This is a preprint of an article accepted for publication in the Journal of Documentation:

Lee, D.J., Stvilia, B., Gunaydin, F. & Pang, Y., (in press, 2024). Developing a data quality assurance ontology for research data repositories. *Journal of Documentation*. https://doi.org/10.1108/JD-09-2024-0212

Developing a Data Quality Assurance Ontology for Research Data Repositories

Dong Joon Lee [a, *], Besiki Stvilia [b], Fatih Gunaydin [c], Yuanying Pang [d]

[a] Mays Business School, Texas A&M University, 4217 TAMU, College Station, TX 77843-4217, United States, djlee@tamu.edu

[b] School of Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100, United States, bstvilia@fsu.edu

[c] School of Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100, United States, fg19d@fsu.edu

[d] School of Information, Florida State University, PO Box 3062100, Tallahassee, FL 32306-2100, United States, yp22c@fsu.edu

[*] Corresponding author

Abstract

Type: Research Paper

Keywords: Data quality, Data quality assurance, Research data repositories, Research data curation, Ontologies, FAIR

Purpose

Data Quality Assurance (DQA) is essential for enabling the sharing and reuse of research data, especially given the increasing focus on data transparency, reproducibility, credibility, and validity in research. Although the literature on research data curation is vast, there remains a lack of theory guided exploration of DQA *modeling* in research data repositories (RDRs).

Design/methodology/approach

This study addresses this gap by examining 12 distinct cases of DQA-related knowledge organization tools, including four metadata vocabularies, three metadata schemas, one ontology, and four standards used to guide DQA work in RDRs.

Findings

The study analyzed the cases utilizing a theoretical framework based on activity theory and data quality literature and synthesized a model and a knowledge artifact, a DQA ontology (DQAO, Lee *et al*., 2024), that encodes a DQA theory for RDRs. The ontology includes 127 classes, 44 object properties, 7 data properties, and 18 instances. The article also uses problem scenarios to illustrate how the DQAO can be integrated into the FAIR ecosystem.

Originality

The study provides valuable insights into DQA theory and practice in RDRs and offers a DQA ontology for designing, evaluating, and integrating DQA workflows within RDRs.

1. Introduction

Ensuring the quality of research data has always been important. Recent developments in federal government policies mandating open access to research data (Marcum and Donohue, 2022; Nelson, 2022), along with the increasing adoption of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles for data products in RDRs, have underscored the critical role of data quality assurance (DQA). The new regulations not only mandate sharing literature, datasets, and associated metadata from federally funded projects but also emphasize the need for discoverability and high-quality data (NSTC, 2022). Quality can be defined as fitness for use (Juran, 1992). Data and metadata quality are critical for research reproducibility and replicability (NASEM, 2019). Research data sharing and reuse face obstacles due to data quality concerns. Data owners worry about the quality and documentation of their data, fearing potential misuse or misinterpretation by others (Stvilia *et al*., 2015). Meanwhile, users seek reliable, valid data that accurately represents the phenomena they study (Davis, 1971; Stvilia and Gibradze, 2022). There are many studies of research data curation communities, and most focus on data curation practices (e.g., Lee and Stvilia, 2017; Witt, 2012) and data management training/instructions/tools in general (e.g., Carlson *et al*., 2011; Lee and Stvilia, 2014; Xu *et al*., 2023). According to Xu's scoping review on research data management (RDM) practice in academic libraries (2022), data quality was not even a topic covered by RDM training programs. Although several general or domain-specific quality assurance standards, models (e.g., ISO 9000, and ISO 19157), and prior studies (Albertoni and Isaac, 2021; Huang *et al*., 2012; Stvilia *et al*., 2015; Wu *et al*., 2012) exist, there is a lack of studies that examine semantic modeling of DQA in RDRs using the lens of information and data quality literature.

Metadata vocabularies, ontologies, and standards encode the needs and requirements of stakeholder communities and representations of their activity workflows (Chandrasekaran *et al*., 1999). By surveying and analyzing existing data quality standards, ontologies, and metadata vocabularies, we can identify the 'state of the art' in DQA knowledge representation. This understanding informs the design of a DQA ontology tailored for research data curation. Currently, we inhabit a 'Digital First' era in which technologies both create and shape human activities and behaviors (Baskerville *et al*., 2020). Often, it is not feasible to manually assess the quality of large-scale, big datasets (Chen and Chen, 2020). RDRs increasingly use automated scripts to perform data quality evaluation and intervention actions. A theoretically sound, machine-processible DQA ontology, together with subject-specific reference sources, can serve as a reusable knowledge base for designing and supporting context-specific automated DQA workflows (Chen and Esmaeilzadeh, 2024). Furthermore, such an ontology could facilitate the standardization of DQA workflows, the interoperability, and integration of data quality information, and ultimately, the interoperability of curated datasets (Chen *et al*., 2020). However, research on the design and use of DQA knowledge representation systems in RDRs is lacking. To design a sound and complete knowledge representation system, the system should be grounded both in the relevant theory and empirical data (Bailey, 1994). There is a need for a data quality ontology that integrates insights from data quality theory and DQA conceptualizations in practice. This paper aims to fill this gap by investigating the following research question: What could be a unified, general DQA ontology that is synthesized from the existing DQA knowledge sources and tailored for use by RDRs?

2. Related Work

Numerous scholarly, community, and governmental initiatives have aimed to enhance research data curation practices, establish standards, and define technical requirements for data repositories (CCSDS, 2011; ISO, 2015; Lee and Stvilia, 2017; NSTC, 2022; Wilkinson *et al*., 2016; Witt, 2012). Despite these efforts, a critical gap remains: a lack of comprehensive studies examining the use of data quality assurance standards and knowledge organization systems in RDRs. The FAIR conceptual framework is widely used in communities of practice, and it defines sets of metadata and repository system requirements to support finding, accessing, linking, and reusing research data (Wilkinson *et al*., 2016). While these principles effectively support research data professionals, they represent high level goals, leaving room for further development (Dunning *et al*., 2017). Indeed, researchers have delved deeper into refining the FAIR principles. Koers *et al*. (2020) highlighted the framework's lack of emphasis on data services, which could enhance quality evaluation and communication among stakeholders. Additionally, a recent technical report from UK JISC underscored the need for systematic development in research data management and stewardship, ensuring efficient, trusted, and sustainable data practices (Moody and Hamilton, 2024).

Data quality remains a well-recognized challenge within the current landscape of web information systems (Loscio *et al*., 2017). As the saying goes, "garbage in, garbage out." If the data stored in an information system lacks good quality, the entire system's reliability and usefulness suffer. Data Quality Vocabulary (DQV) is an information organization tool that facilitates communication of data quality on the Web (Albertoni and Isaac, 2016; 2021). It covers various data quality concepts, such as datasets and their distribution, standards, and policies to conform to, quality measurement, measurement metric, quality dimension, quality annotation data, and quality provenance data.

ISO 9000 Quality Management Systems (ISO, 2015) is a series of international standards that help manage the quality of information systems. This standard covers different aspects of quality management, including quality planning, control, determination, improvement, assurance, and continual improvement. It also emphasizes the importance of understanding the context for given information systems or information products for their quality evaluation. ISO 19157 Geographic Information – Data Quality (ISO, 2023a) is another international standard for geographic information and data, focusing on data quality evaluation. Rogala and Wawak (2021) conducted a study that evaluates the international standards within the ISO 9000 series. Their evaluation results of the quality management standards showed that the standards are generally adopted with positive feedback. However, the study also suggested that the precise definition of quality, enabling better identification of context specific quality requirements and specific tools connected to different activities, would be helpful in better understanding and supporting the quality of information and information systems.

In addition to data quality-specific standards, models, and vocabularies, there is a growing emphasis on defining general data curation practices for digital repositories. These repositories play a vital role in storing and managing research data, enabling its reuse and enhancing its credibility (Heidorn, 2008; Witt and Cragin, 2008). Researchers recognize that well-curated data contributes significantly to the scientific community (Stvilia *et al*., 2015). The process of research data curation involves managing data throughout its lifecycle, ensuring its long-term availability and reusability (Curry *et al*., 2010). Key activities within this process include discovery, selection, verification, analysis, and archiving (Qin *et al*., 2012). By adhering to these practices, repositories enhance the value and trustworthiness of research data (Lee and Stvilia, 2017).

Several frameworks have been proposed for research data services and infrastructure. Notably, the Digital Curation Centre (DCC) in the UK has introduced the DCC Curation Lifecycle Model, which emphasizes data discovery, reuse, and flexible services within diverse data environments (Higgins, 2008). Similarly, the Reference Model for an Open Archival Information System (OAIS) helps identify data assets and curation tasks within organizations (CCSDS, 2012). Other valuable data curation related models and frameworks include DRAMBORA (Digital Repository Audit Method Based On Risk Assessment), the Data Audit Framework, and the Trustworthy Repositories Audit and Certification (TRAC). These frameworks guide data curation activities, architecture components, risk management, and policy development for data archives and repositories (CCSDS, 2011; DRAMBORA Consortium, 2008; Jones *et al*., 2008).

3. Design

This study designed an information artifact, an ontology, that encodes a DQA theory for RDRs. The study used a case study design (Yin, 2018). It was informed by a theoretical framework that combined activity theory (Kaptelinin and Nardi, 2013) with the information quality assessment framework (IQAF) developed by Stvilia *et al*. (2007). Bailey's method of constructing a knowledge representation model stipulates employing both a theory-driven conceptualization and empirical data analysis to develop an effective knowledge representation model (Bailey, 1994). We used the theoretical framework to conceptualize the initial theoretical model of the ontology, to guide the selection of cases for empirical analysis, analyze the selected cases, and interpret and integrate the findings of the analysis into the ontology's theoretical model (Walls *et al*., 1992; Yin, 2018).

Thus, following Bailey's method of knowledge representation model construction (1994), we used a two-step approach to constructing the ontology. The first step involved theorizing the ontology's core conceptual model guided by the theoretical framework. The second phase of the study comprised searching and identifying empirical cases of data quality ontologies, standards, metadata schemas, or vocabularies that included any DQA-related concept theorized in the conceptual model in the first phase. We used the analysis of 127 approved applications for CoreTrustSeal certifications, interviews with 32 curators and repository managers, and data curation-related webpages of their repository websites (109 documents) as the source for identifying those cases. The data was collected from April 2022 to February 2023. The combined dataset represented 146 unique RDRs. This analysis identified nine sources. Next, we used a "snowball" approach to identify three additional standards (ISO 25000s, ISO 8000) and ontology (PROV[1] ontology) that were referenced in the previously identified standards and metadata vocabularies. Thus, in total, we identified and selected 12 cases of standards and knowledge organization systems for the analysis.

The literature includes numerous examples of using case study design for developing theoretical artifacts (e.g., Choi *et al*., 2023; Eisenhardt and Graebner, 2007; Kshetri, 2018; Stvilia *et al*., 2007). This study identified and analyzed twelve cases, which exceeds the number suggested by the literature for theory building (Eisenhardt and Graebner, 2007). The study used an embedded case study design that accommodates the use of multiple units of analysis (Yin, 2018). The units of analysis were defined by the guiding theoretical framework and the conceptual model of the ontology constructed in the first phase. We used qualitative content analysis to analyze the collected cases (four vocabularies, three metadata schemas, one ontology, and four standards; Cresswell and Cresswell, 2018). Two authors of this paper independently analyzed the content of each case for a priori codes matching the components of the conceptual model as well as emerging data quality-related themes. They compared and discussed their coding of the case at a weekly meeting and

[1] https://www.w3.org/TR/prov-o/

resolved differences, if any. Next, they compared their consensus findings of the analysis of the case with the ontology's current iteration. They integrated the findings into the ontology, resulting in a new version. This iterative process of the ontology building was continued until all the empirical cases were exhausted.

4. Theoretical Framework

Before examining how DQA is modeled in different standards and vocabularies, the study used its theoretical framework to conceptualize the process of DQA. The study's theoretical framework included activity theory (Kaptelinin and Nardi, 2012) and the IQAF (Stvilia *et al*., 2007). RDRs are sociotechnical systems. To design a DQA process and its components, such as a process ontology, one needs to jointly attend to both the social and technical considerations of the DQA process. Activity theory provides high-level conceptualizations for the relationships among human agents, their activities, the technologies used, and the organizational and community contexts that mediate those activities (Kaptelinin and Nardi, 2012). Hence, it can serve as a powerful theoretical lens to guide the conceptualization of the DQA process' representation – its ontology. IQAF is grounded in activity theory and provides models of information quality conceptualization and operationalization. The DQA ontology aims to cater to the DQA knowledge needs of different types of RDRs, both subject-specific and domain-agnostic. Hence, it is expected to function as a boundary object (Star and Griesemer, 1989), providing a common yet adaptable conceptualization of DQA process structures. This shared high level conceptual model allows individual repositories to expand and tailor the ontology to their unique contexts, developing bespoke DQA models. Furthermore, this design philosophy aligns with sociotechnical design theory, advocating for minimal essential specifications to ensure adaptability and wide applicability (Cherns, 1987).

IQAF has proven effective in guiding the development of conceptual models across various domains. It outlines common categories of activities that rely on information, identifies the types of data and information quality problems, specifies dimensions or criteria for assessing quality, and explicates the connections between these elements. After identifying the specific activities within a given context, this framework can assist in creating a conceptual model for ensuring data quality tailored to that particular context. While activity theory provides a general framework for activity relationships, IQAF conceptualizes the structure and relationships of a DQA process. According to IQAF, a DQA process involves three key activities: data quality conceptualization, measurement, and intervention (Stvilia *et al*., 2007). An RDR needs first to define a dataset's nature as a product and how its stakeholders perceive its quality. Specifically, the RDR needs to determine what data quality means to its stakeholders using the data quality dimensions or criteria the stakeholders care about. Following that, the RDR needs to operationalize each of those high level quality criteria into measurable (i.e., valid and reliable) metrics that can be used to measure a dataset's quality. Once the RDR has evaluated the dataset's quality, it can intervene to enhance it if the assessment indicates the dataset does not meet the expectations of the RDR or its stakeholder community. In addition to this general structure of a DQA process, IQAF also defines the typologies of DQA process agents, the categories of quality criteria, and the types of data quality problems and data quality dependent activities (Stvilia *et al*., 2007). A DQA analyst can use these IQAF typologies in brainstorming about a DQA process for a specific context defined by a set of activities and theorize a context specific DQA model.

5. A Conceptual Model of the DQA Process Ontology

In this section, we use the above formulated theoretical framework and the related literature to construct a DQA model for RDRs. DQA is a critical component of a repository's data curation process. A DQA model in an RDR should conceptualize key aspects of product management: what is managed, how it is managed, why it is managed, what constitutes success, and who manages it (Wang *et al*., 1998). An RDR's research data management ecosystem is multifaceted, encompassing a variety of agents and processes. It includes individuals or organizations who collect original data or aggregate and recompile data from existing datasets. Intermediaries, such as curators and repository managers, manage data on behalf of researchers, institutions, scholarly communities, or governments. At the end of this spectrum are the end users, who consume data without contributing to its creation, processing, or augmentation (Stvilia and Lee, 2024).

The primary objective of a DQA process in RDRs is to ensure the quality of data and associated information objects (e.g., metadata, data papers, and code). Therefore, a DQA process ontology must include a conceptualization of quality. In data management, quality is generally defined as "fitness for use," highlighting the contextual nature of data quality (Juran, 1992; Strong *et al*., 1997). Users' perceptions and evaluations of quality may depend on the context of the data's use and their individual circumstances. Users may discuss data quality in conventional terms or in a context-specific manner, reflecting the context of their tasks, organization, and/or culture. Their perception and understanding of data quality can also be influenced by their individual contexts, such as their subject literacy. For instance, if a user cannot directly evaluate data quality due to a lack of subject knowledge or time, they might indirectly assess it based on the source's reputation or credibility. Hence, IQAF defines three categories of information quality dimensions to represent these different perspectives on quality: intrinsic, relational, and reputational. Intrinsic quality dimensions, such as accuracy, communicate the conventional quality of a dataset. Relational quality dimensions conceptualize context-specific meanings of data quality, such as relevance or stability. Reputational quality dimensions, such as authority, are used for indirectly evaluating a dataset's quality (Choi and Stvilia, 2015; Stvilia *et al*., 2007). Furthermore, the data quality literature identifies additional dimensions associated with evaluating data as a product, such as usability (Strong et al., 1997).

The data quality literature differentiates between intrinsic quality characteristics and product level quality dimensions like accessibility or ease of understanding (Wang and Strong, 1996). The total data quality management approach proposed by Wang *et al*. (1998) emphasizes treating and managing datasets as information or data products. This approach entails understanding data products' user requirements, monitoring data product production and lifecycle, and assigning a product manager for effective management. As with any product, ensuring the usability and understandability of a data product requires high-quality metadata and documentation. Similar to conceptual models for metadata requirements proposed in the literature (e.g., CIDOC-CRM, FRBR), the data curation and research communities have adopted a conceptual model specifying four action-based quality principles for a research data product and its associated metadata: findability, accessibility, interoperability, and reusability (FAIR; Wilkinson *et al*., 2016). Thus, the quality of a data product and its documentation should be sufficient to make the dataset findable, accessible, interoperable, and reusable.

Thus, an RDR's definition of data quality is shaped by the needs for data quality of the activities the RDR is designed to support. For example, if the purpose of an RDR is to archive datasets to comply with some legal requirements and the datasets are not intended for active use in decision-making, then the RDR's definition of data quality might not include or deemphasize relational quality dimensions such as data timeliness and accessibility. IQAF defines typologies of data quality problem sources and activity types to further facilitate the design of context-specific information quality models. By identifying the types of users' data activities, designers can use the framework to infer the set of criteria that comprise or should comprise the RDR's operational definition of data quality. In particular, IQAF identifies four types of sources of information quality problems. The first type of data quality problems can be linked to faulty mappings between a dataset and the entities it represents (Stvilia *et al*., 2007). The remaining three types of information quality problems are linked to the dynamic nature of an information ecosystem, such as changes in the context of an information product evaluation, the information product itself, and the underlying entities the information product represents, which can lead to quality problems. For example, a landslide changing a river's bed may render a recent geological survey of the river obsolete and inaccurate. IQAF also defines four types of activities that can be affected by one or more of these types of quality problems: representation-dependent, decontextualizing, stability-dependent, and provenance-dependent. For instance, a study of a riverbed can be classified as a representation-dependent activity, as its success can depend on the quality of mapping/measurement of the riverbed's properties encoded in the data the study uses.

Finally, as activity theory and IQAF stipulate, a DQA process in an RDR is mediated by tools (e.g., DQA metrics, software) and its context, including organizational, institutional, and cultural contexts. These contexts mediate the DQA process through rules, conventions, norms, and other reference bases (e.g., best practices, standards) used when evaluating and intervening in a dataset's quality. The context of the DQA process also mediates the process' objective through a division of labor among different roles played by its subjects and their motivations. IQAF defines four types of agents that may interact with an information product and affect its quality. These types are defined by their task goals or intents: user, environmental, malicious, and quality assurance agents. They use and/or modify the information product to carry out their specific tasks and meet their strategic goals. In the context of DQA in RDRs, four types of agents can be defined: providers, quality assurance agents, end users, and environmental agents. While there could be malicious agents intending to disrupt an RDR's operations or vandalize its datasets, the likelihood of such agents gaining access to the RDR's data collections is low. Most RDRs do not utilize an open data curation model similar to peer curation communities like Wikipedia. Agents who submit data to an RDR are providers. Individuals performing DQA actions on the data play the DQA agent role, which may extend beyond RDR managers or curators to include providers, contributing to ensuring the quality of their data. Environmental agents can be any environmental factor that affects the dataset's quality through changes in its underlying entities or the dataset's evaluation context (Stvilia and Gasser, 2008; Stvilia and Lee, 2024).

Thus, based on the above theorizing using the study's theoretical framework, we developed a conceptual model of a DQA process ontology for RDRs that comprised 62 high-level concepts and their relationships (see Figure 1). Next, we used this initial version of the ontology to iteratively analyze 12 cases of DQA-related ontologies, standards, and metadata vocabularies. In particular, we analyzed the content of each case for DQA-related concepts and relationships and mapped them to the extant version of the DQA process ontology (DQAO, Lee *et al.,* 2024) to iteratively validate and expand it. For a case to be eligible for inclusion in the sample, it needed to contain at least one DQA concept or relationship. Some cases were specific to particular domains (e.g., geography), and subject matter-specific components from these cases (e.g., domain-specific data quality metrics) were excluded from DQAO. Only general DQA concepts were considered to maintain a broad, high-level focus for the DQA process model and ontology.

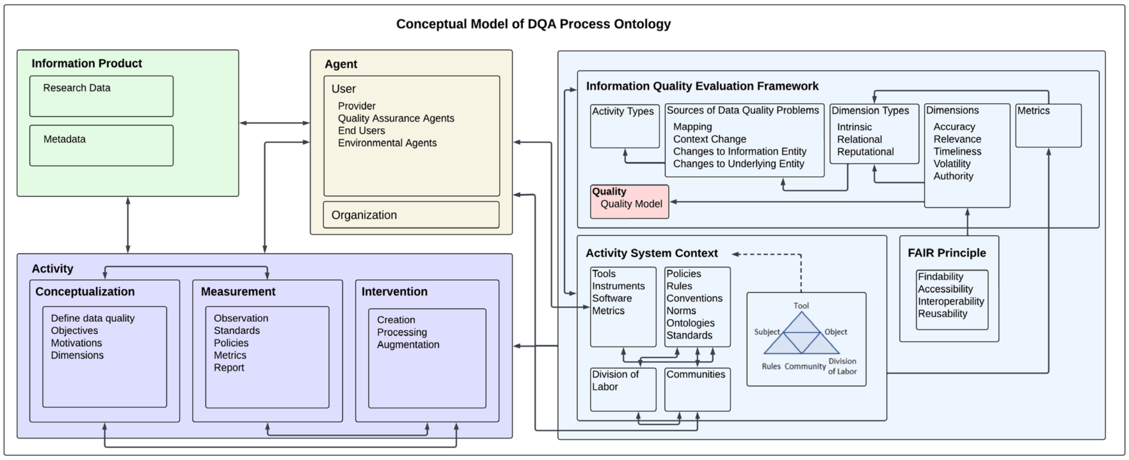

Figure 1. Conceptual model of the DQA ontology

6. Analysis and Findings

An RDR can be conceptualized as a three-module information system: data submission, management, and dissemination (CCSDS, 2012). DQA is applicable to any of these modules, and these modules shape the context of DQA activities. Our analysis of the cases identified three DQA activities referenced in the sample: data and metadata evaluation, intervention, and communication, as well as their structure and components. The following sections present the new DQA process ontology, DQAO, synthesized by mapping and aggregating relevant concepts from the cases (see Figure 2).

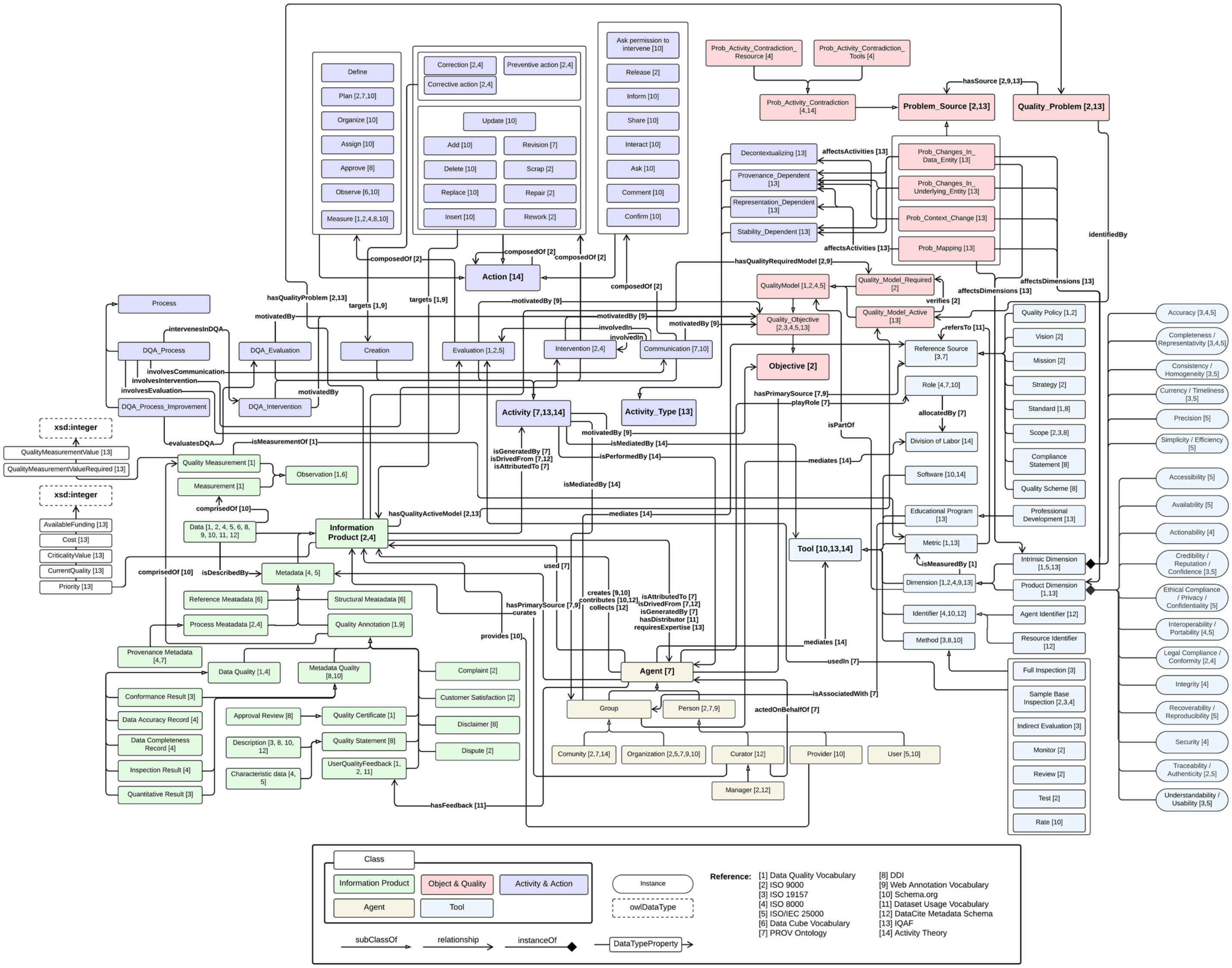

Figure 2. Diagram depicting classes and properties of DQAO

Information Product

In the context of DQA in RDRs, an information or data product comprises data and its associated metadata. Metadata encompasses several sub-classes, including process metadata, quality annotation, structural metadata, and reference metadata. Process metadata, informed by the ISO 8000, describes DQA's iterative processes (see Figure 2). Additionally, this class includes a sub-class known as provenance metadata. Provenance metadata plays a critical role in maintaining transparency in data curation activities. It facilitates the exchange of provenance information among researchers, curators, and users, even across different contexts and reference sources. To describe research data, the Data Cube Vocabulary[2] introduces two types of metadata: structural metadata and reference metadata. Structural metadata provides information on how to interpret the data, including details such as units of measurement. Reference metadata, on the other hand, offers general descriptive information about the dataset, such as creator, publisher, and location. DQAO defines quality annotation as a subclass of metadata and information product. This class can be used to store information about quality issues and evaluation results for data and metadata. It also has subclasses of user feedback, customer satisfaction, complaints, data disclaimers, quality statements, and quality certificate status. Furthermore, DQAO includes data properties that RDRs can utilize to specify their priorities for an information product. When deciding which datasets to prioritize for quality assurance, repositories may consider factors such as the dataset's value or criticality, its current quality level, their subject matter expertise, the cost of maintaining its quality, and the financial support provided by the depositor or funding organizations for quality assurance efforts (Lacagnina *et al*., 2022; Stvilia and Lee, 2024).

DQA Activities and Actions

Our conceptual model defined three DQA activities: conceptualization, measurement, and intervention (see Figure 1), The analysis of the cases, however, revealed three similar DQA activities with a slightly different structure: (1) evaluation, combined with conceptualization, (2) intervention, and (3) communication (see Figure 2). Many of the concepts and relationships from the cases are also aligned with the concepts of IQAF and activity theory, serving as quality requirements, contexts, or mediating factors. These concepts are important in shaping DQAO as tools, quality requirements, or reference sources. The activity of measurement has been renamed to evaluation, following the guidelines of ISO 19157 and DDI. ISO 19157 offers methods for quality evaluation, while DDI provides guidance on various types and processes of evaluation (see Figure 2). Furthermore, we added communication to the DQA activities. Communication, as defined by the PROV ontology, ISO 9000, and Schema.org, not only accompanies an intervention activity but is also associated with the evaluation of data quality. Effective communication among all agents is critical throughout the entire DQA process. Given its iterative nature, continuous quality improvement relies on transparent and ongoing communication between stakeholders.

Evaluation

The Evaluation activity model of DQAO consists of four parts: (1) defining quality, (2) evaluation actions, (3) outputs, and (4) tools, policies, and reference sources. Data quality conceptualization is the first DQA action that produces a conceptual model of data quality (see Figure 1). Evaluation begins by defining and understanding the quality – a high level objective that the activity strives to obtain, and motivations for ensuring data quality specific to the context of an RDR. A DQA activity's objective can be shaped by multiple motivations (Stvilia and Lee, 2024). Quality dimensions or attributes are used to define quality for a particular context. The RDR must identify the quality characteristics or dimensions its stakeholders deem important in their perceptions of data quality. DQAO adopted different quality dimensions from the ISO standards and IQAF (see Table I, Figure 2). The data quality definition is then operationalized into an operational model of data quality evaluation using associated metrics and reference sources. The ontology establishes a parent-child relationship between the quality and the quality model, as proposed by IQAF (Stvilia *et al*., 2007).

A quality problem arises when a data product fails to meet the RDR's quality requirements on one or more dimensions, as outlined by the RDR's data and metadata quality policy or model for the dataset's type. The DQA ontology incorporates quality dimensions from

[2] https://www.w3.org/TR/vocab-data-cube/

the ISO standards analyzed by the study (see Table I). Specifically, we carefully examined the data quality dimensions from the sample for adoption, mapping and merging them where necessary, and categorized them into intrinsic data quality and data product-level quality dimensions. Intrinsic DQ can be assessed by measuring internal/intrinsic characteristics of data in relation to some general or context specific reference source (e.g., scientific reference databases and models, standards, or sources of a particular culture). Thus, we combined the original intrinsic and relational categories of the conceptual model into the intrinsic category of the DQAO. On the other hand, DQAO's product level category integrates the reputational dimension with dimensions grounded in platform and organizational characteristics and requirements for making a data product usable (see Table I). The digital data product perspective on data quality management combines context-specific characteristics of a dataset and its metadata, such as interpretability and legal compliance, with system-level characteristics like security and accessibility, shaping the user's overall experience with the dataset (Wang *et al.*, 1998).

#	DQ Dimension	Dimension Category	Definition	Source
1	Accuracy	Intrinsic	The degree to which data is a correct or valid representation of an object, phenomenon, relation, process, or event.	ISO 19157; 25012; 8000
2	Completeness	Intrinsic	The extent to which data is a complete representation of another object, relation, process, or event	ISO 19157; 25012; 8000
3	Consistency	Intrinsic	The extent of consistency in using the same values or structure for representing an object, relation, process, or event	ISO 19157; 25012
4	Currency / Timeliness	Intrinsic	The age of data; The extent to which the age of data meets the requirements of its usage context	ISO 19157; 25012
5	Precision	Intrinsic	The extent of precision/data of data in representing an object, relation, process, or event.	ISO 25012
6	Simplicity	Intrinsic	The extent of cognitive or representational complexity of data	ISO 25012
7	Accessibility	Product	The extent to which a data product can be accessed within a specific context of use, especially by individuals who require assistive technology or customized settings due to disabilities.	ISO 25012
8	Availability	Product	The ability to perform as and when required.	ISO 19157; 8000; 9000; 25012
9	Actionability	Product	The ability of the identifier system to locate a data product using a unique and actionable identifier	ISO 8000
10	Credibility / Reputation / Confidence	Product	The degree of reputation or credibility of a data product	ISO 19157; 25012; 8000; 9000
11	Ethical Compliance Privacy / Confidentiality	Product	The extent to which a data product complies with the ethical principles of a particular community or culture; The extent to which a data product preserves people's privacy and does not disclose confidential information.	ISO 19157; 25012
12	Interoperability / Portability	Product	The extent of a data product being interoperable and usable in another system; The degree to which a data product has attributes that enable it to be installed, replaced, or moved from one system to another, preserving the existing quality in a specific context of use	ISO 8000; ISO 25012

13	Legal Compliance / Conformity	Product	The extent to which a data product complies with the rules, laws, and regulations of a particular community or a particular country.	ISO 25012; 9000
14	Recoverability	Product	The degree to which a data product has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use	ISO 25012
15	Integrity	Product	The degree to which the integrity of a data product is preserved	ISO 8000
16	Security	Product	The degree to which a data product is appropriately protected from damage or abuse (including unauthorized use or access).	ISO 8000
17	Traceability	Product	The ability to trace the history, application, or location of a data product.	ISO 25012
18	Understandability / Usability	Product	The degree to which a data product has attributes that enable it to be read and interpreted by users.	ISO 25012; 8000

Table I. Data quality dimensions and their dimension categories, definitions, and sources.

In addition to the action of defining quality, a data quality evaluation activity may include planning, organizing, approving, and assigning evaluation tasks to different agents (see Figure 2). The agents with assigned tasks start to observe or assess the quality of research data and their metadata. The execution and outcomes of the evaluation activity are captured in quality annotation and process metadata. In an iterative data quality assurance process involving multiple agents, DQA process metadata plays a critical role in maintaining DQA transparency and facilitating communication among different agents.

Tools and reference sources are essential resources for data curators in DQA. Several evaluation methods have been proposed as tools by standards such as ISO 9000, ISO 19157, and ISO 8000. Depending on their specific contexts, agents can select different methods to evaluate information products. Additionally, they can decide whether to inspect their products thoroughly or partially. Furthermore, they can use indirect evaluation methods utilizing data providers' and/or RDRs' reputation or authority (see Figure 2).

Intervention

When information products require improvement due to quality problems, agents such as researchers or curators undertake one or more intervention actions (e.g., correction, repair, rework, or scrap) on research data products, research data creation process, and/or associated metadata (ISO, 2015; 2016). In addition, effective intervention requires an RDR to have robust and well-designed DQA process models and associated human and technical infrastructure. Therefore, the DQAO defines two additional activity classes: DQA Evaluation and DQA Intervention (see Figure 2). These classes can be used to develop models for evaluating and intervening in a DQA process and its components. For instance, by implementing ISO 9000, ISO 8000, or ISO 25000 in their work, RDR managers and curators enhance their repositories' DQA process, and their own understanding and skills in data quality management. As with other actions, intervention actions are mediated by tools. In some RDRs, DQA can be an iterative process, in line with ISO 9000's concept of 'continual improvement.'

Communication

Communication plays a vital role in both data quality evaluation and intervention. Data curators and their teams communicate and share data quality information during the evaluation and intervention phases of a DQA process. Additionally, they may collaborate and exchange data quality information with data providers. Furthermore, data quality information can be disseminated to and from the public and end-users of datasets. Agents involved in these processes perform various communication actions, such as releasing information, informing stakeholders, interacting, asking questions and permission to intervene, and confirming details. They may use different tools and reference sources to communicate data quality information better (see Figure 2).

Agents

Different roles and responsibilities exist in the DQA process. Providers typically initiate the submission of their research data into RDRs. Managers receive these submissions and assign DQA tasks to curators. Curators then evaluate the quality of the provided data products and communicate with researchers to address any issues or problems identified during the evaluation. If necessary, curators seek permission to make changes or updates to the products. Alternatively, they may request researchers to update their submissions. Once research data is published, users also contribute to the quality of the datasets. They may provide feedback in the form of comments, complaints, or questions, which becomes part of the quality annotation information (see Figure 2).

Groups, including communities and organizations, are another type of agent in the DQA process, represented by the "Group" class in DQAO. DQA in RDRs may involve organizational infrastructure, professional data curation community, digital object management, security, and risk management (CCSDS, 2011), and certification bodies. Communities such as the CoreTrustSeal Board for Core Trustworthy Data Repositories play crucial roles in evaluating and certifying RDRs for trustworthiness (CoreTrustSeal, 2024). These certifications can serve as essential reputation cues for assessing a dataset's quality indirectly. Other communities of practice (e.g., Data Curation Network) may collaborate with and contribute to an RDR's DQA process.

Tools, Policies, and Standards as Reference Sources

Data curators use tools for evaluating the quality of data products, communicate data quality evaluation information, and carry out intervention actions such as data cleaning or editing (Stvilia and Lee, 2024). DQA activities are mediated by different tools and instruments. DQA rules, policies, standards, and best practices help conceptualize a DQA process within RDRs. These guidelines serve as reference sources for agents, enabling them to conceptualize quality assurance procedures, evaluate data and metadata quality, intervene to enhance the quality of information products, and facilitate better communication among agents (see Figure 2). ISO 9000 defines various types of reference sources, including quality policy, vision, mission, and strategy. Additionally, the DQV utilizes standards as essential reference points. ISO 19157 defines the scope of the specific data to which the data quality information applies. Furthermore, DDI provides compliance statements and quality rules, serving as valuable reference sources for DQA processes.

7. Discussion

This study synthesized a general purpose DQA ontology from the existing DQA knowledge sources for use in RDRs. The ontology includes 127 classes, 44 object properties, 7 data properties, and 18 instances. DQAO provides high-level semantic models of the core concepts and relationships of DQA. These include DQA activity models, tools, metrics, baselines, rules, regulations, roles, and division of labor. The ontology is domain-agnostic, allowing it to be adapted for specific domains and RDR contexts by extending its components with context-specific instances and values. Additionally, DQAO serves as a semantic "glue" for integrating multiple DQA processes within a single RDR or across a network of RDRs.

The core concepts of DQA are data quality and information product. As conceptualized by DQAO, an information product includes both data and associated metadata (see Figure 1, 2). For specialized, complex, and/or real-time research data, an information product may also include the software used to generate or utilize the data (Peng *et al*., 2022; Zhou *et al*., 2016). Grounded in activity theory, DQAO conceptualizes quality dimensions as conceptual tools used to mediate DQA activities. An RDR may have multiple models of quality. Some of these models may be specific to data formats, while others may represent the RDR's organizational model of data quality or the models of the research networks and communities the RDR or its stakeholders are associated with (Hodson *et al*., 2018; Stvilia and Lee, 2024). DQAO supports that. In addition, it categorizes quality dimensions into two categories: intrinsic and product level. While data quality at intrinsic dimensions is mainly ensured by data creators, ensuring a data product's quality along product level dimensions is the responsibility of both the creator of the data and the RDR's curator team. Product level dimensions such as Accessibility and Understandability are critical for making a data product produced from often siloed research projects reusable, transformable, and aggregable by external users to create value for the whole research community(s) or the public (Wang *et al*., 1998; Zhou *et al*., 2016). These can be accomplished by supplementing the data product's data with accurate, complete, and consistent metadata, including data quality evaluation metadata that contains information about the data product's intrinsic quality (Peng *et al*., 2022; Zhou *et al*., 2016). DQAO supports that by defining data and metadata as separate subclasses (see Figure 2). DQAO also supports the specification of DQA related roles and associated tasks with two classes: Division of Labor and Role. It is important to note that data product creators are not always the original creators of data. Other members of the research project can contribute to the creation of the data product by enhancing the quality of its data and metadata. Likewise, a curator who works at an RDR or is embedded in the project can shape the quality and content of the product's data and metadata by suggesting relevant models and best practices to use (Kaplan *et al*., 2021). Finally, individuals who use the data product can aggregate it into a different data product and/or enhance its metadata and data quality by providing feedback, cleaning it, and/or transforming it into a more accessible format (Stvilia and Lee, 2024).

DQA is ultimately an organizational management issue (Redman, 2017). Organizations, including universities and research labs, invest in the quality assurance of their data, including research data, to meet the needs of and generate value for their stakeholders and comply with various government mandates and regulations. Hence, the quality expectations and requirements for a data product are shaped by the needs and motivations of those stakeholders and the organization or community's contexts (Stvilia and Lee, 2024). Tensions and misalignment within those components and relationships can lead to activity problems, including DQA activity challenges and contradictions (Kaptelinin and Nardi, 2012). Many studies have discussed how various social factors, such as policies, standards, organizational differences, institutional variations, political pressure, and technical artifacts, shape an organization's data management practices (e.g., Kim and Stanton, 2016; Yang and Maxwell, 2011). DQAO captures these relationships by defining required and actual data quality models and connecting them to the organization's human and technical infrastructures (see Figure 2). In addition, DQAO defines five types of problem sources and four types of activities affected by those problems. The activity types range from decontextualizing to stability dependent activities. Some data quality problems can be complex and challenging to understand. By modeling and capturing those relationships, DQAO can be used not only for identifying the sources of data quality problems but also for predicting data quality problems and their effects on various data activities, including data transformation, aggregation, and sharing. For example, the quality of a data or information product can be evaluated differently in a context that is different from the context of its creation, leading to a data quality problem. This type of problem may affect activities that DQAO classifies as decontextualizing activities, including data sharing and aggregation activities. Finally, DQAO defines Activity Contradictions as a subclass of Problem Source. This class can be used to encode information about problems or contradictions within and between DQA activities as theorized by activity theory and the literature (Kaptelinin and Nardi, 2012; Stvilia *et al*., 2021).

To illustrate and discuss DQAO's potential utility in guiding DQA in RDRs, we can compare it to the FAIR framework. A report on the European Commission's Expert Group on FAIR Data defines the FAIR Digital Object (FDO) and FAIR ecosystem concepts. A FAIR ecosystem shapes and mediates FDO creation, evaluation, and use. FAIR Digital Objects extend beyond data to include software and other research outputs. They operate within a FAIR ecosystem and are mediated by various elements such as data policies, persistent identifiers, data management plans, standards, and personnel. According to the report, for a FAIR ecosystem to thrive, it is crucial to address social aspects like skills enhancement, defining relevant quality metrics, establishing incentive structure for FAIR Digital Object creation, sharing, and quality assurance (Hodson *et al*., 2018). A major recommendation for implementing the FAIR framework is the creation of assessment and certification processes for FDOs and services. Furthermore, a significant emphasis is placed on forming and utilizing community-driven certification entities for RDRs, such as CoreTrustSeal, and developing specialized evaluation models for data products, like data quality maturity models (Hodson *et al*., 2018). The same European Union report also recommended establishing a knowledge base or semantic layer comprising taxonomies, metadata vocabularies, ontologies, or knowledge graphs to systematically describe the FAIR ecosystem and its services. Indeed, the FAIR principles have already shaped the evaluation of the components of the semantic layers of some RDR's data management workflows, enhancing data products' interoperability, and raising their overall quality (Mayernik and Liapich, 2022).

DQAO can serve as a valuable component of FAIR ecosystems in RDRs by providing a standardization and integration layer for siloed DQA processes and data quality information. The Information Product concept of the DQAO can be connected to the FDO concept. Next, DQAO can help an RDR define quality by serving a set of quality dimensions, the typologies of quality problem types, and the types of affected activities. The FAIR framework focuses on data users. Its four principles require a FAIR ecosystem to enable users to find, access, aggregate, and reuse data products (Wilkinson *et al*., 2016). Hence, as discussed above, FAIR operationalizations are expected to prioritize data product availability and quality of data products' metadata to support these user actions. The reusability requirement, however, implies the requirement of ensuring the intrinsic data quality of FAIR data objects and documenting that through data quality metadata. Researchers need to trust the intrinsic quality of data to use it (Stvilia *et al*., 2015; Yoon and Lee, 2019). Table II provides a mapping between the FAIR principles and the quality dimensions defined in the DQAO ontology. For each FAIR principle, we identified specific quality dimensions from DQAO's Dimension class that are essential for fulfilling the principle's goals for data and/or its metadata (see Table II). Our analysis revealed that all the quality dimensions outlined in the DQAO ontology are directly applicable and relevant to the FAIR model. A data product having a problem on any of these quality dimensions can disrupt one or more FAIR goals. Additionally, we provide DQA process scenarios to demonstrate the relationships between FAIR goals and data quality dimensions, showcasing how data curators can leverage the ontology to guide their DQA activities effectively. Scenario-based task analysis proves especially valuable for identifying and exploring potential future applications of new information organization tools or technologies (Go and Carroll, 2004; Hevner *et al*., 2004). These scenarios were derived from qualitative content analysis of larger datasets presented elsewhere (Stvilia and Lee, 2024) and used in this paper to illustrate DQAO's utility to digital data curation.

FAIR Principles (Wilkinson *et al.*, 2016)		**DQA Ontology's Quality Dimensions**	**Scenarios**
Findable	F1. (meta)data are assigned a globally unique and persistent identifier	Accuracy, Actionability, Completeness, Currency, Integrity, Precision	Alex, a research scientist, recently deposited his biomedical research dataset. The RDR curator, John, must ensure that the dataset is findable. The dataset's findability is achieved by the quality of its metadata and the quality of the RDR's indexing algorithm. After consulting the DQAO, John classified the finding activity as Representational Dependent. Hence, John needs to ensure that the Intrinsic and Product Level dimensions of the metadata are the starting point for assembling a metadata quality evaluation model for the dataset. John operationalizes each quality dimension, such as Completeness using the Metric class and by identifying the set of metadata elements used by the stakeholder communities to describe and search datasets and their activity-specific priorities.
	F2. data are described with rich metadata (defined by R1 below)		
	F3. metadata clearly and explicitly include the identifier of the data it describes		
	F4. (meta)data are registered or indexed in a searchable resource		
Accessible	A1. (meta)data are retrievable by their identifier using a standardized communications protocol	Accessibility, Actionability, Integrity, Security, Ethical Compliance, Legal Compliance, Recoverability	Jerome submitted a social science research dataset that includes survey responses and interview transcripts related to the impact of remote work on employee productivity and well-being. John uses the QualityPolicy class of DQAO to ensure that the dataset's metadata specifies clear access protocols and policies that define who can access the data, under what conditions, and how. John also assesses the assembled data product on the Accessibility, and Ethical and Legal Compliance dimensions using the Metric class and assigns it an appropriate accessibility quality rating.
	A1.1 the protocol is open, free, and universally implementable		
	A1.2 the protocol allows for an authentication and authorization procedure, where necessary		
	A2. metadata are accessible, even when the data are no longer available		
Interoperable	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge	Consistency, Interoperability	Tina submitted an environmental research dataset that includes data on air quality measurements, meteorological data, and pollution sources in a large metropolitan area. To ensure broad accessibility and reuse, the semantic layer (i.e., the schema and metadata) and the content values of the dataset need to be interoperable with other datasets created by other labs and used by environmental scientists, public health researchers, and policymakers. The Evaluation activity class also helped John devise an Intervention activity aimed at enhancing dataset interoperability. John advised Tina to recode the data and associated metadata in accordance with relevant standards, formats, and vocabularies recommended by environmental science communities.
	I2. (meta)data use vocabularies that follow FAIR principles		
	I3. (meta)data include qualified references to other (meta)data		

Reusable	R1. meta(data) are richly described with a plurality of accurate and relevant attributes	Accuracy, Completeness, Consistency, Precision, Simplicity, Accessibility, Ethical Compliance, Legal Compliance, Integrity Traceability, Understandability	John received agricultural research datasets on crop yields, soil health, irrigation practices, and weather conditions across various farms over the past decade. The dataset needs to be reusable by researchers, agronomists, policymakers, and farmers to optimize agricultural practices and inform policy decisions. Reusability is a high-level, composite DQA objective. It does not have a direct match in the dimensions' taxonomy. However, reuse an action for any data dependent activity. After consulting the DQAO, John determined that the reuse goal is part of the objectives of all four activity types: Representation-Dependent, Decontextualizing, Stability-Dependent, and Provenance-Dependent. Hence, the quality of the dataset and associated metadata objects must be ensured along all categories of dimensions (i.e., Intrinsic and Product level). For example, to ensure a dataset's Integrity and Traceability, the metadata object must include the dataset's complete Provenance Metadata detailing how the data was collected, evaluated, processed, and curated, any quality control checks and interventions performed.
	R1.1. (meta)data are released with a clear and accessible data usage license		
	R1.2. (meta)data are associated with detailed provenance		
	R1.3. (meta)data meet domain-relevant community standards		

Table II. Mapping between FAIR principles and DQA process ontology's data quality dimensions.

8. Conclusion

This study devised a theory based DQA process ontology for RDRs. Guided by a theoretical framework and using a case study design, it analyzed twelve distinct DQA-related knowledge tools and used the findings to develop a DQA model for RDRs. The model was then encoded as a machine-readable ontology. The ontology, DQAO, comprises 127 classes, 44 object properties, 7 data properties, and 18 instances.

Research data curators and research data service librarians can use DQAO to develop strategies, actions, and tools for conceptualizing, assessing, and intervening in data quality. DQAO can also be part of a semantic layer to integrate data quality information across disparate research projects and repositories. Additionally, the study's findings and ontology can serve as valuable educational resources in data curation and data science training/certificate programs in library and information schools and communities of practice.

The study has a limitation. It relied on a theoretical framework, literature analysis, and empirical examples of data quality ontologies, standards, and metadata vocabularies to design DQAO and demonstrate its relevance. While scenarios were employed to illustrate the ontology's utility for RDRs, the efficiency and effectiveness of the ontology's design have yet to be validated through field studies. There remains a need to evaluate how well the DQAO addresses the knowledge needs of RDR curators and managers in designing and managing data quality assurance processes within their organizations (Hevner *et al.*, 2004).

This limitation suggests several directions for future research. One important direction is partnering with RDRs to implement the DQAO in real-world data quality assurance operations and assess its practical utility in meeting the DQA process design and management knowledge needs of curators and managers, using qualitative and mixed-method approaches.

Another promising area of research is the integration of the DQAO with existing RDR software systems. This includes evaluating its effectiveness in enhancing data curation workflow automation tools and quality monitoring dashboards. The paper has already discussed how the DQAO can be incorporated into a FAIR ecosystem. Future studies could further explore how the DQAO might be adapted and integrated into evolving data management paradigms, such as AI-ready data frameworks (NIST, 2024) and computational reproducibility systems.

9. Acknowledgements

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.

10. References

1. Albertoni, R. and Isaac, A. (2016). Data on the web best practices: data quality vocabulary. *W3C Working Group Note*. available at: https://www.w3.org/TR/vocab-dqv/ (accessed 6 September 2024).
2. Albertoni, R. and Isaac, A. (2021). Introducing the data quality vocabulary (DQV). *Semantic Web,* Vol. 12 No. 1, pp.81-97. doi: 10.3233/SW-200382
3. Bailey, K. D. (1994). Typologies and taxonomies: an introduction to classification techniques. Sage.
4. Baskerville, R., Myers, M. and Yoo, Y. (2020). Digital first: the ontological reversal and new challenges for information systems research, *MIS Quarterly,* Vol. 44 No. 2, pp.509-523. doi: 10.25300/MISQ/2020/14418
5. Carlson, J., Fosmire, M., Miller, C. and Nelson, M.S. (2011). Determining data information literacy needs: a study of students and research faculty. *Portal: Libraries and the Academy,* Vol. 11 No. 2, pp.629-657. doi: 10.1353/pla.2011.0022
6. Chandrasekaran, B., Josephson, J. R. and Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems and Their Applications,* Vol. 14 No. 1, pp.20-26. doi: 10.1109/5254.747902
7. Chen, Q., Britto, R., Erill, I., Jeffery, C. J., Liberzon, A., Magrane, M., Onami, J., Robinson-Rechavi, M., Sponarova, J., Zobel, J. and Verspoor, K. (2020). Quality Matters: biocuration Experts on the Impact of Duplication and Other Data Quality Issues in Biological Databases. *Genomics, Proteomics & Bioinformatics*, Vol. 18 No. 2, pp.91–103. doi: 10.1016/j.gpb.2018.11.006
8. Chen, S. and Chen, B. (2020). Practices, challenges, and prospects of Big Data curation: A case study in geoscience. *International Journal of Data Curation*, Vol. 14 No. 1, pp.275-291.
9. Chen, Y. and Esmaeilzadeh, P. (2024). Generative AI in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research,* Vol. 26, e53008. doi: 10.2196/53008
10. Cherns, A. (1987). Principles of sociotechnical design revisited. *Human relations*, Vol. 40 No. 3, pp.153-161. doi: 10.1177/00187267760290080
11. Choi, W., Stvilia, B. and Lee, H. S. (2023). Developing a platform-specific framework for web credibility assessment: A case of social Q&A sites. *Information Processing & Management*, Vol. 60 No. 3, 103321. doi: 10.1016/j.ipm.2023.103321
12. Choi, W. and Stvilia, B. (2015). Web credibility assessment: conceptualization, operationalization, variability, and models, *JASIST*, Vol. 66 No. 12, pp.2399-2414. doi: 10.1002/asi.23543
13. Consultative Committee for Space Data Systems (CCSDS) (2011). “Audit and certification of trustworthy digital repositories”, available at: https://public.ccsds.org/Pubs/652x0m1.pdf (accessed 6 September 2024).
14. Consultative Committee for Space Data Systems (CCSDS). (2012). “Reference model for an open archival information system (OAIS)”, available at: http://public.ccsds.org/publications/archive/650x0m2.pdf (accessed 6 September 2024).
15. CoreTrustSeal (2024), CoreTrustSeal Trustworthy Data Repositories Requirements, available at: https://www.coretrustseal.org/ (accessed 1 December 2024).

16. Creswell, J. W. and Creswell, J. D. (2018). Research design: qualitative, quantitative, and mixed methods approaches (5th ed.). Sage.

17. Curry, E., Freitas, A. and O'Riáin, S. (2010). The role of community-driven data curation for enterprises, Wood, D. (Ed*.), Linking Enterprise Data*, Springer, pp.25–47. doi: 10.1007/978-1-4419-7665-9_2

18. DataCite Metadata Working Group. (2024). DataCite metadata schema documentation for the publication and citation of research data and other research outputs. Version 4.5. DataCite, doi: 10.14454/g8e5-6293

19. Davis, M. S. (1971). That's interesting!: towards a phenomenology of sociology and a sociology of phenomenology, *Philosophy of the Social Sciences,* Vol. 1 No. 2. pp.309-344. doi: 10.1177/004839317100100211

20. DRAMBORA Consortium. (2008). "Welcome to DRAMBORA interactive: log in or register to use the toolkit", available at: http://www.repositoryaudit.eu/ (accessed 6 September 2024).

21. Dunning, A., De Smaele, M. and Böhmer, J. (2017), "Are the FAIR data principles fair?", *International Journal of Digital Curation*, Vol. 12 No. 2, pp.177-195. doi: 10.2218/ijdc.v12i2.567

22. Eisenhardt, K. and Graebner, M. (2007). Theory building from cases: opportunities and challenges. *Academy of Management Journal,* Vol. 50 No. 1, pp.25–32. doi: 10.5465/amj.2007.24160888

23. Go, K. and Carroll, J. (2004). Scenario-based task analysis. Diaper, D. & Stanton, N. (Eds.), *The handbook of task analysis for human-computer interaction*, Lawrence Erlbaum, pp.117–133.

24. Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science, *Library Trends,* Vol. 57 No. 2, pp.28099. https://muse.jhu.edu/article/262029 (accessed 6 September 2024).

25. Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, Vol. 28, No. 1 pp.75-105. doi: 10.2307/25148625

26. Higgins, S. (2008). The DCC curation lifecycle model, *International Journal of Digital Curation*, Vol. 3 No. 1, pp.134-140. available at: http://www.dcc.ac.uk/resources/curation-lifecycle-model (accessed 6 September 2024).

27. Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., ... and Wittenburg, P. (2018). Turning FAIR into reality. Final report and action plan from the European Commission expert group on FAIR data, available at: https://data.europa.eu/doi/10.2777/1524 (accessed 6 September 2024).

28. Huang, H., Stvilia, B., Jörgensen, C. and Bass, H., (2012). Prioritization of data quality dimensions and skills requirements in genome annotation work. *JASIST,* Vol. 63 No. 1, pp.195-207. doi: 10.1002/asi.21652

29. International Organization for Standardization. (2015). "Quality management systems – fundamentals and vocabulary (ISO 9000:2015)", available at: https://www.iso.org/standard/45481.html (accessed 6 September 2024).

30. International Organization for Standardization. (2016). "Data quality – part 61: data quality management: process reference model (ISO 8000-61)", available at: https://www.iso.org/standard/63086.html (accessed 6 September 2024).

31. International Organization for Standardization. (2023a). "Geographic information – data quality – part 1: General Requirement (ISO 19157-1:2023)", available at: https://www.iso.org/standard/78900.html (accessed 6 September 2024).

32. International Organization for Standardization. (2023b). "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality-in-use model (ISO/IEC 25019:2023)", available at https://www.iso.org/standard/78177.html (accessed 6 September 2024).

33. Jones, S., Ross, S. and Ruusalepp, R. (2008). The data audit framework: a toolkit to identify research assets and improve data management in research led institutions, available at: http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf (accessed 6 September 2024).

34. Juran, J. M. (1992). *Juran on quality by design*, The Free Press, New York, NY.

35. Kaplan, N. E., Baker, K. S. and Karasti, H. (2021). Long live the data! Embedded data management at a long-term ecological research site. *Ecosphere,* Vol. 12 No. 5, e03493.

36. Kaptelinin, V. and Nardi, B. (2012). *Activity theory in HCI: fundamentals and reflections*, Morgan & Claypool Publishers.

37. Kim, Y. and Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: a multilevel analysis. *JASIST,* Vol. 67 No. 4, pp.776–799. doi: 10.1002/asi.23424

38. Kshetri, N. (2018). Blockchain's roles in meeting key supply chain management objectives. *International Journal of Information Management,* Vol. 39, pp.80–89. doi: 10.1016/j.ijinfomgt.2017.12.005

39. Koers, H., Gruenpeter, M., Herterich, P., Hooft, R., Jones, S., Parland-von Essen, J. and Staiger, C. (2020), Assessment report on 'FAIRness of services', *Zenodo*, available at https://doi.org/10.5281/zenodo.3688762

40. Lee, D. J. and Stvilia, B. (2014). Developing a data identifier taxonomy, *Cataloging & Classification Quarterly,* Vol. 52 No. 3, pp.303-336, doi: 10.1080/01639374.2014.880166

41. Lee, D. J. and Stvilia, B. (2017). Practices of research data curation in institutional repositories: a qualitative view from repository staff, *PLoS One*, Vol. 12 No. 3, e0173987, doi: 10/1371/journal.pone.0173987

42. Lee, D.J., Stvilia, B., Gunaydin, F., Pang, Y., 2024. *Data Quality Assurance Ontology (DQAO).* https://github.com/stvilia/Data-Quality-Assurance-Ontology.git

43. Lacagnina, C., Doblas-Reyes, F., Larnicol, G., Buontempo, C., Obregón, A., Costa Surós, M., Bretonnière, P.A., Llabrés Brustenga, A. and Pérez-Zanón, N., 2022. Quality Management Framework for Climate Datasets. Data Science Journal, 21(1).

44. Loscio, B. F., Burle, C. and Calegari, N. (2017). "Data on the Web best practices", W3C, available at: https://www.w3.org/TR/dwbp/ (accessed 6 September 2024).

45. Marcum, C. S. and Donohue, R. (2022). *New guidance to ensure federally funded research data equitably benefits all of American*, available at: https://www.whitehouse.gov/ostp/news-updates/2022/05/26/new-guidance-to-ensure-federally-funded-research-data-equitably-benefits-all-of-america/ (accessed 6 September 2024).

46. 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,* Vol. 11 No. 2. e619. doi: 10.7191/jeslib.619

47. Moody, V. and Hamilton, M. (2024). *Mapping federation journeys for optimizing the UK digital research infrastructure,* Report for UK Research and Innovation (UKRI), available at: https://repository.jisc.ac.uk/9516/1/federated-digital-research-infrastructure-full-report.pdf (accessed 6 September 2024).

48. National Academies of Sciences, Engineering, and Medicine (NASEM). (2019). Reproducibility and replicability in science. The National Academies Press, Washington, DC, doi: 10.17226/25303

49. National Science and Technology Council (NSTC). (2022). *Desirable characteristics of data repositories for federally funded research*, available at: https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf (accessed 6 September 2024).

50. Nelson, A. (2022). *OSTP Memo: ensuring free, immediate, and equitable access to federally funded research*. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf (accessed 6 September 2024).

51. National Institute of Standards and Technology (NIST), 2024. Trustworthy and Responsible AI: Artificial Intelligence Risk Management Framework - Generative Artificial Intelligence Profile. NIST AI 600-1. Available at: https://doi.org/10.6028/NIST.AI.600-1 [Accessed 13 Dec. 2024].

52. Qin, J., Ball, A. and Greenberg, J. (2012). Functional and architectural requirements for metadata: supporting discovery and management of scientific data, *International Conference Dublin Core Metadata Application*, pp.62–71. available at:/ http://dcpapers.dublincore.org/pubs/article/view/3660 (accessed 6 September 2024).

53. 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*, Vol. 21 No. 8, 20. doi: 10.5334/dsj-2022-008

54. Redman, T. C. (2017). Seizing opportunity in data quality. *MIT Sloan Management Review*, Vol. 29. available at: https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/ (accessed 6 September 2024).

55. Rogala, P. and Wawak, S. (2021). Quality of the ISO 9000 series of standards-perceptions of quality management experts, *International Journal of Quality and Service Sciences*, Vol. 13 No. 4, pp.509-525, doi: 10.1108/IJQSS-04-2020-0065

56. Star, S. L. & Griesemer, J. R. (1989). Institutional ecology, translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, Vol. 19 No. 3. pp.387-420. available at: https://www.jstor.org/stable/285080 (accessed 6 September 2024).

57. Strong, D. M., Lee, Y. W. and Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, Vol. 40 No. 5, pp.103-110. doi: doi.org/10.1145/253769.253804

58. Stvilia, B. and Gasser, L. (2008). An activity theoretic model for information quality change. *First Monday,* Vol. 13 NO. 4. doi: 10.5210/fm.v13i4.2126

59. Stvilia, B. and Gibradze, L. (2022). Seeking and sharing datasets in an online community of data enthusiasts. *Library & Information Science Research*, Vol. 44 No. 3, 101160. doi: 10.1016/j.lisr.2022.101160

60. Stvilia, B., Gasser, L., Twidale, M. and Smith L. C. (2007). A framework for information quality Assessment. *JASIST, Vol. 58 No.* 12, pp.1720-1733. doi: 10.1002/asi.20652

61. Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M. and Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *JASIST*, Vol. 66 No. 2, pp.246-263, doi: 10.1002/asi.23177

62. Stvilia, B. and Lee, D.J. (2024). Data quality assurance in research data repositories: A theory-guided exploration and model. *Journal of Documentation*, Vol. 80 No. 4, pp.793-812. doi: 10.1108/JD-09-2023-0177

63. Stvilia, B., Lee, D. J. and Han, N. (2021). "Striking out on your own" - a study of research information management problems on university campuses. *JASIST,* Vol. 72 No. 8, pp.963-978. doi: 10.1002/asi.24464

64. Walls, J. G., Widmeyer, G. R. and El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information systems research*, Vol. 3 No. 1, pp.36-59. Available at: https://www.jstor.org/stable/23010780 (accessed 6 September 2024).

65. Wang, R. Y., Lee, Y. W., Pipino, L. L. and Strong, D. M. (1998). Manage your information as a product. *MIT Sloan Management Review*, Vol. 39 No. 4, 95. Available at: https://sloanreview.mit.edu/article/manage-your-information-as-a-product/ (accessed 6 September 2024).

66. Wang, R.Y. and Strong, D.M. (1996). Beyond accuracy: what data Quality means to data consumers. *Journal of Management Information Systems,* Vol. 12 No. 4, pp.5-33.

67. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A. ... and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship, *Scientific Data*, Vol. 3 No. 1, pp.1-9, doi: 10.1038/sdata.2016.18

68. Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service, *Journal of Library Administration,* Vol. 52 No. 2, pp.172-188. doi: 10.1080/01930826.2012.655607

69. Witt, M. and Cragin, M. (2008). "Introduction to institutional data repositories workshop", Libaries Research Publications, available at: http://docs.lib.purdue.edu/lib_research/83 (accessed 6 September 2024).

70. Xu, Z. (2022). Research data management training in academic libraries: a Scoping Review, *Journal of Librarianship and Scholarly Communication,* Vol. 10 No. 1, eP13700, doi: 10.31274/jlsc.13700

71. Xu, Z., Zhou, X. and Lee, D. J. (2023). A pilot study on social science graduate students' data core competency, *The Journal of Academic Librarianship*, Vol. 49 No. 3, pp.102715. doi: 10.1016/j.acalib.2023.102715

72. Yang, T. M. and Maxwell, T. A. (2011). Information-sharing in public organizations: A literature review of interpersonal, intra-organizational and inter-organizational success factors. *Government Information Quarterly,* Vol. 28 No. 2, pp.164–175. doi: 10.1016/j.giq.2010.06.008

73. Yin, R. K. (2018). Case study research and applications. Sage.

74. Yoon, A. and Lee, Y.Y. (2019). Factors of trust in data reuse. *Online Information Review,* Vol. 43 No. 7, pp.1245-1262. doi: 10.1108/OIR-01-2019-0014

75. Zhou, L., Divakarla, M. and Liu, X. (2016). An overview of the Joint Polar Satellite System (JPSS) science data product calibration and validation. *Remote Sens, Vol. 8 No.* 2, pp.139. doi: 10.3390/rs8020139