Informationspraxis Bd. 8, Nr. 1 (2022)
Berichte und Kritik

DOI: https://doi.org/10.11588/ip.2022.1.93173

Maxi Kindling, Dorothea Strecker, Lea Maria Ferguson, Hervé L'Hours, Cristina Magder, Rouven Schabinger, Seta Štuhec, Paul Vierkant, Nina Weisweiler

Report on re3data COREF / CoreTrustSeal Workshop on Quality Management at Research Data Repositories

Zusammenfassung

Am 5. Oktober 2022 fand der gemeinsam von re3data COREF und CoreTrustSeal ausgerichtete Online-Workshop "Quality Management at Research Data Repositories" statt, den über 70 Teilnehmende besuchten. Ziel des Workshops war es, Aktivitäten von Forschungsdatenrepositorien zur Sicherung, Bewertung und Verbesserung der Datenqualität zu diskutieren.

Der Workshop begann mit einem Input der Workshop-Organisatoren: re3data COREF präsentierte die Ergebnisse einer kürzlich durchgeführten Umfrage zum Qualitätsmanagement bei Repositorien und CoreTrustSeal stellte die Perspektive einer Zertifizierungsorganisation vor. Anschließend präsentierten Repositorien verschiedener Disziplinen ihre Ansätze zum Qualitätsmanagement. Der Workshop schloss mit einer Breakout-Session und einer Diskussion über Möglichkeiten, Informationen zur Datenqualitätssicherung besser sichtbar zu machen.

Abstract

On October 5, 2022, the "Workshop on Quality Management at Research Data Repositories" – jointly organized by re3data COREF and CoreTrustSeal – was held online with more than 70 participants attending. The objective of the workshop was to discuss activities research data repositories perform to assure, assess, and improve data quality.

The workshop started with input from the workshop organizers: re3data COREF presented results of a recent suvey on quality management at repositories, and CoreTrustSeal shared the perspective of a certification organization. Then, repositories from different disciplinary backgrounds presented their approaches to quality management. The workshop concluded with a breakout session and a plenary discussion on options for making information on data quality assurance more visible.

Keywords

research data, research data repositories, quality assurance

Veröffentlichung
04.02.2023 in Informationspraxis Bd. 8, Nr. 1 (2022)

Dieses Werk steht unter der Lizenz Creative Commons Namensnennung 4.0 International

Inhaltsverzeichnis

1 Introduction **2**
1.1 Research data and quality assurance 2
1.2 Workshop format 2

2 Part I: Input from workshop organizers **3**
2.1 re3data COREF 3
2.2 CoreTrustSeal 4

3 Part II: Repositories' approaches to data quality management **5**
3.1 PANGAEA 5
3.2 ARCHE 5
3.3 UK Data Service 6

4 Part III: Breakout Sessions **6**

5 Plenary Discussion **7**

6 Conclusion **7**

7 Workshop Resources **8**
7.1 Presentation slides 8
7.2 Recordings 8
7.3 Zotero Bibliography 8

References **8**

1 Introduction

On October 5, 2022, the „Workshop on Quality Management at Research Data Repositories" – jointly organized by re3data COREF and CoreTrustSeal – was held online. More than 70 participants attended the workshop and most of them are actively involved with a repository or a similar service. The objective of the workshop was to discuss and highlight the numerous activities research data repositories perform to assure, assess, and improve data quality.

1.1 Research data and quality assurance

Definitions of research data quality are often context-dependent; data quality is commonly conceptualized as „fitness for use". In this understanding, quality is determined by whether research data are suitable to be used for a specific purpose. Sometimes, these definitions result in the specification of criteria – i.e., characteristics that data must exhibit to be considered fit

for use. Data quality assurance involves several activities, including a) the assessment of data, b) necessary actions taken to ensure that data meet these criteria, and c) the documentation of requirements and quality levels achieved by those measures. It is a fundamental issue for repositories that seek to ensure the usability of the data they hold and trust in their collection and services. However, information on the specific measures repositories undertake to ensure data quality is currently limited.

1.2 Workshop format

The workshop comprised three parts. In part 1, the workshop organizers gave an overview of the current developments in quality management for research data. Results of a recent survey among research data repository operators were presented to outline the status quo of quality management for research data. In the following presentation, the perspective of the certification organization CoreTrustSeal on data quality assurance was shared. In part 2, operators of repositories with different scopes (earth and environmental sciences, humanities, and social sciences) presented their approaches to quality assurance. In part 3, workshop participants were invited to breakout sessions to share measures for quality assurance that they have implemented and challenges they are facing in this context. The workshop concluded with a plenary discussion on options for making information on data quality assurance more visible at the level of repositories. Presentation slides and recordings of part 1 and 2 are linked below.

2 Part I: Input from workshop organizers

2.1 re3data COREF

Maxi Kindling from re3data COREF at Berlin School of Library and Information Science, Humboldt-Universität zu Berlin gave an overview of research carried out within her PhD work and the re3data COREF project (funded by German Research Foundation, DFG). Overall goals of the work package „Quality & Transparency“ within the re3data COREF project are to explore current activities of quality assurance and determine how the re3data Metadata Schema can be revised to better reflect the process of quality management at repositories. Following a mixed-methods approach, the research included qualitative analyses of data journal guidelines and CoreTrustSeal self-assessment documents, as well as a survey on data quality management among 332 repository operators. Based on the various analyses, a framework on data quality assurance at research data repositories was developed. This framework guided the understanding of data quality assurance practices used for the workshop. The framework further provides a holistic concept of quality assurance and outlines six categories spanning the entire data life cycle from (pre-)ingest (quality definition; quality development) to curation activities (quality control; quality improvement) to access and (re-)use (quality evaluation; quality documentation) (see Figure 1).

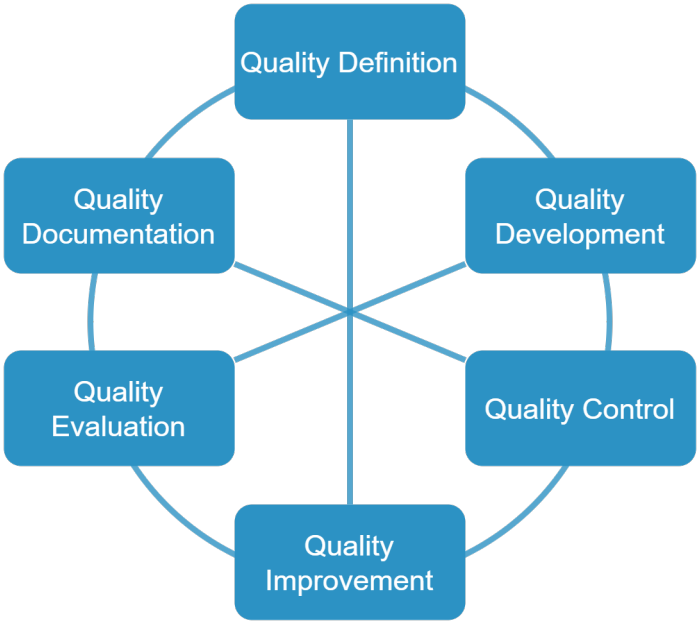

Fig. 1: Categories of the Framework for Quality Assurance of Data Publications at Research Data Repositories

The survey results have been described in a separate publication (Kindling & Strecker 2022) and will therefore not be discussed in detail here. The questionnaire covered various aspects of quality management, including types of data quality assessment, quality criteria, responsibilities, details of the review process, and data quality information. In summary, the results of the survey show that repositories play a significant role in the quality assurance of data publications. The survey also revealed that the prevailing approaches to quality assurance are diverse. Repositories perform a variety of measures and practices, but not necessarily from all categories described in the framework. Furthermore, quality assurance in general is not relevant to all repositories equally. The relevance of quality assurance measures depends, among other factors, on the type and scope of the repository. The results of our study also show that the quality assurance measures that are currently being performed need to be better acknowledged, this may result in an increased visibility of these efforts.

2.2 CoreTrustSeal

Hervé L'Hours, Preservation and Repository Manager at the UK Data Archive and Vice Chair of the CoreTrustSeal Board, presented the perspective of CoreTrustSeal on data quality management. CoreTrustSeal is a standards and certification organization that evaluates research data repositories based on a core set of 16 requirements. These requirements are regularly reviewed and reflect characteristics of trustworthy data repositories. Applicants participate in a self-assessment of their service and practices against the requirements set by

CoreTrustSeal; applicants then provide evidence for their claims, and these are subsequently verified by selected reviewers. Since the publication of the first set of requirements in 2017, 136 repositories have received CoreTrustSeal certification. The CoreTrustSeal Trustworthy Digital Repositories Requirements 2023-2025 have recently been published. (CoreTrustSeal Standards And Certification Board 2022) From the perspective of CoreTrustSeal, repositories are enablers and communicators, rather than guarantors of data quality. Several CoreTrustSeal requirements touch on aspects of data quality, but most prominently Quality Assurance (R10). R10 refers to technical quality and standards compliance, which means that formal aspects rather than the scientific value of repository collections are evaluated. Data quality assurance explicitly includes metadata; therefore, repositories must also provide sufficient information for the designated community to make quality-related decisions. This approach to data quality is a) context-dependent, as it is centered around the expectations of the specific community a repository serves, and b) pragmatic, as it acknowledges that even flawed (meta-)data may retain value for re-use if it is sufficient for supporting decision-making in the designated community. As a certification organization, CoreTrustSeal is interested in the long-term perspective of quality assurance. A major challenge for CoreTrustSeal is the lack of best practices for quality assessment or its outcomes, be it in general or for a specific discipline. In addition, it remains to be seen which aspects of data quality management can be automated and which will continue to require human mediation.

3 Part II: Repositories' approaches to data quality management

3.1 PANGAEA

Janine Felden, PANGAEA Group Leader at the Alfred Wegener Institute in Bremerhaven, Germany, presented quality assurance measures implemented at PANGAEA, a repository for earth and environmental science data. PANGAEA staff comprises a back office representing the core team, responsible for IT, content management, and training, as well as a front office of trained data managers and editors. This structure makes it possible to cater to the needs of a diverse audience and a wide range of data types. Data management processes are organized internally using a ticket system (Jira by Atlassian). This allows for these processes to be meticulously tracked and resulting tasks to be shared efficiently. PANGAEA also has developed several tools that assist data management processes. Data review at PANGAEA follows a two-step process: 1) initially, the relevance of the respectively submitted data is evaluated; 2) if the submission is considered suitable for publication in PANGAEA, an in-depth review of the dataset follows. PANGAEA requires data providers to submit a minimum set of metadata for each dataset. Automated metadata consistency checks are performed periodically. Results of the review process are included in the metadata of a dataset. Supporting documents for data providers are available, for example in the form of wiki entries and templates. Currently,

changes in the operation of the ticket system Jira is a cause of concern for PANGAEA since Atlassian is moving to a cloud-based business model, and costs will increase significantly. As a result, many institutions, including PANGAEA, are looking for alternatives.

3.2 ARCHE

Seta Štuhec, responsible data curator at the Austrian Academy of Sciences in Vienna, Austria presented quality assurance measures performed at ARCHE ARCHE, a repository for the humanities.
ARCHE has developed a comprehensive collection policy and guidelines for data deposit and metadata preparation. ARCHE uses an open source ticket management system (Redmine) to organize data management tasks and track processes. Data review is a two-step process at ARCHE. The first step in the ingest phase is automated: a „doorkeeper“ script validates metadata against the schema. In the second step, two curators perform manual checks following a checklist that includes criteria for structure, reusability, and legal aspects of the submitted data. Data providers are included in this process. ARCHE organizes regular „metadata cleanup weeks“, where curators retroactively check and improve metadata records. Quality management processes are documented using the ticket system, resulting in a detailed curation log.

3.3 UK Data Service

Cristina Magder, Data Collections Development Manager for UK Data Archive at the University of Essex, UK, gave insights into quality management measures at UK Data Service, a nationally funded research infrastructure for the social sciences. UKDA was established in 1967 as a center of expertise in acquiring, curating and providing access to social science data. In October 2012 UKDA became the lead partner of the UK Data Service. UK Data Service brings together seven host institutions to support high quality social science research, teaching and learning. Data quality is well defined at UK Data Service; several documents record the requirements for data depositors, including the collections development policy and appraisal criteria, deposit user guides, and templates for documentation. Internal quality management processes are also documented, including pre-ingest procedures, data curation standards and de-identification and disclosure review. Quality control at UK Data Service follows a curation workflow that includes semi-automated and manual steps. Data review includes checking data validity and anonymisation as well as documentation and metadata. Processing standards range from level A* to C and guide the extent of curation activities performed, including quality assurance measures. UK Data Service regularly organizes training for data producers and secondary data users and provides necessary guides and resources. Bespoke and standard training sessions ensure valuable feedback from repository users. Data curation activities, including quality assurance measures, are recorded and retained for long-term preservation: For each dataset, a README file is created that outlines data processing for data users.

4 Part III: Breakout Sessions

In the breakout sessions, participants were invited to share and discuss experiences with quality management at their repositories. The framework for quality assurance of data publications was used to guide the session. Participants reported a variety of quality assurance measures at their repositories, many specifically tailored to the type, scope, and collection of the repository. Many repositories automate specific quality management processes; all include manual checks by repository staff. Guidance of data providers, including individual consultations or training sessions in larger groups, is seen as an important preventative measure to ensure data quality. Few participants report approaches that involve users in the process of quality evaluation, e.g., in the form of user surveys or case studies.

The discussion revealed that most repositories create documentation of their data quality management processes; however, the resulting documents are only made public if they are perceived to be useful for data depositors. Among the challenges that were reported by repository managers, a prominent aspect is finding a shared understanding of data quality. Currently, neither terminology nor quality criteria are well defined, and disciplinary boundaries can further compound this issue. Another significant challenge are scarce repository resources, mainly relating to staff and time. There is a shortage of staff in general, and in some cases of staff with a particular disciplinary background. Time constraints of both repository staff and data depositors limit data quality management activities. Participants also reported that managing staff is a challenge, for example to ensure good communication in distributed teams. Automation is seen as a solution for scarce resources, but repositories with heterogeneous collections find it difficult to automate processes. Participants also listed changing expectations with regards to data quality among the challenges. Changing expectations can lead to „legacy data“ which needs to be re-reviewed to meet current standards. Some repositories perform retroactive quality checks, but are mainly restricted to checking samples due to limited resources.

5 Plenary Discussion

The workshop concluded with a plenary discussion, which centered around the topic of data quality information in repository registries. Workshop participants indicated that they would find it useful if information on data quality management activities was displayed in repository registries, such as re3data, as this could help potential repository users to see which particular services can be expected from a particular repository. The information could be delivered similar to the process at CoreTrustSeal that currently distinguishes between „levels of curation“: CoreTrustSeal delineates repositories based on the extent of curation and long-term preservation that these repositories provide. The CoreTrustSeal levels of curation are currently under review. While improving information on data quality assurance measures in registries is perceived as useful overall, participants expressed concerns: For example,

transferring information from the dataset level (specific quality assurance measures performed for individual datasets) to the repository level requires an evaluation by registries and can pose challenges due to the extensive resources required for this task. In addition, repositories have their unique mission and policies, often informed by the institution or community they serve. Therefore, not all quality assurance measures are relevant to all repositories, and it might be difficult to communicate that in registry entries.

6 Conclusion

The workshop has shown that approaches to data quality management at repositories are diverse. This also extends to the terminology used to describe data quality. The framework for quality assurance of data publications described above could be the first step towards a common vocabulary of data quality management, but it is necessarily abstract and does not cover individual measures and procedures. Presentations and discussions have also demonstrated that repositories can learn from each other at events like this workshop. Over the years, the repository community has developed knowledge and expertise and a variety of reusable materials and resources. Communication and collaboration can foster the diffusion of quality assurance measures. Results of workshop discussions will inform the revision of the re3data Metadata Schema (CoreTrustSeal Standards And Certification Board 2022), with the intention to make quality management at research data repositories more visible.

7 Workshop Resources

7.1 Presentation slides

- Maxi Kindling & Dorothea Strecker: Quality Management at Research Data Repositories. Results from a survey and Framework of Quality Assurance for Data Publications at Research Data Repositories: https://doi.org/10.5281/zenodo.7142736
- Hervé L'Hours: Data Quality Assurance from the Perspective of CoreTrustSeal: https://doi.org/10.5281/zenodo.7228059
- Seta Štuhec: Quality Assurance at ARCHE: https://doi.org/10.5281/zenodo.7228083
- Cristina Magder: Quality Assurance at UK Data Service: https://doi.org/10.5281/zenodo.7228029

7.2 Recordings

- Maxi Kindling: Quality Management at Research Data Repositories. Results from a survey and Framework of Quality Assurance for Data Publications at Research Data Repositories: https://doi.org/10.5446/59609

- Hervé L'Hours: Data Quality Assurance from the Perspective of CoreTrustSeal: https://doi.org/10.5446/59610
- Seta Štuhec: Quality Assurance at ARCHE: https://doi.org/10.5281/zenodo.7228083
- Cristina Magder: Quality Assurance at UK Data Service: https://doi.org/10.5446/59763

7.3 Zotero Bibliography

This reference collection was initiated on the occasion of the „re3data COREF / CoreTrustSeal Workshop on Data Quality Management at Resesearch Data Repositories“ (2022-10-05). The collection is continuously curated by Dorothea Strecker (re3data COREF at Humboldt-Universität zu Berlin) and Maxi Kindling (Open-Access-Büro Berlin). It is open for participation by interested colleagues. https://www.zotero.org/groups/4779941/data_quality_management_at_repositories

References

CoreTrustSeal Standards And Certification Board 2022. CoreTrustSeal Requirements 2023-2025. https://doi.org/10.5281/ZENODO.7051012

Kindling, Maxi & Strecker Dorothea 2022. Data Quality Assurance at Research Data Repositories. *Data Science Journal*, 21(1), 18. https://doi.org/10.5334/dsj-2022-018

Strecker, Dorothea, Bertelmann, Roland, Cousijn, Helena, Elger, Kirsten, Ferguson, Lea Maria, Fichtmüller, David, Goebelbecker, Hans-Jürgen, Kindling, Maxi, Kloska, Gabi, Nguyen, Thanh Binh, Pampel, Heinz, Petras, Vivien, Schabinger, Rouven, Schnepf, Edeltraut, Semrau, Angelika, Trofimenko, Margarita, Ulrich, Robert, Upmeier, Arne, Vierkant, Paul, Weisweiler, Nina Leonie, Wang, Yi, Witt, Michael 2021. Metadata Schema for the Description of Research Data Repositories: Version 3.1. https://doi.org/10.48440/re3.010

Authors

- Maxi Kindling, maxi.kindling@open-access-berlin.de

  Open-Access-Büro Berlin

  https://orcid.org/0000-0002-0167-0466

- Dorothea Strecker, dorothea.strecker@hu-berlin.de

  Humboldt-Universität zu Berlin

  https://orcid.org/0000-0002-9754-3807

- Lea Maria Ferguson

Helmholtz Association, Helmholtz Open Science Office

https://orcid.org/0000-0002-7060-3670

- Hervé L'Hours

  UK Data Archive

  https://orcid.org/0000-0001-5137-3032

- Cristina Magder

  University of Essex

  https://orcid.org/0000-0001-5937-8188

- Rouven Schabinger

  Karlsruhe Institute of Technology (KIT)

  https://orcid.org/0000-0002-0249-7917

- Seta Štuhec

  University of Vienna

  https://orcid.org/0000-0002-1218-9635

- Paul Vierkant

  DataCite

  https://orcid.org/0000-0003-4448-3844

- Nina Weisweiler

  Helmholtz Association, Helmholtz Open Science Office

  https://orcid.org/0000-0001-6967-9443