skills

- Knowledge of data management
- Technical skills
- Research skills
- Soft skills
- Domain expertise

Policies, Standards, Tools

- DQ evaluation, intervention, and communication tools (e.g., metrics, quality checks, guides, reports, statistical analysis, and visualization tools, quality badges, and metadata vocabularies)
- DQA standards: ISO 9000, ISO 19157, ISO 14721, ISO 16363, ISO 16919, ISO 25012, ISO 25024, ISO 19115

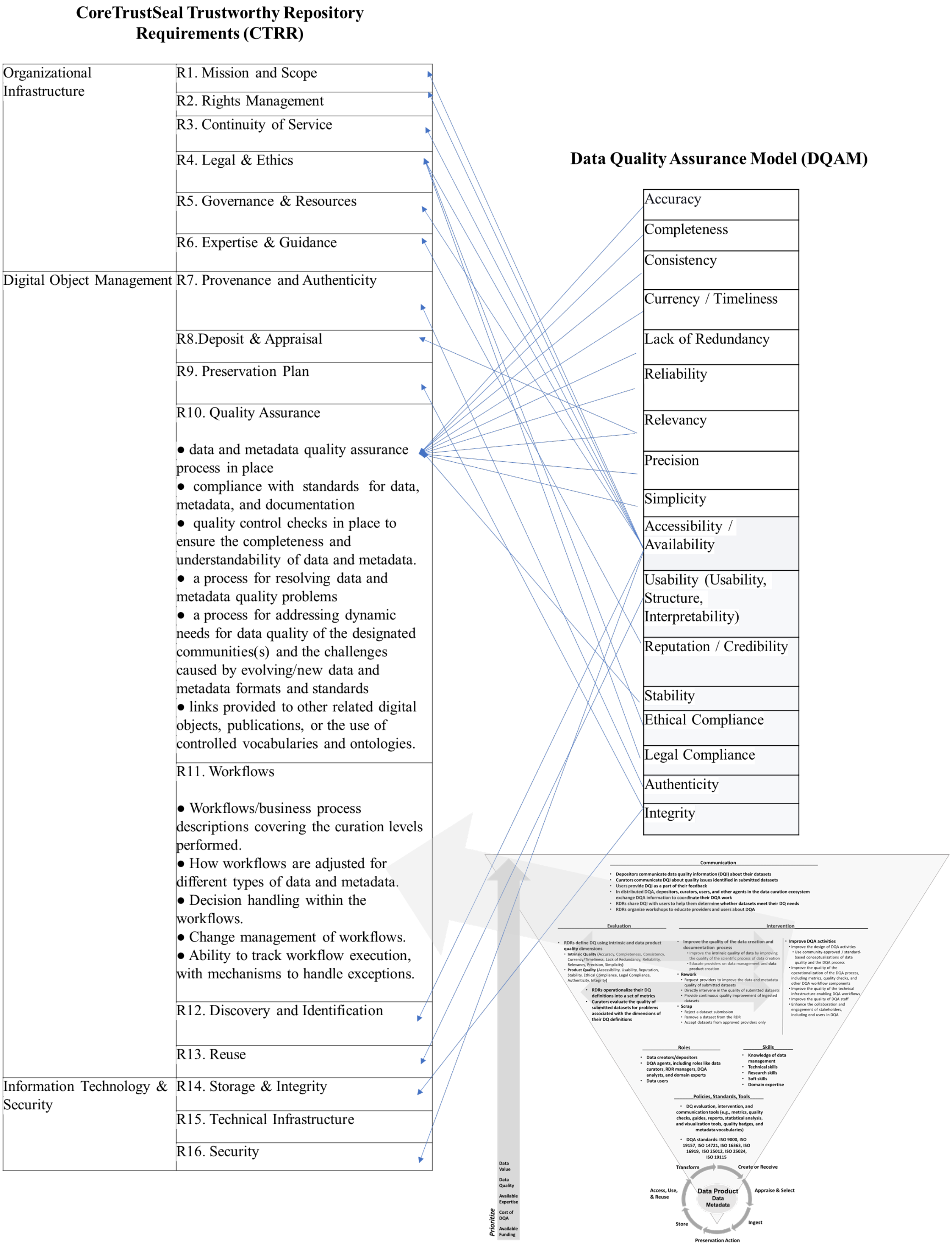
Figure 3. The synthesized model of DQA (DQAM) for RDRs.

The second research question of this study investigated how RDRs ensured data quality. The product level quality characteristics of datasets included in DQAM may affect the success of users' actions in RDRs, such as finding, accessing, interoperating, and reusing conceptualized as the FAIR framework (Wilkinson et al., 2016). The FAIR framework has been widely accepted by research data curator communities. Supporting users' FAIR actions for data products means ensuring the quality of data products' metadata and documentation and the quality of the RDR system (Peng, Gross, & Edmunds, 2022). In 2018, the European Commission's Expert Group on FAIR Data published a report aimed at facilitating the translation of FAIR principles into practical implementations (ECDRI, 2018). The report provided detailed recommendations and specific actions for various stakeholder groups. Key recommendations included defining FAIR Digital Objects and the FAIR ecosystem, understanding the social aspects that drive the system, and exploring their interactions. FAIR Digital Objects encompass not only data but also associated software and other research products and outcomes. Within the FAIR ecosystem, these digital objects interact with the ecosystem components such as data policies, persistent identifiers, data management plans, standards, and people. To create a sustainable FAIR ecosystem, it's crucial to consider social aspects like skills development, appropriate metrics, incentive structures, and continuous resource investment (ECDRI, 2018). Aligning repositories with FAIR principles has resulted in increased openness and accessibility of data, ultimately improving data quality and facilitating data reuse. Furthermore, the FAIR principles have impacted the assessment of metadata and vocabulary standards to support scientific data interoperability, further bolstering data quality in repositories (Mayernik & Liapich, 2022). Overall, integrating FAIR principles into repository practices has been crucial in advancing data quality, as well as encouraging the sharing and reuse of research data across RDRs (Aguilar Gómez, 2023).

Efforts to enhance the FAIR ecosystem are ongoing, with projects like FAIRsFAIR ("Fostering FAIR Data Practices In Europe") receiving funding from the European Union's Horizon 2020 program. FAIRsFAIR aimed to develop recommendations for FAIRness assessment within the FAIR ecosystem, including related services (e.g., metadata documentation, data transformation) and infrastructure (e.g., persistent identifiers, sustainable and trustworthy repositories). The project emphasized the importance of clear scoping and purpose when initiating data services, given the diverse goals of stakeholders (Koers et al., 2020). Additionally, the above mentioned European Union report recommended the development of a knowledge base (e.g., taxonomy, ontology, or classification scheme) to formally describe the FAIR ecosystem and its related services. DQAM can be a component of that knowledge base, providing conceptualizations of DQA concepts and relationships. The Data Product component of DQAM corresponds to the concept of the FAIR digital object (see Figure 3). One of the main recommendations of FAIR operationalization efforts is to develop assessment and certification mechanisms for FAIR Digital Objects and services. An emphasis has been placed on developing and utilizing community-based certification bodies for RDRs, such as CoreTrustSeal, and subject-specific evaluation models for data products, including data quality maturity models (ECDRI, 2018).

To further illuminate our findings for the second research question, we compared DQAM to the CoreTrustSeal Trustworthy Repository Requirements (CTRR, Figure 4). CTRR is a trustworthy data curation requirements model for RDRs (CTSC, 2022). It comprises sixteen facets of a reliable data management system, ranging from a repository providing rights management to DQA and security services. The facets are divided into three categories: organizational infrastructure, digital object management, and information technology and security.

The quality assurance requirement of CTRR is focused on assessing an RDR's ability to ensure a dataset's technical quality, including ensuring that the dataset's format, metadata schema, content, and identifiers are up to standard. However, the requirement also stipulates that enough information about a dataset's intrinsic or scientific quality must be communicated to potential users so they can make informed decisions about whether the dataset meets their scientific quality requirements (CTSC, 2022). Data product quality is not limited to using the right format and metadata quality only. Our analysis showed that it may include system quality characteristics such as accessibility, authenticity, integrity, and ethical and legal compliance. These virtues of a data product are enabled by system modules grouped in the organizational infrastructure and information technology categories of CTRR. Since CTRR is a model of information system quality, it is not surprising that the system quality support modules are separated from the data quality module (see Figure 4). The DQAM mapping to CTRR shows that the CTRR model of an RDR does have modules to support data quality assurance along the quality attributes identified by DQAM, including both the intrinsic and product level quality criteria (see Figure 4). At the same time, CTRR would benefit by adding DQAM's specification of the DQA activity structure (i.e., assessment, intervention, communication) to its quality assurance and workflow modules. The CTRR model could also be enhanced by including the list of data product quality criteria from DQAM (see Figure 3). It would make applicant repositories' self-evaluation for DQA workflows more consistent.



CoreTrustSeal Trustworthy Repository Requirements (CTRR)

Organizational Infrastructure	R1. Mission and Scope
	R2. Rights Management
	R3. Continuity of Service
	R4. Legal & Ethics
	R5. Governance & Resources
	R6. Expertise & Guidance
Digital Object Management	R7. Provenance and Authenticity
	R8.Deposit & Appraisal
	R9. Preservation Plan
	R10. Quality Assurance ● data and metadata quality assurance process in place ● compliance with standards for data, metadata, and documentation ● quality control checks in place to ensure the completeness and understandability of data and metadata. ● a process for resolving data and metadata quality problems ● a process for addressing dynamic needs for data quality of the designated communities(s) and the challenges caused by evolving/new data and metadata formats and standards ● links provided to other related digital objects, publications, or the use of controlled vocabularies and ontologies.
	R11. Workflows ● Workflows/business process descriptions covering the curation levels performed. ● How workflows are adjusted for different types of data and metadata. ● Decision handling within the workflows. ● Change management of workflows. ● Ability to track workflow execution, with mechanisms to handle exceptions.
	R12. Discovery and Identification
	R13. Reuse
Information Technology & Security	R14. Storage & Integrity
	R15. Technical Infrastructure
	R16. Security

Data Quality Assurance Model (DQAM)

Accuracy
Completeness
Consistency
Currency / Timeliness
Lack of Redundancy
Reliability
Relevancy
Precision
Simplicity
Accessibility / Availability
Usability (Usability, Structure, Interpretability)
Reputation / Credibility
Stability
Ethical Compliance
Legal Compliance
Authenticity
Integrity

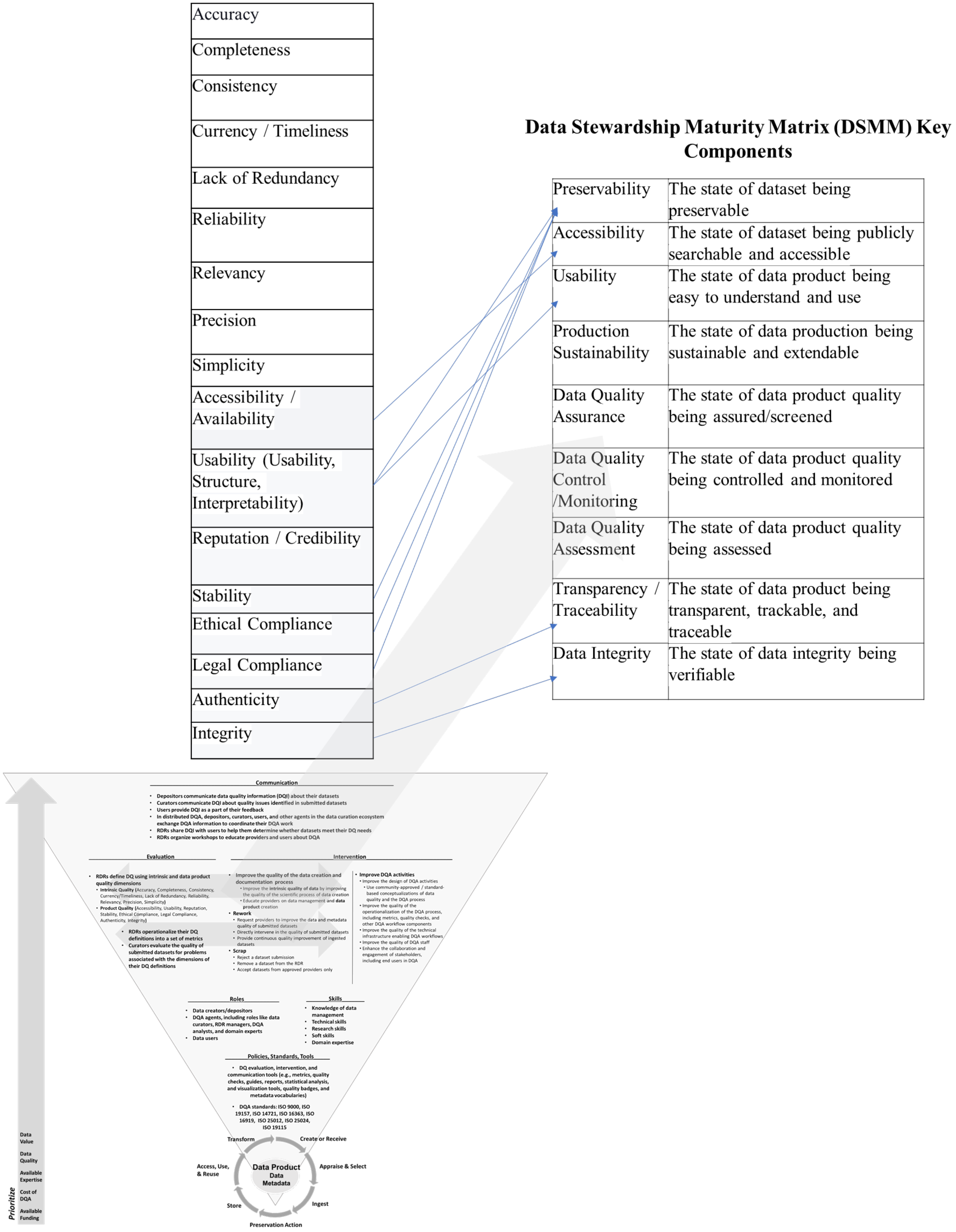
Communication
Depositors communicate data quality information (DQI) about their datasets
Curators communicate DQI about quality issues identified in submitted datasets
Users provide DQI as a part of their feedback
In distributed DQA, depositors, curators, users, and other agents in the data curation ecosystem exchange DQA information to coordinate their DQA work
RDRs share DQI with users to help them determine whether datasets meet their DQ needs
RDRs organize workshops to educate providers and users about DQA
Evaluation
RDRs define DQ using intrinsic and data product quality dimensions
Intrinsic Quality (Accuracy, Completeness, Consistency, Currency/Timeliness, Lack of Redundancy, Reliability, Relevancy, Precision, Simplicity)
Product Quality (Accessibility, Usability, Reputation, Stability, Ethical Compliance, Legal Compliance, Authenticity, Integrity)
RDRs operationalize their DQ definitions into a set of metrics
Curators evaluate the quality of submitted datasets for problems associated with the dimensions of their DQ definitions
Intervention
Improve the quality of the data creation and documentation process
Improve the intrinsic quality of data by improving the quality of the scientific process of data creation
Educate providers on data management and data product creation
Rework
Request providers to improve the data and metadata quality of submitted datasets
Directly intervene in the quality of submitted datasets
Provide continuous quality improvement of ingested datasets
Scrap
Reject a dataset submission
Remove a dataset from the RDR
Accept datasets from approved providers only
Improve DQA activities
Improve the design of DQA activities
Use community-approved / standard-based conceptualizations of data quality and the DQA process
Improve the quality of the operationalization of the DQA process, including metrics, quality checks, and other DQA workflow components
Improve the quality of the technical infrastructure enabling DQA workflows
Improve the quality of DQA staff
Enhance the collaboration and engagement of stakeholders, including end users in DQA
Roles
Data creators/depositors
DQA agents, including roles like data curators, RDR managers, DQA analysts, and domain experts
Data users
Skills
Knowledge of data management
Technical skills
Research skills
Soft skills
Domain expertise
Policies, Standards, Tools
DQ evaluation, intervention, and communication tools (e.g., metrics, quality checks, guides, reports, statistical analysis, and visualization tools, quality badges, and metadata vocabularies)
DQA standards: ISO 9000, ISO 19157, ISO 14721, ISO 16363, ISO 16919, ISO 25012, ISO 25024, ISO 19115
Transform
Create or Receive
Appraise & Select
Ingest
Preservation Action
Store
Access, Use, & Reuse
Data Product
Data
Metadata
Prioritize
Data Value
Data Quality
Available Expertise
Cost of DQA
Available Funding

Figure 4. Comparison of DQAM to CTRR.

Another model we compared DQAM to is the Data Stewardship Maturity Matrix (DSMM; Peng, Gross, & Edmunds, 2022; Peng et al., 2015). DSMM is a DQA model grounded in the OAIS Reference Model. It was developed collaboratively by the National Centers for Environmental Information (NCEI) of NOAA and the Cooperative Institute for Climate and Satellites in North Carolina. DSMM evaluates an RDR's data curation practice along nine dimensions of RDR success (Figure 5; Peng et al., 2022).

The mapping of DQAM to DSMM revealed that DSMM includes all the product level quality dimensions from DQAM except reputation (see Figure 5). Furthermore, the preservability facet of DSMM is a broader concept than individual quality dimensions. Indeed, multiple dimensions of DQAM can be mapped to the preservability facet of DSMM, including the usability dimension (see Figure 5). Unlike CTRR, DSMM includes DQA process related facets such as data quality assurance, data quality control, and data quality assessment. It is important to note that data quality assurance is a broader concept and comprises the other two. The process facets are used as a practical tool for assessing a dataset's quality. As this study's findings show, the dataset's quality can be assured at different levels, from basic quality assessment to continuous quality maintenance. Therefore, it is understandable that DSMM may need to include these three processes of DQA to indicate a data product's maturity level, even if one process encompasses the other two. At the same time, it would benefit users if the DSMM explicated the structure of DQA by including the DQA process model and the intrinsic quality dimensions from DQAM. That way, users could apply DSMM more consistently. Furthermore, the DSMM model includes a DQA sustainability facet (i.e., production sustainability; Figure 5). The facet is assessed based on the level of commitment to a dataset's DQA, ranging from individual to national and international commitment levels (Figure 5; Peng et al., 2015). As our literature review indicated, DQA priorities are shaped not just by a dataset's value to an individual researcher or a research community but also by the available funding, RDR staff expertise, and cost of DQA. If DSMM is to be applied for DQA in different types of RDRs, then the production sustainability facet of DSMM could be further expanded with the DQA prioritization facets from DQAM.

Data Quality Assurance Model (DQAM)



Data Stewardship Maturity Matrix (DSMM) Key Components

Accuracy
Completeness
Consistency
Currency / Timeliness
Lack of Redundancy
Reliability
Relevancy
Precision
Simplicity
Accessibility / Availability
Usability (Usability, Structure, Interpretability)
Reputation / Credibility
Stability
Ethical Compliance
Legal Compliance
Authenticity
Integrity

Preservability	The state of dataset being preservable
Accessibility	The state of dataset being publicly searchable and accessible
Usability	The state of data product being easy to understand and use
Production Sustainability	The state of data production being sustainable and extendable
Data Quality Assurance	The state of data product quality being assured/screened
Data Quality Control /Monitoring	The state of data product quality being controlled and monitored
Data Quality Assessment	The state of data product quality being assessed
Transparency / Traceability	The state of data product being transparent, trackable, and traceable
Data Integrity	The state of data integrity being verifiable

Communication

- Depositors communicate data quality information (DQI) about their datasets
- Curators communicate DQI about quality issues identified in submitted datasets
- Users provide DQI as a part of their feedback
- In distributed DQA, depositors, curators, users, and other agents in the data curation ecosystem exchange DQA information to coordinate their DQA work
- RDRs share DQI with users to help them determine whether datasets meet their DQ needs
- RDRs organize workshops to educate providers and users about DQA

Evaluation

- RDRs define DQ using intrinsic and data product quality dimensions
  - **Intrinsic Quality** (Accuracy, Completeness, Consistency, Currency/Timeliness, Lack of Redundancy, Reliability, Relevancy, Precision, Simplicity)
  - **Product Quality** (Accessibility, Usability, Reputation, Stability, Ethical Compliance, Legal Compliance, Authenticity, Integrity)
- RDRs operationalize their DQ definitions into a set of metrics
- Curators evaluate the quality of submitted datasets for problems associated with the dimensions of their DQ definitions

Intervention

- Improve the quality of the data creation and documentation process
  - Improve the **intrinsic quality of data** by improving the quality of the scientific process of data creation
  - Educate providers on data management and **data product** creation
- Rework
  - Request providers to improve the data and metadata quality of submitted datasets
  - Directly intervene in the quality of submitted datasets
  - Provide continuous quality improvement of ingested datasets
- Scrap
  - Reject a dataset submission
  - Remove a dataset from the RDR
  - Accept datasets from approved providers only

- Improve DQA activities
  - Improve the design of DQA activities
    - Use community-approved / standard-based conceptualizations of data quality and the DQA process
  - Improve the quality of the operationalization of the DQA process, including metrics, quality checks, and other DQA workflow components
  - Improve the quality of the technical infrastructure enabling DQA workflows
  - Improve the quality of DQA staff
  - Enhance the collaboration and engagement of stakeholders, including end users in DQA

Roles

- Data creators/depositors
- DQA agents, including roles like data curators, RDR managers, DQA analysts, and domain experts
- Data users

Skills

- Knowledge of data management
- Technical skills
- Research skills
- Soft skills
- Domain expertise

Policies, Standards, Tools

- DQ evaluation, intervention, and communication tools (e.g., metrics, quality checks, guides, reports, statistical analysis, and visualization tools, quality badges, and metadata vocabularies)
- DQA standards: ISO 9000, ISO 19157, ISO 14721, ISO 16363, ISO 16919, ISO 25012, ISO 25024, ISO 19115

Figure 5. Comparison of DQAM to DSMM.

6. Conclusion

This study examined DQA practices in RDRs using systematic literature analysis guided by a theoretical framework grounded in activity theory and the data quality literature. The findings of the literature analysis were then used to synthesize a theoretical model of DQA in RDRs—DQAM. DQAM conceptualizes a DQA process structure comprising three activities: evaluation, intervention, and communication. The literature analysis identified 17 quality dimensions RDRs used to define data quality. The model classifies the dimensions into the intrinsic and product level data quality categories. The intervention activity is further divided into data product creation process improvement, data product rework and scrap, and DQA process improvement activities. In addition, DQAM defines DQA roles and skills, as well as the standards and categories of tools used in DQA work at RDRs. Finally, the paper compares and contrasts DQAM to two DQA models used in practice: CTRR and DSMM. The comparison reveals that DQAM can supplement the CTRR model by providing a specific DQA activity structure to its quality assurance and workflow modules. The analysis also showed that DSMM can be extended by adding DQAM's DQA activity structure to make DSMM application in practice more consistent. Furthermore, DSMM's sustainability facet can be further qualified using DQAM's DQA prioritization criteria.

The theoretical implications of this study include DQAM expanding the understanding of data quality by systematically categorizing quality dimensions and linking them to quality assurance activities. This model helps in conceptualizing how various facets of data quality, like intrinsic and product level dimensions, can be integrated into a broader DQA framework for RDRs. The practical implications of the study and DQAM are to facilitate a more comprehensive and systematic approach to the design and development of DQA workflows and tools that are grounded in the DQA and information systems literature. Future related research may examine the application and evaluation of DQAM to guide the design of DQA work in different domains.

7. Acknowledgement

We would like to express our gratitude to the anonymous reviewers and editors for their invaluable comments and suggestions on the previous version of this manuscript.

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. Aerts, H., Kalra, D., Sáez, C., Ramírez-Anguita, J. M., Mayer, M.-A., Garcia-Gomez, J. M., Durà-Hernández, M., Thienpont, G., & Coorevits, P. (2021). Quality of Hospital Electronic Health Record

(EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study. *JMIR Medical Informatics*, *9*(8), e27842. https://doi.org/10.2196/27842

2. Aguilar Gómez, F., & Bernal, I. (2023). FAIR EVA: Bringing institutional multidisciplinary repositories into the FAIR picture. *Scientific Data*, *10*(1), 764.
3. Araujo-Pradere, E., Weatherhead, E. C., Dandenault, P. B., Bilitza, D., Wilkinson, P., Coker, C., Akmaev, R., Beig, G., Bureŝová, D., Paxton, L. J., Hernández-Pajares, M., Liu, J.-Y., Lin, C. H., Habarulema, J. B., & Paznukhov, V. (2019). Critical issues in ionospheric data quality and implications for scientific studies. *Radio Science*, *54*(5), 440–454. https://doi.org/10.1029/2018RS006686
4. Arkhangelskiy, T., Hedeland, H., & Riaposov, A. (2020). Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data. *CLARIN Annual Conference*, 1–7. https://doi.org/10.3384/ecp1801
5. Austin, C. C., Brown, S., Fong, N., Humphrey, C., Leahey, A., & Webster, P. (2016). Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. *IASSIST Quarterly*, *39*(4), 24-24.
6. Bailey, K.D., 1994. Typologies and taxonomies: An introduction to classification techniques, (No. 102). Sage.
7. Ball, A. 2012. Review of data management lifecycle models. University of Bath, IDMRC.
8. Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*, *44*(4), 462-484.
9. 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.
10. Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers in Ecology and Evolution*, *6*. https://www.frontiersin.org/articles/10.3389/fevo.2018.00239
11. Borries, C., Gordon, A. D., & Koenig, A. (2013). Beware of Primate Life History Data: A Plea for Data Standards and a Repository. *PLOS ONE*, *8*(6), e67200. https://doi.org/10.1371/journal.pone.0067200
12. 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.
13. Chen, Q., Britto, R., Erill, I., Jeffery, C. J., Liberzon, A., Magrane, M., Onami, J., Robinson-Rechavi, M., Sponarova, J., Zobel, J., & Verspoor, K. (2020). Quality Matters: Biocuration Experts on the Impact of Duplication and Other Data Quality Issues in Biological Databases. *Genomics, Proteomics & Bioinformatics*, *18*(2), 91–103. https://doi.org/10.1016/j.gpb.2018.11.006
14. Chen, S. & Chen, B. (2020). Practices, challenges, and prospects of Big Data curation: A case study in geoscience. *International Journal of Data Curation*, *14*(1), 275-291.
15. Cho, M. K., Taylor, H., McCormick, J. B., Anderson, N., Barnard, D., Boyle, M. B., ... & Wilfond, B. S. (2015). Building a central repository for research ethics consultation data: a proposal for a standard data collection tool. *Clinical and translational science*, *8*(4), 376-387.

16. Cichy, C., & Rass, S. (2019). An Overview of Data Quality Frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

17. Colquhoun, D. A., Shanks, A. M., Kapeles, S. R., Shah, N., Saager, L., Vaughn, M. T., Buehler, K., Burns, M. L., Tremper, K. K., Freundlich, R. E., Aziz, M., Kheterpal, S., & Mathis, M. R. (2020). Considerations for Integration of Perioperative Electronic Health Records Across Institutions for Research and Quality Improvement: The Approach Taken by the Multicenter Perioperative Outcomes Group. *Anesthesia and Analgesia*, *130*(5), 1133–1146. https://doi.org/10.1213/ANE.0000000000004489

18. Cook, H. (1997). *Product management: Value, quality, cost, price, profit and organization* (p. 411). London: Chapman & Hall.

19. CoreTrustSeal Standards and Certification Board (CTSC) (2022). Core Trustworthy Data Repository Requirements 2023–2025. Zenodo. https://doi.org/10.5281/zenodo.7051011

20. Corrado, E. M. (2019). Repositories, trust, and the CoreTrustSeal. *Technical Services Quarterly*, *36*(1), 61-72.

21. Corrales, D. C., Ledezma, A., & Corrales, J. C. (2018). From Theory to Practice: A Data Quality Framework for Classification Tasks. *Symmetry (20738994)*, *10*(7), 248. https://doi.org/10.3390/sym10070248

22. Cosley, D., Frankowski, D., Kiesler, S., Terveen, L., & Riedl, J. (2005, April). How oversight improves member-maintained communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 11-20).

23. De Rosa, A. P., Esposito, F., Valsasina, P., d'Ambrosio, A., Bisecco, A., Rocca, M. A., Tommasin, S., Marzi, C., De Stefano, N., Battaglini, M., Pantano, P., Cirillo, M., Tedeschi, G., Filippi, M., Gallo, A., Altieri, M., Borgo, R., Capuano, R., Storelli, L., … the INNI Network. (2023). Resting-state functional MRI in multicenter studies on multiple sclerosis: A report on raw data quality and functional connectivity features from the Italian Neuroimaging Network Initiative. *Journal of Neurology*, *270*(2), 1047–1066. https://doi.org/10.1007/s00415-022-11479-z

24. 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), 9-30.

25. Devaraju, A., & Herterich, P. (2020). D4. 1 Draft Recommendations on Requirements for FAIR Datasets in Certified Repositories. *Zenodo*.

26. Dunning, A., De Smaele, M., & Böhmer, J. (2017). Are the FAIR data principles fair?. *International Journal of digital curation*, *12*(2), 177-195.

27. Eppler, M. (2003). Managing information quality: Increasing the value of information in knowledge-intensive products and processes. Berlin, Germany: Springer-Verlag.

28. Estiri, H., Stephens, K. A., Klann, J. G., & Murphy, S. N. (2018). Exploring completeness in clinical data research networks with DQe-c. *Journal of the American Medical Informatics Association*, *25*(1), 17–24. https://doi.org/10.1093/jamia/ocx109

29. European Commission. (2020). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions: A European strategy for data*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066

30. European Commission. (n.d.). *Data management: H2020 Online Manual*. Retrieved July 12, 2024, from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

31. European Commission, Directorate-General for Research and Innovation (ECDRI) (2018). Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data, Publications Office. https://data.europa.eu/doi/10.2777/1524

32. Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

33. Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, *75*(6), 1274-1297.

34. Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science & Technology*, *67*(6), 1404–1416. https://doi.org/10.1002/asi.23480

35. Freeman, E., Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Angel, W. E., Berry, D. I., Brohan, P., Eastman, R., Gates, L., Gloeden, W., Ji, Z., Lawrimore, J., Rayner, N. A., Rosenhagen, G., & Smith, S. R. (2017). ICOADS Release 3.0: A major update to the historical marine climate record. *International Journal of Climatology*, *37*(5), 2211–2232. https://doi.org/10.1002/joc.4775

36. García-de-León-Chocano, R., Sáez, C., Muñoz-Soler, V., & García-Gómez, J. M. (2015). Construction of quality-assured infant feeding process of care data repositories: definition and design (Part 1). *Computers in Biology and Medicine*, *67*, 95-103.

37. Gonzalez-Vidal, A., Ramallo-González, A. P., & Skarmeta, A. F. (2022). Intrinsic and extrinsic quality of data for open data repositories. *ICT Express*, *8*(3), 328–333. https://doi.org/10.1016/j.icte.2022.06.001

38. Grossberg, A. J., Mohamed, A. S. R., Elhalawani, H., Bennett, W. C., Smith, K. E., Nolan, T. S., Williams, B., Chamchod, S., Heukelom, J., Kantor, M. E., Browne, T., Hutcheson, K. A., Gunn, G. B., Garden, A. S., Morrison, W. H., Frank, S. J., Rosenthal, D. I., Freymann, J. B., & Fuller, C. D. (2018). Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Scientific Data*, *5*(1), Article 1. https://doi.org/10.1038/sdata.2018.173

39. Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., & Piattini, M. (2021). Data quality certification using ISO/IEC 25012: Industrial experiences. *Journal of Systems and Software*, *176*, 110938. https://doi.org/10.1016/j.jss.2021.110938

40. Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. *Earth System Science Data*, *10*(2), 787–804. https://doi.org/10.5194/essd-10-787-2018

41. Gururangan, S., Card, D., Dreier, S., Gade, E., Wang, L., Wang, Z., ... & Smith, N. A. (2022, December). Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2562-2580).
42. Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal, 3*(0), 209-221. https://doi.org/10.2481/dsj.3.209
43. Hall, C., & Jensen, R. E. (2022). USACE Coastal and Hydraulics Laboratory Quality Controlled, Consistent Measurement Archive. *Scientific Data, 9*(1). https://doi.org/10.1038/s41597-022-01344-z
44. Heidebrecht, C. L., Kwong, J. C., Finkelstein, M., Quan, S. D., Pereira, J. A., Quach, S., & Deeks, S. L. (2014). Electronic immunization data collection systems: Application of an evaluation framework. *BMC Medical Informatics and Decision Making*, *14*(1), 5. https://doi.org/10.1186/1472-6947-14-5
45. Hirdes, J. P., Poss, J. W., Caldarelli, H., Fries, B. E., Morris, J. N., Teare, G. F., Reidel, K., & Jutan, N. (2013). An evaluation of data quality in Canada's Continuing Care Reporting System (CCRS): Secondary analyses of Ontario data submitted between 1996 and 2011. *BMC Medical Informatics and Decision Making*, *13*(1), 27. https://doi.org/10.1186/1472-6947-13-27
46. Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. W. (2012). Prioritization of data quality dimensions and skills requirements in genome annotation work. Journal of the American Society for Information Science and Technology, 63(1), 195-207.
47. Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2018). How important are data curation activities to researchers? Gaps and opportunities for academic libraries. *Journal of Librarianship and Scholarly Communication*, *6*(1). DOI: https://doi.org/10.7710/2162-3309.2198
48. Juárez, D., Schmidt, E. E., Stahl-Toyota, S., Ückert, F., & Lablans, M. (2019). A generic method and implementation to evaluate and improve data quality in distributed research networks. *Methods of Information in Medicine*, *58*(02/03), 086–093. https://doi.org/10.1055/s-0039-1693685
49. Juran, J. M. (1992). *Juran on quality by design: the new steps for planning quality into goods and services*. Simon and Schuster.
50. Kaplan, N. E., Baker, K. S., & Karasti, H. (2021). Long live the data! Embedded data management at a long-term ecological research site. *Ecosphere, 12*(5), e03493.
51. Kapsner, L. A., Mang, J. M., Mate, S., Seuchter, S. A., Vengadeswaran, A., Bathelt, F., Deppenwiese, N., Kadioglu, D., Kraska, D., & Prokosch, H.-U. (2021). Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository. *Applied Clinical Informatics*, *12*(4), 826–835. https://doi.org/10.1055/s-0041-1733847
52. Kaptelinin, V., & Nardi, B. (2012). *Activity theory in HCI: Fundamentals and reflections*. Morgan & Claypool Publishers.
53. Kindling, M., & Strecker, D. (2022). Data Quality Assurance at Research Data Repositories. *Data Science Journal*, *21*, 18-18.

54. Koers, H., Gruenpeter, M., Herterich, P., Hooft, R., Jones, S., Parland-von Essen, J., & Staiger, C. (2020). Assessment report on "FAIRness of services.". *FAIRsFAIR.* https://zenodo. org/record/3688762.

55. Koshoffer, A., Neeser, A. E., Newman, L., & Johnston, L. R. (2018). Giving datasets context: A comparison study of institutional repositories that apply varying degrees of curation. *International Journal of Digital Curation*, *13*(1), 15-34.

56. Kulmukhametov, A., Rauber, A., & Becker, C. (2021). Improving data quality in large-scale repositories through conflict resolution. *International Journal on Digital Libraries*, *22*(4), 365–383. https://doi.org/10.1007/s00799-021-00311-0

57. Lacagnina, C., Doblas-Reyes, F., Larnicol, G., Buontempo, C., Obregón, A., Costa Surós, M., ... & Pérez-Zanón, N. (2022). Quality management framework for climate datasets. *Data Science Journal*, *21*(1).

58. Larsen, S., Hamilton, S., Lucido, J., Garner, B., & Young, D. (2016). Supporting Diverse Data Providers in the Open Water Data Initiative: Communicating Water Data Quality and Fitness of Use. *JAWRA Journal of the American Water Resources Association, 52*(4), 859–872. https://doi.org/10.1111/1752-1688.12406

59. Lavery, J. A., Lepisto, E. M., Brown, S., Rizvi, H., McCarthy, C., LeNoue-Newton, M., Yu, C., Lee, J., Guo, X., Yu, T., Rudolph, J., Sweeney, S., Park, B. H., Warner, J. L., Bedard, P. L., Riely, G., Schrag, D., & Panageas, K. S. (2022). A Scalable Quality Assurance Process for Curating Oncology Electronic Health Records: The Project GENIE Biopharma Collaborative Approach*.* *JCO Clinical Cancer Informatics*, *6*, e2100105. https://doi.org/10.1200/CCI.21.00105

60. Lee, C. P., Dourish, P., & Mark, G. (2006, November). The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 483-492).

61. Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PloS one*, *12*(3), e0173987 .

62. Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, *40*(2), 133-146.

63. Liao, J., Wang, H., Zhou, Z., Liu, Z., Jiang, L., & Yuan, F. (2021). Integration, quality assurance, and usage of global aircraft observations in CRA. *Journal of Meteorological Research*, *35*(1), 1-16.

64. Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 144. https://doi.org/10.1038/s41597-020-0486-7

65. Longman, R. J., Giambelluca, T. W., Nullet, M. A., Frazier, A. G., Kodama, K., Crausbay, S. D., Krushelnycky, P. D., Cordell, S., Clark, M. P., Newman, A. J., & Arnold, J. R. (2018). Compilation of climate data from heterogeneous networks across the Hawaiian Islands. *Scientific Data*, *5*, 180012. https://doi.org/10.1038/sdata.2018.12

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

67. Marinov, G. K., Kundaje, A., Park, P. J., & Wold, B. J. (2014). Large-scale quality analysis of published ChIP-seq data. *G3: Genes, Genomes, Genetics*, *4*(2), 209-223.

68. Mayernik, M. S. and Liapich, Y. (2022). The role of metadata and vocabulary standards in enabling scientific data interoperability: a study of earth system science data facilities. *Journal of eScience Librarianship, 12*(1). https://doi.org/10.7191/jeslib.619

69. Mason, R. (1986). Four ethical issues of the information age. *Management Information Systems Quarterly*, *10*(1), 5-12. https://doi.org/10.2307/248873

70. McFarland, J., Helmich, E., & Valentijn, E. (2013). The Astro-WISE approach to quality control for astronomical data. *Experimental Astronomy*, *35*(1/2), 79–102. https://doi.org/10.1007/s10686-012-9296-z.

71. McGrath, N., Foley, B., Hurley, C., Ryan, M., & Flynn, R. (2022). A multi-method quality improvement approach to systematically improve and promote the quality of national health and social care information. *Health Information Management Journal*, *51*(1), 50-56. https://journals.sagepub.com/doi/full/10.1177/1833358320926422

72. Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

73. Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

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

75. 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-practicesstakeholder-perspectives-on-incentives-and-disincentives.

76. National Institutes of Health (NIH). (2024). Data management and sharing policy overview.

77. National Science Foundation (NSF). (2024). *Preparing your data management and sharing plan*. https://new.nsf.gov/funding/data-management-plan

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

79. Nelson, A. (2022). Desirable characteristics of data repositories for federally funded research. *Tech. Rep.*

80. Ng, A. (2021). AI doesn't have to be too complicated or expensive for your business. *Harvard Business Review*.

81. Orr, K. (1998). Data quality and systems theory. *Communications of the ACM*, *41*(2), 66-71.

82. Owens, D., Abeysirigunawardena, D., Biffard, B., Chen, Y., Conley, P., Jenkyns, R., Kerschtien, S., Lavallee, T., MacArthur, M., Mousseau, J., Old, K., Paulson, M., Pirenne, B., Scherwath, M., & Thorne, M. (2022). The Oceans 2.0/3.0 Data Management and Archival System. *Frontiers in Marine Science*, *9*. https://www.frontiersin.org/articles/10.3389/fmars.2022.806452

83. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71.

84. Parr, C., Gries, C., O'Brien, M., Downs, R. R., Duerr, R., Koskela, R., ... & Stall, S. (2019). A discussion of value metrics for data repositories in earth and environmental sciences. *Data Science Journal*, *18*, 58-58.

85. Peer, L., Green, A., & Stephenson, E. (2014). Committing to Data Quality Review. *International Journal of Digital Curation*, *9*(1), Article 1. https://doi.org/10.2218/ijdc.v9i1.317

86. Peng, G., Gross, W. S., & Edmunds, R. (2022). Crosswalks among stewardship maturity assessment approaches promoting trustworthy FAIR data and repositories. *Scientific Data*, *9*(1), 576. https://doi.org/10.1038/s41597-022-01683-x

87. Peng, G., Lacagnina, C., Downs, R.R., Ganske, A., Ramapriyan, H.K., Ivánová, I., Wyborn, L., Jones, D., Bastin, L., Shie, C.-L. and Moroni, D.F. (2022). Global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. *Data Science Journal*, *21*(8), 20. DOI: http://doi.org/10.5334/dsj-2022-008

88. Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S. (2015). A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, *13*, 231-253. https://www.jstage.jst.go.jp/article/dsj/13/0/13_14-049/_article/-char/ja/

89. Petter, S., DeLone, W., & McLean, E. R. (2013). Information systems success: The quest for the independent variables. *Journal of management information systems*, *29*(4), 7-62.

90. Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide. Blackwell Publishing*. https://doi.org/10.1002/9780470754887

91. Rajan, N. S., Gouripeddi, R., Mo, P., Madsen, R. K., & Facelli, J. C. (2019). Towards a content agnostic computable knowledge repository for data quality assessment. *Computer Methods and Programs in Biomedicine*, *177*, 193–201. https://doi.org/10.1016/j.cmpb.2019.05.017

92. Randles, A., Junior, A.C., & O'Sullivan, D. (2020). A framework for assessing and refining the quality of R2RML mappings. *iiWAS: Proceedings of the 22nd International Conference on Information Integration and Web-based Application & Services*, 347–351, doi: 10.1145/3428757.3429089

93. Reimer, A. P., Milinovich, A., & Madigan, E. A. (2016). Data quality assessment framework to assess electronic medical record data for use in research. *International Journal of Medical Informatics*, *90*, 40–47. https://doi.org/10.1016/j.ijmedinf.2016.03.006

94. Sáez, C., Zurriaga, O., Pérez-Panadés, J., Melchor, I., Robles, M., & García-Gómez, J. M. (2016). Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. *Journal of the American Medical Informatics Association*, *23*(6), 1085–1095. https://doi.org/10.1093/jamia/ocw010

95. Samuel-Rosa, A., Dalmolin, R. S. D., Moura-Bueno, J. M., Teixeira, W. G., & Alba, J. M. F. (2019). Open legacy soil survey data in Brazil: Geospatial data quality and how to improve it. *Scientia Agricola*, *77*, e20170430. https://doi.org/10.1590/1678-992X-2017-0430

96. Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., ... & Jinks, C. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity*, *52*, 1893-1907.

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

98. Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schäffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *arXiv preprint arXiv:2402.13635*.

99. Shekhovtsov, V. A., & Eder, J. (2022). Metadata Quality for Biobanks. *Applied Sciences*, *12*(19). https://doi.org/10.3390/app12199578

100. Singh, H., Kaur, R., Saluja, S., Cho, S. J., Kaur, A., Pandey, A. K., Gupta, S., Das, R., Kumar, P., Palma, J., Yadav, G., & Sun, Y. (2020). Development of data dictionary for neonatal intensive care unit: Advancement towards a better critical care unit. *JAMIA Open*, *3*(1), 21–30. https://doi.org/10.1093/jamiaopen/ooz064

101. Smith, M., Lix, L. M., Azimaee, M., Enns, J. E., Orr, J., Hong, S., & Roos, L. L. (2018). Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. *Journal of the American Medical Informatics Association*, *25*(3), 224-229. https://doi.org/10.1093/jamia/ocx078

102. Stamnas, E., Lammert, A., Winkelmann, V., & Lang, U. (2016). The HD(CP)2 Data Archive for Atmospheric Measurement Data. *ISPRS International Journal of Geo-Information*, *5*(7). https://doi.org/10.3390/ijgi5070124

103. Star, S. L., & Ruhleder, K. (1994, October). Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 253-264).

104. Stodden, V. (2013). Re-use and reproducibility: Opportunities and challenges. *Open Repositories 2013*. http://or2013.net/sites/or2013.net/files/OR2013-July92013-STODDEN.pdf

105. Strauss, A., & Corbin, J. (1990). *Basics of qualitative research*. Sage publications.

106. 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), 246-263.

107. Stvilia, B. & Lee, D.J. (2024). Data quality assurance in research data repositories: A theory-guided exploration and model. Journal of Documentation, 80(4), 793-812. https://doi.org/10.1108/JD-09-2023-0177 7

108. Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American society for information science and technology*, *58*(12), 1720-1733.

109. Stvilia, B., Twidale, M., Smith, L. C., Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983-1001.

110. 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*. https://par.nsf.gov/servlets/purl/10113364

111. Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one*, *10*(8), e0134826.

112. Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., ... & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one, 15*(3), e0229003.

113. Thomer, A. K., Akmon, D., York, J. J., Tyler, A. R. B., Polasek, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 414:1-414:29. https://doi.org/10.1145/3555139

114. Tian, Q., Han, Z., Yu, P., An, J., Lu, X., & Duan, H. (2021). Application of openEHR archetypes to automate data quality rules for electronic health records: a case study. *BMC Medical Informatics and Decision Making*, *21*, 1-11. https://doi.org/ 10.1186/s12911-021-01481-2

115. Trisovic, A., Mika, K., Boyd, C., Feger, S., & Crosas, M. (2021). Repository approaches to improving the quality of shared data and code. *Data*, *6*(2), 15.

116. U.S. Congress. (2002). Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745.

117. Urbano, F., Cagnacci, F., & Euromammals Collaborative Initiative. (2021). Data management and sharing for collaborative science: Lessons learnt from the Euromammals Initiative. *Frontiers in Ecology and Evolution*, *9*. https://www.frontiersin.org/articles/10.3389/fevo.2021.727023

118. Vignolo, S. M., Diray-Arce, J., McEnaney, K., Rao, S., Shannon, C. P., Idoko, O. T., Cole, F., Darboe, A., Cessay, F., Ben-Othman, R., Consortium, E., Tebbutt, S. J., Kampmann, B., Levy, O., & Ozonoff, A. (2021). A cloud-based bioinformatic analytic infrastructure and Data Management Core for the Expanded Program on Immunization Consortium. *Journal of Clinical and Translational Science*, *5*(1), e52. https://doi.org/10.1017/cts.2020.546

119. Vihinen, M., Hancock, J. M., Maglott, D. R., Landrum, M. J., Schaafsma, G. C. P., & Taschner, P. (2016). Human Variome Project Quality Assessment Criteria for Variation Databases. *Human Mutation*, *37*(6), 549–558. https://doi.org/10.1002/humu.22976

120. Wang, R. Y., Lee, Y. W., Pipino, L. L., & Strong, D. M. (1998). Manage your information as a product. *MIT Sloan Management Review, 39*(4), 95.

121. Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.

122. Westenbrink, S., Roe, M., Oseredczuk, M., Castanheira, I., & Finglas, P. (2016). EuroFIR quality approach for managing food composition data; where are we in 2014?. *Food Chemistry*, *193*, 69–74. https://doi.org/10.1016/j.foodchem.2015.02.110

123. 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), 1-9.

124. Windhager, F., Salisu, S., & Mayr, E. (2019). Exhibiting Uncertainty: Visualizing Data Quality Indicators for Cultural Collections. *Informatics*, *6*(3), Article 3. https://doi.org/10.3390/informatics6030029

125. Yoon, A. (2014). End users' trust in data repositories: Definition and influences on trust development. Archival Science. https://doi.org/10.1007/s10502-013-9207-8

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

127. Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering data quality problems: the case of repurposed data. *Business & Information Systems Engineering*, *61*, 575-593. https://doi.org/10.1007/s12599-019-00608-0

128. Zhou, L., Divakarla, M., & Liu, X. (2016). An overview of the Joint Polar Satellite System (JPSS) science data product calibration and validation. *Remote Sens*, *8*(2), 139.