Data Quality Assurance at Research Data Repositories

MAXI KINDLING (ID)

DOROTHEA STRECKER (ID)

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

This paper presents findings from a survey on the status quo of data quality assurance practices at research data repositories.

The personalised online survey was conducted among repositories indexed in re3data in 2021. It covered the scope of the repository, types of data quality assessment, quality criteria, responsibilities, details of the review process, and data quality information and yielded 332 complete responses.

The results demonstrate that most repositories perform data quality assurance measures, and overall, research data repositories significantly contribute to data quality. Quality assurance at research data repositories is multifaceted and nonlinear, and although there are some common patterns, individual approaches to ensuring data quality are diverse. The survey showed that data quality assurance sets high expectations for repositories and requires a lot of resources. Several challenges were discovered: for example, the adequate recognition of the contribution of data reviewers and repositories, the path dependence of data review on review processes for text publications, and the lack of data quality information. The study could not confirm that the certification status of a repository is a clear indicator of whether a repository conducts in-depth quality assurance.

CORRESPONDING AUTHOR:

Maxi Kindling

Open-Access-Büro Berlin, Freie Universität Berlin, Germany

maxi.kindling@open-access-berlin.de

1 INTRODUCTION

Upon collection, research data are rarely fit for analysis or publication in a repository; instead, additional processing and other measures are necessary to ensure that data conform to quality expectations. In the context of research data, quality is an ubiquitous yet elusive concept. It is stated as the motivation behind data curation (Johnston et al. 2018) and the FAIR Principles (Wilkinson et al. 2016), often without describing or defining the concept in more detail.

So far, data quality assurance practices at research data repositories have not been researched systematically. Ensuring data quality is sometimes conceptualised as a part of data curation, which makes it difficult to get specific insights into data quality assurance processes. In addition, the role of research data repositories in the quality assurance process remains unclear. Given the expertise and resources repositories provide, it must be assumed that they contribute to data quality assurance. However, little is known about how they define data quality or what quality assurance measures they take; as a result, their contributions remain largely invisible.

Certificates evaluate certain aspects of research data repositories, including quality assurance measures. So far, it is unknown whether there is a relationship between the certification status of a repository and the data quality assurance measures it performs.

To address these gaps, this study aims at analysing the status quo of data quality assurance at research data repositories. It investigates the measures of data quality assurance that repositories take as well as the influence of repository certification on the prevalence of these measures.

2 LITERATURE

2.1 DATA QUALITY

The term *quality* can refer to inherent or essential characteristics of an object, but it can also be used in the context of evaluating, rating, or comparing objects (Merriam-Webster 2022). In this paper, we focus on quality in the second sense. Definitions of *quality* sometimes refer to intrinsic characteristics of an object that are universal (Wang & Strong 1996), but more often, they are context-dependent and situational. For example, widely used context-dependent definitions describe the quality of a thing based on its conformance to a set of requirements (ISO 2015) or in relation to the needs of a stakeholder intending to use it for a specific purpose; this idea is commonly referred to as *fitness for use* (Juran 1951). In the context of data, quality is often conceptualised as fitness for use, highlighting the need to take the perspective of data users into account (Wang & Strong 1996).

A stated objective in many definitions of research data quality is the reusability of data (Peer, Stephenson & Green 2014). For example, the FAIR Principles conceptualise data quality as a 'function of its ability to be accurately and appropriately found, re-used, and cited over time, by all stakeholders, both human and mechanical' (Wilkinson et al. 2016: 3).

Definitions of *data quality* are often supplemented by dimensions that outline general aspects of data quality and criteria that specify what characteristics make data fit for use in a certain context. Wang and Strong (1996) published the most widely cited framework for data quality to date, and it remains a milestone in describing quality criteria from the perspective of data users. It includes 20 quality dimensions grouped into four categories. Although the framework is applied in the context of research data, its original scope was business data, and it remains unclear whether all criteria are applicable to research data (RfII 2020; Koltay 2020). Theoretical reflections on data quality also started evolving around this time (Lee et al. 2002; Madnick et al. 2009). In the current discourse around data quality, the FAIR Principles have become central (Peng et al. 2022) as well as aspects of openness (Koltay 2020).

It is important to note that quality dimensions and criteria mentioned in the literature are not always congruent nor do they always coincide (Lee et al. 2002), highlighting the context dependence of research data quality. In addition, definitions of concepts and the use of terminology in sources also varies. Therefore, in a pragmatic approach, this study focuses on quality criteria as an expression of characteristics that make data fit for use.

The literature mentions a variety of data quality criteria: for example, accuracy, appropriate use of methods, consistency, coverage, or reuse potential. Table 1 lists quality criteria that are

QUALITY CRITERIA	DESCRIPTIONS	SOURCE (EXAMPLE)
Accessibility	Restrictions to accessing data are kept at a minimum.	Wang & Strong 1996
Accuracy	Data truly and unambiguously represent the phenomena they describe.	Cai & Zhu 2015
Appropriate and correct use of methods	Research methods are appropriately and correctly applied for data collection and processing.	RfII 2020
Appropriateness of metadata/data documentation	Metadata and data documentation appropriately describe data.	Wilkinson et al. 2016
Completeness	All necessary components are present in the data.	CASRAI 2022b
Consistency	Properties of data are homogeneous and constant.	Batini & Scannapieco 2016
Coverage	Data have the necessary temporal or spatial coverage.	Peng et al. 2022
Open data format	Data are available in an open, nonproprietary format.	OKF n.d.
Open data licence	Data are assigned an open licence.	OKF n.d.
Reuse potential	The dataset is of value for future analysis by others.	Palmer, Weber & Cragin 2011

Table 1 Examples of data quality criteria mentioned in the literature.

mentioned in the literature. The list is not exhaustive, but it gives examples of criteria used for the evaluation of data quality.

Metadata and data documentation are important factors of data quality (Austin et al. 2016; Lafia et al. 2021) because datasets require context to be useful. Therefore, Assante et al. (2016) argue that if data quality is conceptualised as fitness for use, repositories should prioritise providing sufficient metadata and documentation to enable data users to evaluate the fitness of a dataset for their use case (Assante et al. 2016). In that sense, metadata quality and data quality are strongly connected. Lawrence et al. (2011) even state that 'quality data is not possible without quality metadata' (Lawrence et al. 2011: 15).

2.2 DATA QUALITY ASSURANCE

Data quality assurance is a concept that is associated with processes and techniques for assessing, measuring, and improving quality. In the context of data publications, quality assurance is seen as the process of assessing data and taking necessary actions to make sure that they meet the requirements of the purpose for which they are used (Peer, Stephenson & Green 2014). This process spans the entire research data life cycle (RfII 2020). Following a contextual approach to data quality as fitness for use, assessing data quality needs to take into account both the dataset and the context in which it would be used (Canali 2020).

It is important to note that there is some overlap between the concepts *data quality assurance*, *data stewardship*, and *data curation*. Peng et al. (2015) describe data quality assurance as a component of data stewardship (the responsible safeguarding of data) that contributes to the usefulness of data over time. Definitions of *data quality assurance* and *data curation* also partially intersect; for example, data curation is also often tied to the idea of producing data that are fit for a specific purpose (CASRAI 2022a). Aspects of quality assurance are sometimes subsumed under data curation activities (Lafia et al. 2021). However, conceptualising data quality assurance as simply an aspect of data stewardship or data curation makes it difficult to analyse and understand specific characteristics of data quality assurance. Overall, more research on the intersection of these concepts is required.

Data quality assurance includes multiple activities, of which the assessment of data quality is one. Often, the literature divides data quality assessment into two processes: evaluating formal or technical aspects of data and evaluating aspects related to the content or scientific value of datasets (Austin et al. 2017). This idea is grounded in the multifaceted nature of the quality assurance process that may require several reviewers with different skill sets, for example, domain experts and data curators, and in the observation that repositories provide varying degrees of review, for example, by only considering technical aspects of quality (Mayernik et al. 2015). Practices and norms are sometimes adopted from the peer review of text publications, with the assumption that this will produce scientific output—data publications—of similar value and trustworthiness (Parsons & Fox 2013).

2.3 DATA QUALITY ASSURANCE AND REPOSITORIES

Repositories are important actors in ensuring data quality, but they follow different approaches (Peer, Stephenson & Green 2014). Some adopt a *self-deposit* model, where most responsibilities for quality assurance lie with the data depositor (Austin et al. 2016), whereas others take on a more active role. The level of data curation performed at repositories and, as a result, the quality of metadata varies (Koshoffer et al. 2018).

Repository features and functionalities support certain dimensions of data quality: for example, increasing the usability of data by providing comprehensive data documentation (Trisovic et al. 2021). Nevertheless, implementing data quality assurance is a complex task for research data repositories because of the continuous nature of the process, shared responsibilities involving multiple stakeholders, and the many facets of data quality (Assante et al. 2016). Quality assurance incurs costs for research data repositories but contributes to the efficiency of data management and the long-term usability and reuse value of data (Parr et al. 2019). In focus groups, researchers have stated that they consider quality assurance among the most important curation activities at research data repositories (Johnston et al. 2018).

Repositories adopt different strategies for meeting staffing needs of quality assurance. For example, discipline-specific repositories may rely on data curators with a background in the respective discipline, whereas data curators at institutional repositories without a clear disciplinary focus may collaborate with subject specialists (Lee & Stvilia 2017).

Some repositories seek formal certification to increase users' trust in their services. Certificates can take quality assurance measures into account; for example, CoreTrustSeal asks applicants to describe their approach to ensuring (meta)data quality (CTS 2019). However, repository certification cannot and does not intend to guarantee that all individual datasets published with a service are of high quality (Assante et al. 2016). So far, the relationship between repository certification and the degree of quality assurance has not yet been investigated by systematic studies.

2.4 DATA QUALITY INFORMATION

To ensure transparency and assist repository users in making informed decisions about data reuse, documentation of data quality and quality assurance measures performed at the level of individual datasets is necessary (Downs et al. 2021; Peng et al. 2022). Currently, the availability of data quality information is limited, whereas, ideally, it should be published in a machine-readable format, taking both researchers' and service providers' perspectives into account (Assante et al. 2016). This could soon change, as the development of tools checking the compliance of data publications with the FAIR Principles facilitates certain aspects of data quality assessment, therefore making quality estimations more widely available (Mangione, Candela & Castelli 2022). Some disciplines, like the earth sciences, are already taking steps towards making information on the quality of individual datasets visible (Peng et al. 2022). A potential reason for the current lack of quality information overall might be the notion of repositories achieving *pristine* datasets through cleaning data. Plantin (2019) argues that to maintain this perception, repositories may choose to make interventions performed as part of the quality assurance process invisible to the outside. Repositories should also provide information on the quality assurance processes they generally apply (Peer, Stephenson & Green 2014). Registries such as re3data record aspects of quality assurance measures that research data repositories perform (Kindling et al. 2017). As mentioned above, certification might be an indicator that a repository conducts quality assurance, but this relationship has not yet been examined in detail.

Overall, there is a lack of systematic research into whether or how repositories share quality information, both on the repository and the dataset levels.

3 METHODOLOGY

This study aims at analysing aspects of data quality assurance at research data repositories indexed in re3data. Following an exploratory approach, it covers the scope of the repository, types of data quality assessment, quality criteria, responsibilities, details of the review process, and data quality information.

3.1 SURVEY DESIGN

The study was conducted as a personalised online survey with a combination of closed- and open-ended questions. Each participant received a personalised invitation link to the survey. The questionnaire's design was based on the findings from qualitative analyses of (1) quality assurance measures described by repositories in CoreTrustSeal self-assessment documents and (2) guidelines of data journals (Kindling & Strecker 2021). These preliminary studies identified a set of quality criteria and quality assurance measures for data publications applied by repositories and data journals.

Following a pretest among 11 repository operators and experts in the field, the questionnaire was restructured, questions were worded more clearly, and ambiguous terms were defined in explanatory texts. The final questionnaire comprised 24 questions; 21 questions were mandatory, and 3 were optional. Eleven questions were only displayed if the participant had selected certain answers in previous questions. To cover the diverse approaches to quality assurance more comprehensively, participants were frequently offered to choose the option 'not applicable' and invited to describe aspects not foreseen in the survey design in free-text fields. In total, 13 questions included free-text fields, and 4 were free-text only. Supplementary File 1: Appendix: Overview of Survey Questions provides an overview of the question and response types.

In the survey tool, each personalised invitation link was paired with a variable with the re3data ID of the repository, making it possible to combine survey results with re3data metadata in the analysis.

3.2 SURVEY ADMINISTRATION

On October 13, 2020, contact information for all repositories indexed in re3data at the time (2674) was extracted from the elements *repositoryContact* and *institutionContact*. If the information was available for a repository, values from *repositoryContact* were used preferentially; otherwise, values from *institutionContact* were added. Additional contact information could be obtained from contact pages of some remaining repositories with English- or German-language websites. The list of contact information was updated after a newsletter was sent to the repositories. Where possible, alternative contact information for invalid email addresses was added. After this process, contact information was available for 1893 repositories as of January 29, 2021. Four additional repositories asked to be included after becoming aware of the survey. In total, 1897 repositories were invited. Invitations for the survey were sent out on May 18, 2021, followed by reminders on June 1 and June 7. The survey was closed on June 15, 2021.

3.3 RESPONSE

Of the 1897 repositories that were invited, 332 completed the questionnaire. For a population of 1897 at a confidence interval of 95%, the minimal sample size is 320. The sampling error with 332 responses is 4.89%. Therefore, the results of the survey can be considered robust.

As Figure 1 demonstrates, compared to all repositories indexed in re3data at the time (2674), disciplinary repositories are slightly under-represented and institutional repositories slightly over-represented in the sample. However, because all repository types are present in all combinations in the sample, the issue is not considered severe.
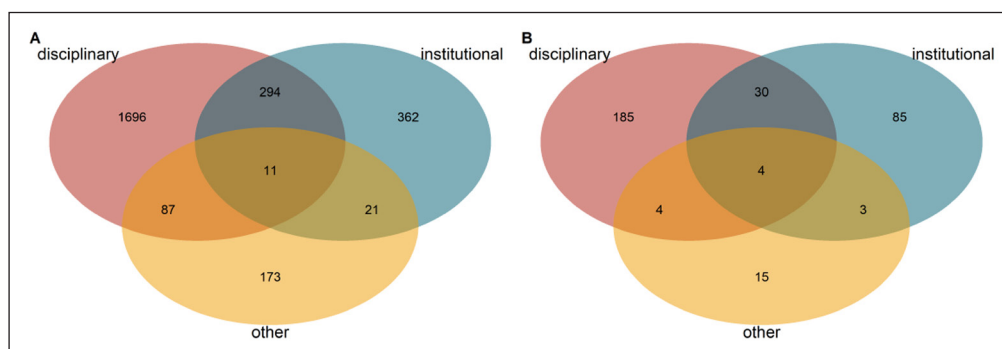

Figure 1 Types of all repositories indexed in re3data (**A**; NA: 30) and repositories included in the analysis (**B**; NA: 6).

3.4 DATA PROCESSING

Prior to data analysis, incomplete responses were removed; 332 complete responses were included in the analysis.

Participants selected the variable *Other* 328 times for 14 questions. The analysis of corresponding free-text fields revealed that in 49 cases, the content of free-text fields specifying the selected variable *Other* matched one of the predefined variables. These variables were reassigned and, where applicable, replaced *Other*.

Survey data was supplemented by information on repository certification from a re3data database dump that was generated on April 22, 2021. The variable *certification status* is derived from the element *certificate* in the re3data metadata schema (Strecker et al. 2021) and describes whether a repository has obtained any type of formal certification (53 in the sample), for example, from CoreTrustSeal. Differences across certification status were evaluated using chi-square (χ^2) tests, and effect sizes are reported using Cramér's V (V). After anonymisation, the data, codebook, and survey instrument were made openly available (Kindling, Strecker & Wang 2022).

4 RESULTS

The following section outlines the findings of the analysis.

4.1 SCOPE OF THE REPOSITORY

The repositories participating in the survey vary in scope, both in terms of the extent of the services they offer (Q01, N = 332, n = 568) and in terms of the types of data they hold (Q02, N = 332, n = 1471). Services are extended to the hosting institution (n = 193, 58.1%), other institutions or projects (n = 158, 47.6%), or to any source (n = 110, 33.1%). Some repositories aggregate metadata from other data providers (n = 86, 25.9%) On average, the repositories selected 5.4 types of data that fall within their scope. Among these data types, the most widespread are measured values (n = 146, 44%), images (n = 110, 33.1%), data from analysed sample material (n = 107, 32.2%), and databases (n = 107, 32.2%). Some participants state that the repository does not focus on a specific data type (n = 77, 23.2%).

Repositories apply different criteria to ensure a homogenous collection (Q03, N = 332, n = 805), for example, based on collection profiles or policies. Most repositories check whether data fit the scope of the repository in general (n = 237, 71.4%). Other criteria include that data passed formal assessment before deposit (n = 106, 31.9%), that data are described in a publicly accessible document (n = 93, 28%), and that data correspond to a peer-reviewed publication (n = 91, 27.4%). Some participants state that a collection policy is not applicable to the repository (n = 26, 7.8%). In the free-text field, additional criteria were listed, including technical suitability or data availability.

Repositories report offering a wide range of support services (Q04, N = 332, n = 1461). On average, repositories selected 4.8 distinct services. Most frequently, repositories offer direct, individualised support to data depositors (n = 244, 73.5%). Other common types of support services include data deposit guidelines (n = 208, 62.7%) and data format recommendations (n = 204, 61.4%). Some repositories state that support for data depositors is not applicable to the repository (n = 23, 6.9%). In the free-text field, guidelines for specific aspects of data curation (data protection, anonymisation, data management plans) were mentioned as additional types of support services.

4.2 TYPES OF DATA QUALITY ASSESSMENT

The survey distinguished between formal assessment of data (Q05, N = 332, n = 332) and data review (Q10, N = 332, n = 332). Formal assessment refers to technical, administrative, and access-related aspects of data, whereas data review refers to the process by which experts, either from the hosting institution or from other institutions, evaluate the scientific quality of datasets.

As Figure 2 A shows, the majority of participants report that formal aspects of data are assessed at the repository (n = 207, 62.3%). Others do not conduct formal assessment (n = 65, 19.6%) or formal assessment is not applicable (n = 36, 10.8%). The analysis revealed no statistically

significant relationship between the formal assessment of data and the certification status of a repository. About half (n = 171, 51.5%) of the responding repositories perform data review either for all (n = 105, 31.6%) or some (n = 66, 19.9%) datasets (see Figure 2 B). About a quarter do not conduct data review (n = 90, 27.1%), and for others, data review is not applicable (n = 52, 15.7%). The association between performing data review and certification status of the repository is statistically significant (χ^2 (3, N = 332) = 9.8, p = 0.02) but weak (V = 0.18).
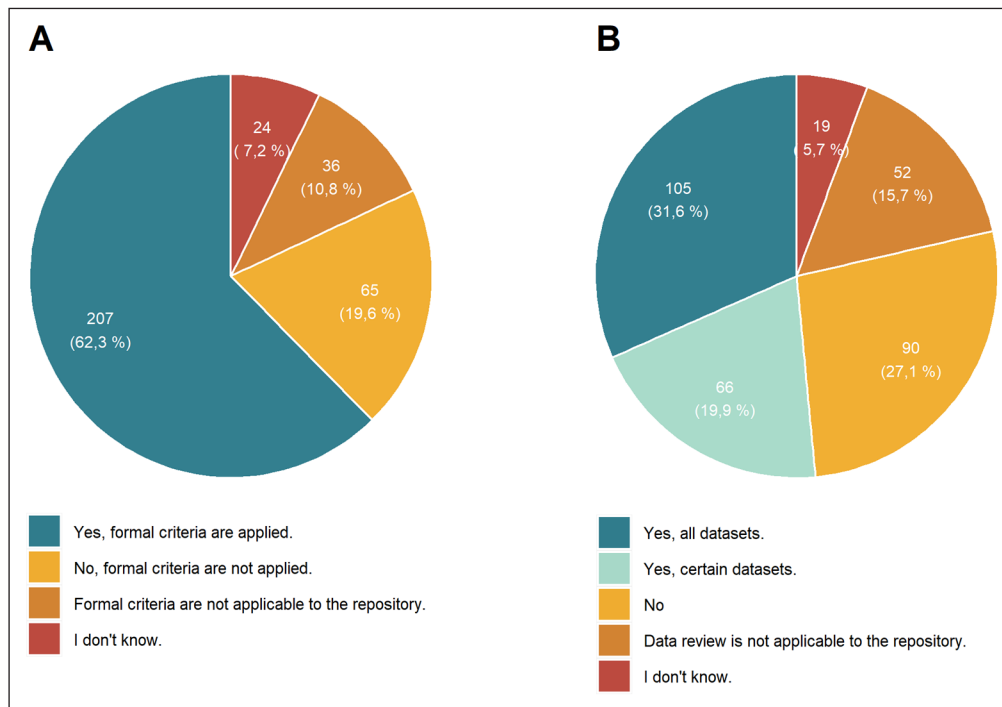
Figure 2 Question 05: Are formal criteria applied to data before publication? **(A)**; Question 10: Are data reviewed beyond the application of formal criteria? **(B)**.

Overall, 22.9% (n = 76) of the responding repositories perform neither formal assessment nor review of data, whereas 77.1% (n = 256) conduct at least one type of data quality assessment. Of these, 134 perform either formal assessment (n = 85) or review of data (n = 49), and 122 perform both, as Figure 3 shows.
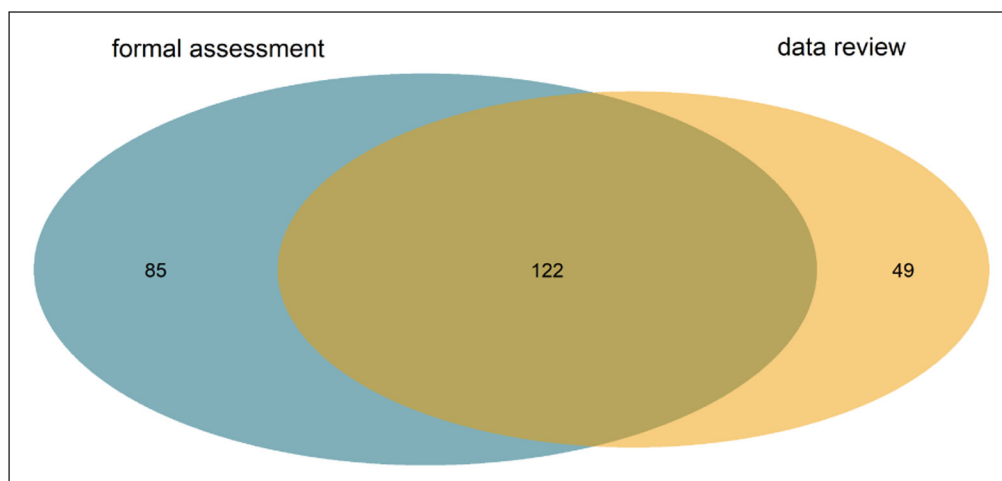

Figure 3 Types of data quality assessment performed at responding repositories.

4.3 CRITERIA FOR THE FORMAL ASSESSMENT OF DATA

Repositories that perform formal assessment of data were asked what criteria guide their process. Figure 4 shows the criteria repositories apply when assessing formal aspects of data (Q06, N = 207, n = 3519). Almost all respondents (n = 201, 97.1%) state that either the repository, the data provider, or both check for a basic description of data. Other widely applied criteria include the clarification of copyright and usage rights (n = 186, 89.9%), compliance with a metadata schema (n = 185, 89.4%), provision of provenance information (n = 184, 88.9%), and compliance with the FAIR Principles (n = 181, 87.5%). The criteria applied least frequently are that data pass statistical tests (n = 64, 30.9%) and the declaration of conflicts of interests (n = 72, 34.8%).
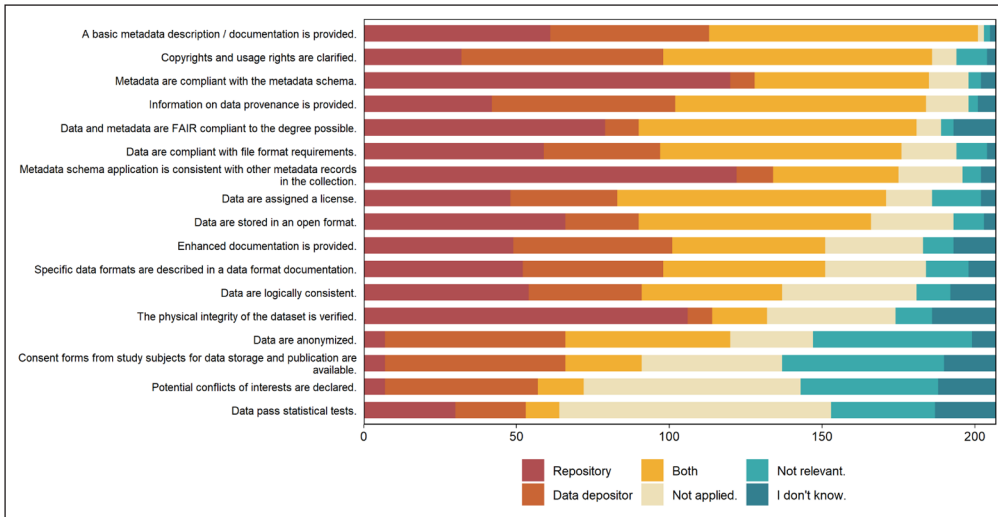
Respondents added a variety of additional formal criteria in the subsequent free-text field (Q07, N = 63, n = 63). For example, responses mentioned the adherence to community standards, automatic quality checks, and fingerprinting.

4.4 CRITERIA FOR THE REVIEW OF DATA

The repositories performing data review were asked what criteria the process was based on. Figure 5 shows relevancy ratings of criteria for the review of data (Q11, N = 171, n = 3591). Most respondents (n = 163, 95.3%) state that the overall data and documentation quality is *very relevant* or *relevant* for data review at their repository. Other criteria that were commonly rated *very relevant* or *relevant* include appropriate data documentation (n = 155, 90.6%), suitability to the scope of the repository (n = 143, 83.6%), and accuracy (n = 135, 79%). The criteria rated *very relevant* or *relevant* least frequently are the novelty (n = 41, 23.9%) and timeliness (n = 58, 33.9%) of data.
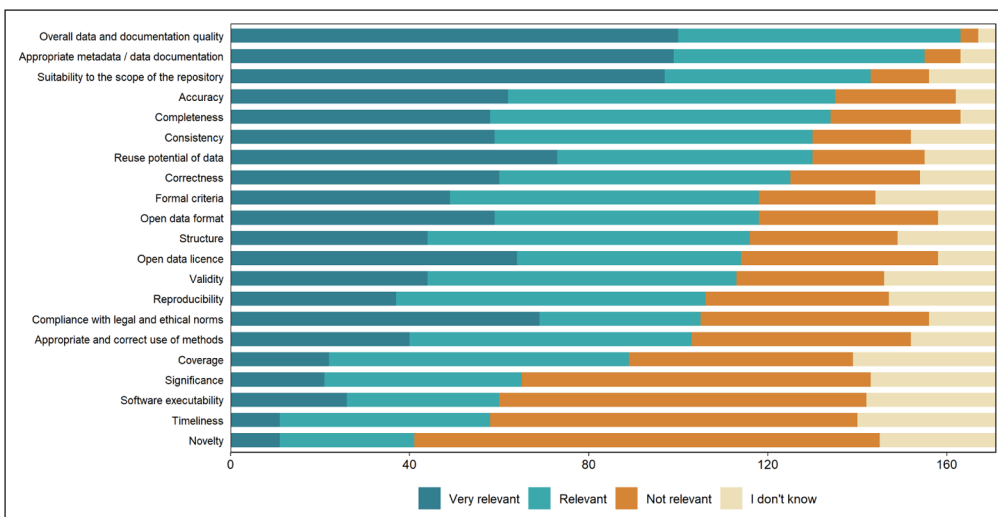
Respondents were encouraged to list any additional criteria for review of data in a free-text field (Q12, N = 28, n = 28). Responses mentioned data anonymity and included laboratory protocols or references to corresponding publications.

4.5 RESPONSIBILITY

A number of questions focused on identifying responsibilities for quality assurance activities. At the repositories that perform a formal assessment of data (Q08, N = 207, n = 518), the responsibility for the process mainly falls to the staff at the institution hosting the repository. Most respondents report that data curators (n = 137, 66.2%) or data managers (n = 109, 52.7%) at the hosting institution conduct the formal assessment, followed by technical administrators

(n = 76, 36.7%) and subject experts (n = 75, 36.2%). Data providers are also regularly involved in this step of data quality assurance (n = 75, 36.2%). Subject experts from other institutions contribute to the formal review less frequently (n = 31, 15%). The process is rarely outsourced to external partners (n = 3, 1.4%). Overall, multiple different roles seem to be involved in assessing formal criteria: respondents selected up to 6 different roles, with an average of 2.5 roles per repository.

On a more detailed level, responsibilities for the formal assessment of data vary across criteria (Q06, N = 207, n = 3519). Besides the general application of criteria for the formal assessment of data, Figure 4 also shows who is responsible for applying these criteria. These results show that the application of some criteria is seen more within the responsibility of either the data depositor or the repository. For example, providing enhanced documentation, anonymising data, making consent forms available, and the declaration of potential conflicts of interest are mainly the responsibility of the data provider. On the other hand, repositories are mainly responsible for ensuring that metadata are compliant with a metadata schema and that the metadata schema application is consistent with other metadata records in the collection and for verifying the physical integrity of datasets. The application of other criteria appears to be a shared responsibility of data depositor and repository, including ensuring a basic description of data, clarifying copyrights and usage rights, providing information on data provenance, seeking compliance with the FAIR Principles or file format requirements, assigning licences, and storing data in open formats.

Similar to the formal assessment of data, the institution hosting the repository mainly assumes responsibilities for data review (Q13, N = 171, n = 435). Most respondents report that data curators (n = 116, 67.8%) or data managers (n = 101, 59.1%) at the hosting institution review data, followed by subject experts (n = 69, 40.4%) and technical administrators (n = 52, 30.4%). Data providers are also regularly (n = 52, 30.4%) involved in reviewing data. Subject experts from other institutions contribute to the process less frequently (n = 33, 19.3%), and it is rarely outsourced to external partners (n = 2, 1.1%). Several roles within a repository tend to contribute to data review: respondents selected up to six different roles, with an average of 2.5 roles per repository. Responses in the free-text field listed additional responsibility mechanisms, including the responsibility of data depositors, the peer review process of journals, and community review of data.

4.6 DATA REVIEW PROCESS

The survey covered a number of aspects of the data review process, including the openness of the process, the acknowledgement of reviewers, and the consequences of data not meeting quality expectations.

Open processes for reviewing data are not common (Q14, N = 171, n = 171). Only a few repositories offer an open process for data review (n = 18, 10.5%). The majority of repositories do not conduct open review of datasets (n = 147, 86%). Some respondents specified details of the open review process in the free-text field, for example, descriptions of community review or references to the review process at the journals of corresponding text publications.

Overall, the recognition of reviewers is rare (Q15, N = 171, n = 175). Most repositories have not implemented measures to acknowledge reviewers (n = 99, 57.9%). Some repositories publish reviewers' names (n = 19, 11.1%), and a few compensate reviewers by paying them a fee per review (n = 3, 1.8%) or a fixed fee rate (n = 2, 1.2%). Some responses in the free-text field indicated that data review is considered a standard task of repository staff. Other respondents mentioned co-authorship or appreciative emails.

Final decisions on publishing data after the data review process is concluded (Q17, N = 171, n = 245), given that the data depositor agrees, are frequently made by repository staff (n = 132, 77.2%). In other cases, the decision is made by the data depositor (n = 57, 33.3%) or collection editor (n = 33, 19.3%). Responses in the free-text field reflect the diversity of approaches to data review. They name [...] being responsible.

Most repositories would consider taking additional steps if submitted datasets do not meet quality expectations (Q18, N = 332, n = 483). Most repositories state that data and metadata are revised until they fulfil required criteria (n = 216, 65.1%); others would consider rejecting

data deposit (n = 110, 33.1%). Quality reports are published alongside datasets at some repositories (n = 37, 11.1%), and others recommend alternative repositories (n = 33, 9.9%). Some respondents (n = 58, 17.5%) report that the scenario is not applicable to the repository. In the free-text field, some respondents stressed the responsibility of the data depositor for ensuring data quality. In this self-deposit model, datasets that do not meet quality criteria might still be published.

Of the repositories that would consider rejecting data deposit, some provided an estimation of the rate of rejected datasets in the last two years (Q19, N = 117, n = 117). On average, the respondents reported a rejection rate of 8.2% (see Figure 6). The median rejection rate is 3%. Six repositories reached or surpassed a rejection rate of 50%, and one respondent stated that 70% of datasets offered to the repository were rejected. Thirty-one respondents reported rejection rates of 0%.
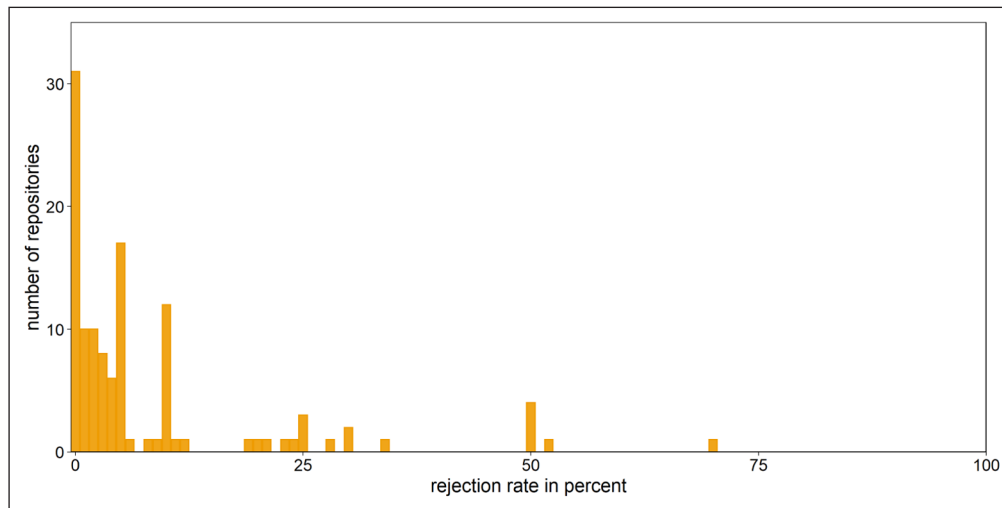
Repositories were offered the opportunity to share any additional thoughts on quality assurance at research data repositories (Q23, N = 84, n = 84). Some respondents emphasised the effort that quality assurance entails and the need for adequate recognition. Other comments described the dependence of quality assurance on various aspects, such as the scope of the repository or data types. Several respondents indicated that there are plans for developing or expanding quality assurance measures and workflows at the repository.

Overall, there is no significant association between the certification status of a repository and the aspects of the review process reported in this section.

4.7 DOCUMENTING AND PUBLISHING RESULTS

A series of questions addressed aspects of data quality information. One question focused on the documentation of the formal assessment of data (Q09, N = 207, n = 207). Only a few repositories make results of this process public (n = 27, 13%). Most respondents state that their repository does not publish the results of formal assessment (n = 154, 74.4%), even though it is documented at almost half of the responding repositories (n = 96, 46.4%). In the free-text field, some respondents stated that they consider the documentation of the review obsolete once data is published. The association between documenting the process of formal assessment of data and certification status of the repository is statistically significant (χ^2 (2, N = 207) = 6.4, p = 0.041) but weak (V = 0.19).

A similar pattern emerged for sharing the results of reviewing data (Q16, N = 171, n = 171). Results are frequently shared with the data depositor (n = 108, 63.2%), but only a few repositories publish a review report alongside the data (n = 9, 5.3%). In the free-text field, some respondents described that review reports are mainly shared internally to facilitate the review process. Others stated that review reports become obsolete with the elimination of quality issues, and subsequent data publication is not shared for this reason. The analysis found no significant relationship between sharing results of data review and the certification status of a repository.

Overall, only 9.3% (n = 31) of the responding repositories publish results of any data quality assurance process. Of these, most (n = 22) publish results of formal assessment only, as Figure 7 shows. Few repositories share results of only data review (n = 4) or of both types of quality assessment (n = 5).
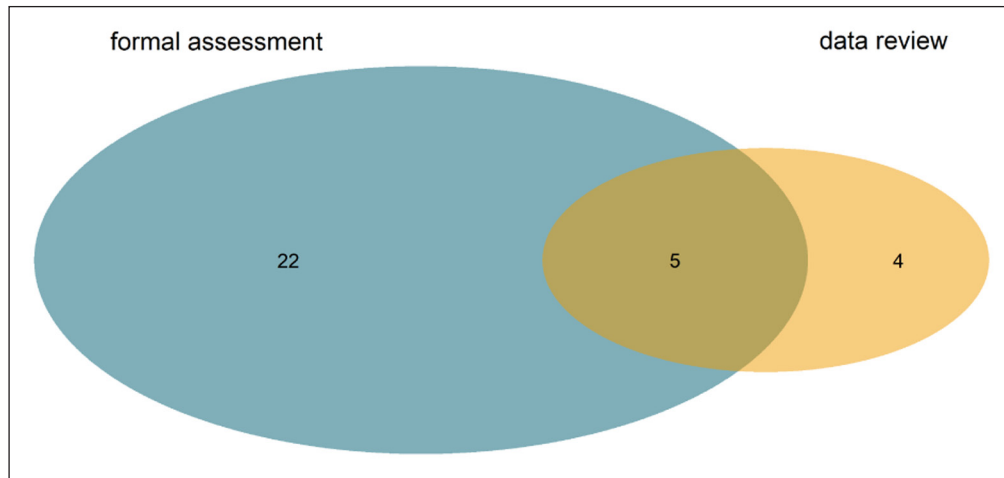
Repository users are rarely involved in the evaluation of data (Q20, N = 322, n = 387); about a third of participating repositories do not include them (n = 115, 34.6%). Several repositories receive textual feedback. Most do not make this information publicly available (n = 111, 33.4%), but some do (n = 22, 6.6%), for example, in the form of comments. Others conduct user surveys (n = 35, 10.5%) or offer data rating (n = 5, 1.5%). The involvement of repository users in data evaluation is not applicable to some repositories (n = 77, 23.2%). Responses in the free-text field detail a variety of approaches to involving repository users: for example, by organising workshops, enabling the reporting of errors in the data, or documenting conversations with colleagues about datasets at conferences.

Repositories adopt different strategies for communicating indicators of data quality to repository users (Q21, N = 322, n = 803). Most repositories use one or more indicators to communicate aspects of data quality. Most commonly, references to corresponding publications are added (n = 232, 69.9%). Other repositories display data versions (n = 169, 50.9%) or usage statistics (n = 119, 35.8%). Some repositories include data quality information in metadata (n = 88, 26.5%), display data-related citations (n = 72, 21.7%) or quality badges (n = 56, 16.9%). Less common is the publication of user survey results (n = 7, 2.1%) or data ratings (n = 5, 1.5%). Thirty-three (9.9%) respondents stated that public indicators of data quality are not applicable to the repository. Some responses in the free-text field described data quality reports, descriptions of characteristics and limitations that complement published datasets at these repositories.

5 DISCUSSION

5.1 VARIETY OF DATA QUALITY ASSURANCE MEASURES

The survey showed that approaches to and the maturity of quality assurance measures for data repositories are diverse. While most responding repositories perform a form of data quality assurance, not all assume quality assurance responsibilities. Some repositories state that they follow a self-deposit model, where data depositors are responsible for quality assurance.

Some repositories have already put in place a variety of data quality assurance measures. At these repositories, there is some indication of clear workflows, for example, support for data depositors in the form of guidelines or checklists or revisions or rejections if (meta)data of insufficient quality are submitted. Some data quality assurance practices seem to be very common; for example, some criteria for formal assessment and review of data are widely used.

The processes of data quality assurance conducted at repositories are not uniform. The formal assessment and review of data appear to be largely independent processes—repositories do not necessarily perform both. The survey demonstrated that not all measures for ensuring data quality are relevant for all research data repositories. Throughout the survey, some respondents

indicated that certain quality assurance measures are not applicable to their repository or mentioned additional quality assurance measures in free-text fields. This suggests that there is no universal approach to quality assurance but that repositories implement measures depending on scope and context.

The survey confirmed that data quality assurance at research data repositories is multifaceted and comprises a variety of activities. Based on the survey results, a framework of data quality assurance at repositories is being developed, which is intended to serve as a theoretical foundation for approaches to quality assurance of data publications at research data repositories (Kindling et al. 2022).

5.2 RESPONSIBILITIES AND THE ROLE OF REPOSITORIES IN SUPPORTING DATA QUALITY

The survey revealed that repositories significantly contribute to data quality, which is demonstrated, for example, by the surprisingly high rejection rates. Many repositories have implemented quality assurance measures, with repository staff assuming essential responsibilities. Responsibilities for data quality assurance are often organised based on a division of labour, as the number of roles involved in the formal assessment and review of data per repository shows. At some repositories, staff with very different backgrounds and skills are involved in quality assurance. These examples challenge the idea that quality assurance at repositories is based on data curators conducting formal assessment and subject specialists being responsible for data review. This raises questions about a clear separation between formal assessment and review of data, which is discussed in more detail in Section 5.4.

Several respondents emphasised the effort data quality assurance entails, yet adequate recognition is still lacking. Initiatives to properly acknowledge the contributions to quality assurance could remedy this imbalance, thereby making research data quality assurance in general and the impact of research data repositories in particular more visible.

The survey demonstrated how multifaceted data quality is. Repositories cannot be realistically expected to apply the full spectrum of quality assurance measures. Instead, repositories need to have a clear understanding of data quality assurance measures they can offer, informed by their mission, scope, and user base.

5.3 PATH DEPENDENCE

The survey confirms a path dependence of data review on the review process of text publications. Overall, the review of data appears to follow established processes for reviewing text publications. So far, few repositories have implemented an open review process, and forms of *post-publication data review* by inviting public feedback on datasets from repository users are still rare. The evaluation of data is often connected to corresponding text publications; for example, data corresponding to a peer-reviewed publication is a common factor for including datasets in the collection of a repository. The most widely used indicator for data quality is a link to the corresponding publication. Data quality assurance at research data repositories also faces similar challenges to the review processes for text publications; for example, despite their valuable contributions, reviewers are often not acknowledged.

The survey also sheds light on friction in implementing data review processes modelled after peer review for text publications. Most importantly, responsibilities for both formal assessment of data and data review currently lie almost exclusively with the institution hosting the repository. Outsourcing of data quality assurance is very rare. The strong reliance on repository staff for data review might raise questions about the independence of the peer review process, as outlined by Lawrence et al. (2011), for data archives.

Responses to free-text fields indicate that some repositories consider data review a standard task of repository staff. These expectations demand a lot of resources at the hosting institution and might not match the reality. In addition, a data review process that mainly relies on repository staff can be time-consuming and slow, potentially delaying the data publication process.

Repositories and other stakeholders should reconsider whether it is useful to emulate aspects of the review process of text publications, and if other mechanisms, such as post-publication user feedback, can be implemented.

5.4 NO CLEAR DISTINCTION BETWEEN FORMAL ASSESSMENT AND DATA REVIEW

The survey indicated that some assumptions described in the literature do not apply to all repositories. For example, it challenges a clear distinction between the formal assessment and review of data. Contrary to descriptions sometimes found in the literature, the survey demonstrates that data curators and managers do not have a clear focus on formal assessment and subject experts are not mainly responsible for data review. Instead, both roles tend to share responsibility for both tasks.

A clear chronological sequence of quality assurance measures, from an initial assessment before the ingest phase to the assessment of formal criteria to data review, was already questioned in the survey pretest. Some repositories reported that they perform data review before ingest, for example, in the context of research projects where repositories assisted in the management of data. Quality assurance should not be conceptualised as a linear process with distinct phases but should be adapted to the respective context and needs.

5.5 NO CLEAR DISTINCTION BETWEEN DATA QUALITY AND METADATA QUALITY

The survey confirms that data and metadata quality are enmeshed; there is no clear distinction between the concepts. The most widely applied criteria for both formal assessment and review of data are related to metadata or data documentation. This observation matches the fact that repositories traditionally have a strong focus on metadata, as metadata underpin essential repository functions, for example, dataset search and retrieval. The results suggest that repositories might be entering into data quality assurance, which might be a relatively new task for some services, by addressing metadata quality first, an area in which they have a lot of experience. Criteria related to metadata quality are also likely more common because they are easier to measure.

Overall, implementing quality criteria related more clearly to data (as opposed to metadata) might require a more mature service because these tasks go beyond traditional repository responsibilities. Further research could explore how repositories assess data-related quality criteria and whether these approaches can be generalised to fit other repositories.

5.6 LACK OF DATA QUALITY INFORMATION

Only about a quarter of participating repositories publish results of the formal assessment or review of data alongside the dataset. This is surprising because a lot of repositories (1) conduct data quality assessment and (2) internally document aspects of these processes.

The survey revealed that repositories seem to be more willing to share quality information if issues with data remain after publication: the number of respondents stating that their repository would consider publishing a quality report alongside a dataset if it did not fully meet quality standards is higher than the number of repositories publishing results of formal assessment and/or review of data. Responses in the free-text fields indicate that some repositories question the necessity of providing data quality information once quality issues are resolved. The discussion about if and when data quality information should be shared has far-reaching implications: for example, in the context of making repositories' successful contributions to data quality visible, tracking data provenance, and other Open Science principles. These questions should therefore be explored further.

The survey revealed that, at the moment, most repositories do not make feedback from data (re-)users public, for example, in the form of comments or ratings. Although public feedback could support decision making by potential data (re-)users, repositories would also need to take steps to prevent potential misuse.

Overall, it is not clear what format is most suitable for reporting data quality information. For example, consistent practices have not yet been established, and not all metadata schemas support reporting data quality information in structured metadata. More research in this area is required to make data quality information more widely available.

5.7 WEAK ASSOCIATION BETWEEN FORMAL CERTIFICATION AND DATA QUALITY ASSURANCE

Combining survey data with re3data metadata made it possible to analyse the association between formal certification and data quality assurance. Overall, the data revealed only two

statistically significant associations with formal certification—performing data review and documenting the formal assessment of data—and these associations were weak. These results suggest that data quality assurance and formal repository certification are largely independent. The survey could not confirm that the certification status of a repository is a clear indicator of whether a repository conducts in-depth quality assurance measures.

The main reason for this is likely that most available repository certificates do not focus on data quality assurance specifically. For example, the goal of CoreTrustSeal certification is to evaluate repositories based on a core set of requirements intended to reflect sustainability and trustworthiness of the service, making certification more widely available. Although it is not the primary focus in CoreTrustSeal, data quality is addressed in several of these requirements, most notably in Requirement 11: 'The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations' (CTS 2019: 15).

A successful application for CoreTrustSeal certification therefore requires repositories to conduct some level of quality assessment and documentation of data processing and curation, but CoreTrustSeal does not stipulate exact data quality assurance measures that must be performed. That would not be reasonable for a certificate that evaluates manifold aspects of repository operations as well as a broad spectrum of repository types; as outlined above, approaches to quality assurance are not uniform and depend on the scope and context of the individual repository. Survey results reflect this: all but two of the participating CoreTrustSeal-certified repositories (n = 33, 94.3%) conduct some form of data quality assessment—either formal assessment, review of data, or both. However, statistical analysis revealed no strong associations between specific measures of data quality assurance and CoreTrustSeal certification.

These observations could start further reflections on ways to effectively communicate information on the quality assurance measures a repository performs: who might be interested in this information and what entities besides certification organisations might deliver it. Certificates like CoreTrustSeal could try to cover data quality assurance more thoroughly, but that might be difficult given the current lack of consensus on the topic in the repository community.

At the level of individual datasets, there are more suitable indicators for signalling data quality to repository users—for example, badges or ratings—but they are not yet widely adopted. Initiatives for measuring the FAIR compliance of datasets might change this by making indicators more widely available. More research is necessary to study the prevalence of these quality indicators and their usefulness for repository users.

6 CONCLUSION

The survey demonstrated that quality assurance at research data repositories is multifaceted and nonlinear. Although there are some common patterns, individual approaches to ensuring data quality are diverse. In the context of research data, data quality and metadata quality are enmeshed and cannot be clearly separated.

Research data repositories significantly contribute to data quality. However, data quality assurance sets high expectations for repositories and requires a lot of resources. These challenges need to be addressed, for example, by critically assessing the path dependence of data review on review processes for text publications. Other approaches might be more suitable to ensuring the quality of data and should be explored further, for example, involvement of repository users in the form of post-publication data review.

Information on results of the formal assessment and review of individual datasets is not yet widely available. Approaches to publishing data quality information should be explored—for the benefit of repository users, for making the labour of data review visible, and for fostering the recognition of data publications as scientific records. How this information can be captured and exposed in a meaningful way needs to be discussed.

Similarly, information on data quality assurance measures repositories perform is currently lacking. The analysis has demonstrated that the certification status of a repository is not a clear indicator of whether it conducts in-depth data quality assurance measures. The project re3data COREF is currently evaluating how information on data quality assurance measures can be described in registries at the repository level.

Overall, a deeper understanding of data quality assurance at research data repositories can lead to a better recognition of efforts and allocation of resources.

Although this paper constitutes the first systematic and comprehensive survey of quality assurance practices at research data repositories, more research is needed to capture individual approaches to data quality assurance.

6.1 LIMITATIONS

The qualitative studies conducted before the survey identified a variety of quality criteria and quality assurance measures applied by repositories and data journals. In contrast, a questionnaire limits the number of possible responses. To obtain structured statements about a large number of repositories and at the same time capture the diversity of individual approaches, the questionnaire was supplemented by free-text fields.

The scope of the questionnaire was limited to capturing the status quo of quality assurance measures at research data repositories. Therefore, this study neither evaluates the success of these measures nor takes into account future plans.

Although the survey was distributed to a large number of repositories as possible, the results still represent a convenience sample of repositories listed in re3data. As a result, there might be a self-selection bias towards repositories already performing data quality assurance measures. Data quality assurance might also be considered a sensitive subject; therefore, some repositories may have been hesitant to participate. This issue was addressed by assuring participants of full anonymity in the survey invitation and in the privacy statement. However, it is possible that repositories where quality assurance is not applicable or not feasible might be under-represented in this paper.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary File 1: Appendix.** Overview of Survey Questions. DOI: https://doi.org/10.5334/dsj-2022-018.s1

ACKNOWLEDGEMENTS

FUNDING INFORMATION

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Maxi Kindling   orcid.org/0000-0002-0167-0466
Open-Access-Büro Berlin, Freie Universität Berlin, Germany

Dorothea Strecker   orcid.org/0000-0002-9754-3807
Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Germany

REFERENCES

Assante, M, et al. 2016. Are scientific data repositories coping with research data publishing? *Data Science Journal*, 15(6). DOI: https://doi.org/10.5334/dsj-2016-006

Austin, C, et al. 2016. Research data repositories: Review of current features, gap analysis, and recommendations for minimum requirements. *IASSIST Quarterly*, 39(4): 24. DOI: https://doi.org/10.29173/iq904

Austin, C, et al. 2017. Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, 18(2): 77–92. DOI: https://doi.org/10.1007/s00799-016-0178-2

Batini, C and **Scannapieco, M.** 2016. Data quality dimensions. In: Batini, C and Scannapieco, M (eds.), *Data and information quality: Dimensions, principles and techniques*. Berlin: Springer. pp. 21–51. DOI: https://doi.org/10.1007/978-3-319-24106-7_2

Cai, L and **Zhu, Y.** 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(2). DOI: https://doi.org/10.5334/dsj-2015-002

Canali, S. 2020. Towards a contextual approach to data quality. *Data*, 5(4): 90. DOI: https://doi.org/10.3390/data5040090

CASRAI (Consortia Advancing Standards in Research Administration Information). 2022a. Curation. Available at https://casrai.org/term/curation/ [Last accessed 30 August 2022].

CASRAI (Consortia Advancing Standards in Research Administration Information). 2022b. Data quality. Available at https://casrai.org/term/data-quality/ [Last accessed 30 August 2022].

CTS (CoreTrustSeal Standards and Certification Board). 2019. CoreTrustSeal trustworthy data repositories requirements 2020–2022. DOI: http://doi.org/10.5281/zenodo.3638211

Downs, R, et al. 2021. Perspectives on citizen science data quality. *Frontiers in Climate*, 3. DOI: https://doi.org/10.3389/fclim.2021.615032

ISO (International Organization for Standardization). 2015. Quality management systems—Fundamentals and vocabulary (ISO 9000:2015). Available at https://www.iso.org/standard/45481.html [Last accessed 30 August 2022].

Johnston, L, et al. 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

Juran, JM. 1951. *Quality-control handbook*. New York: McGraw-Hill.

Kindling, M and **Strecker, D.** 2021. How to ensure 'good' data? A presentation at Open Repositories 2021. Available at https://coref.project.re3data.org/blog/how-to-ensure-good-data-a-presentation-at-open-repositories-2021 [Last accessed 30 August 2022].

Kindling, M, Strecker, D and **Wang, Y.** 2022. Data quality assurance at research data repositories: Survey data (1.0) [data set]. *Zenodo*. DOI: https://doi.org/10.5281/ZENODO.6457849

Kindling, M, et al. 2017. The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, 23(3/4). DOI: https://doi.org/10.1045/march2017-kindling

Kindling, M, et al. 2022. Data quality assurance at research data repositories—Results from a survey. In: *International Digital Curation Conference* on 13–16 June 2022. DOI: https://doi.org/10.5281/ZENODO.6638409

Koltay, T. 2020. Quality of open research data: Values, convergences and governance. *Information*, 11(4): 175. DOI: https://doi.org/10.3390/info11040175

Koshoffer, A, et al. 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. DOI: https://doi.org/10.2218/ijdc.v13i1.632

Lafia, S, et al. 2021. Leveraging machine learning to detect data curation activities. In: *IEEE 17th International Conference on eScience*, Innsbruck, Austria, 20–23 September 2021. DOI: https://doi.org/10.1109/eScience51609.2021.00025

Lawrence, B, et al. 2011. Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*, 6(2). DOI: https://doi.org/10.2218/ijdc.v6i2.205

Lee, DJ and **Stvilia, B.** 2017. Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE*: e0173987. DOI: https://doi.org/10.1371/journal.pone.0173987

Lee, Y, et al. 2002. AIMQ: A methodology for information quality assessment. *Information & Management*, 40(2): 133–146. DOI: https://doi.org/10.1016/S0378-7206(02)00043-5

Madnick, S, et al. 2009. Overview and framework for data and information quality research. *Journal of Data and Information Quality*, 1(1): 1–22. DOI: https://doi.org/10.1145/1515693.1516680

Mangione, D, Candela, L and **Castelli, D.** 2022. A taxonomy of tools and approaches for FAIRification. In: *CEUR Workshop Proceedings, Padova*, Italy, 24–25 February 2022. Available at http://ircdl2022.dei.unipd.it/downloads/papers/IRCDL_2022_paper_18.pdf [Last accessed 30 August 2022].

Mayernik, M, et al. 2015. Peer review of datasets: When, why, and how. *Bulletin of the American Meteorological Society*, 96(2): 191–201. DOI: https://doi.org/10.1175/BAMS-D-13-00083.1

Merriam-Webster. 2022. Quality. Available at https://www.merriam-webster.com/dictionary/quality [Last accessed 30 August 2022].

OKF (Open Knowledge Foundation). n.d. Open Definition: Version 2.1. Available at http://opendefinition.org/.

Palmer, C, Weber, N and **Cragin, M.** 2011. The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1–10. DOI: https://doi.org/10.1002/meet.2011.14504801174

Parr, C, et al. 2019. A discussion of value metrics for data repositories in earth and environmental sciences. *Data Science Journal*, 18: 58. DOI: https://doi.org/10.5334/dsj-2019-058

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

Peer, L, Stephenson, E and **Green, A.** 2014. Committing to data quality review. *International Journal of Digital Curation*, 9(1): 263–291. DOI: https://doi.org/10.2218/ijdc.v9i1.317

Peng, G, et al. 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13: 231–253. DOI: https://doi.org/10.2481/dsj.14-049

Peng, G, et al. 2022. Global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. *Data Science Journal*, 21: 8. DOI: https://doi.org/10.5334/dsj-2022-008

Plantin, J -C. 2019. Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values*, 44(1): 52–73. DOI: https://doi.org/10.1177/0162243918781268

RfII (German Council for Scientific Information Infrastructures). 2020. The data quality challenge: Recommendations for sustainable research in the digital turn. Göttingen. Available at https://nbn-resolving.org/urn:nbn:de:101:1-2020041412321918717265

Strecker, D, et al. 2021. Metadata schema for the description of research data repositories: Version 3.1. DOI: https://doi.org/10.48440/re3.010

Trisovic, A, et al. 2021. Repository approaches to improving the quality of shared data and code. *Data*, 6(2): 15. DOI: https://doi.org/10.3390/data6020015

Wang, R and **Strong, D.** 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4): 5–33. DOI: https://doi.org/10.1080/07421222.1996.11518099

Wilkinson, M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018. DOI: https://doi.org/10.1038/sdata.2016.18