Data quality in research

A framework of the roles and processes in research data quality, and areas of focus for the future.

## 

## Introduction

### Data-dependent scientific research

We live in an increasingly data-centric world where data have become more prevalent and pervasive than ever before. Technological advances have led to the proliferation of data and accessible datasets, which contain details about any and all subjects, both public and private. This has led to fundamental changes in the way science and research is conducted. We are amid a transition from predominantly narrative-based scientific research to one that is more data-dependent. While data produced from research were once small and usually set in the tables or figures of an article, data volume has increased, and more fields have more types of data available to them.

The era of Big Data and data-dependent research has created new untapped potential. Data can now be used in novel contexts other than in which they were generated, not in the least through the application of innovative technologies such as artificial intelligence.

Increased access to data can play a vital role in efforts to increase the reproducibility and the integrity of science. Access to the data underlying research reported in scientific articles and books allows researchers to better assess the integrity of the work and reproduce and replicate the published findings.

Nowhere is the importance of scientific data more visible than in the research into COVID-19 – which in all aspects seems to have accelerated existing research trends rather than created new ones. The development of the pandemic, the effectiveness of public responses, the way it was reported, how public understanding developed, and how attitudes changed over time all demonstrate the need for and expectations of high-quality data, and how research has adapted to deliver that.

Research has begun to adapt to this new data-intensive world. In the process, the challenges of how to collect, share and present data to a range of audiences must be integrated with questions as to whether the data are accurate and actually address the questions researchers wish to answer. We must better understand what constitutes quality data, how it can be ensured, and the processes and roles involved.

Put directly: improving, assessing, vetting, and communicating the level of quality of data in the context of its purpose is pivotal in today's scientific environment. We need to understand the levels of quality alongside its dimensions, and the various roles and responsibilities that play a part in the process.

Sharing and making data openly available has not been normal practice historically, and hence the attempt to do so is not just a technological problem of data packaging but a sociological and cultural challenge involving disruption of barriers inherent in scientific practice.

Understanding of the importance of these issues emerged from group discussions at the [STM Research Data Program](http://stm-researchdata.org) (RDP), aimed at boosting the effective sharing, linking and citing of research data alongside publications. This White Paper has developed from desk-based research and in consultation with leading experts, all of whom are involved with aspects of data quality across scholarly communications and research.

This Research Paper articulates a framework for thinking about data quality as it operates within scientific research[[1]](#footnote-2). It is primarily meant to serve to start a broad conversation among the various players in the research ecosystem, who will be shown to all play a significant role in research data quality. We hope this report offers a valuable perspective on the current transition to data-dependent science and we look forward to seeing more initiatives and partnerships of the types highlighted in this Research Paper emerge and spread. We have tried to highlight what we believe are the most critical areas of focus and development and indicate where the most fruitful partnerships may emerge.

### Overview of the Research Paper

The data quality framework we propose in this Research Paper is not new. It draws on existing work and initiatives on how the term ‘data quality’ is being used in an increasingly complex setting, highlighting the characteristics of data quality and the processes and initiatives that operate across research and scholarly communications, and on how these aspects connect to data quality and to one another.

Section 1 sets out a framework of quality characteristics appropriate for research data.

Section 2 provides an overview of diverse processes and mechanisms operated to ensure research data quality throughout the research cycle. It highlights how these processes overlap and interlink and provides examples and brief case studies.

Section 3 offers some areas of focus for the research community suggested by the framework and the successful developments various groups have already put in place.

### Acknowledgments

STM would like to thank the following experts for their generosity in contributing their expertise and time to the consultation conducted in the research phase of this Research Paper.

Dr Astrid Wittmann – Data and Project Manager, PANGEA

Dr Daniel Hook – CEO, Digital Science

Daniella Lowenberg – Product Manager, Dryad & Principal Investigator, Make Data Count California Digital Library

Dr Elizabeth Ford – Senior Lecturer in Primary Care Research, University of Sussex

Dr Helena Cousijn – Community Engagement Director, DataCite

Iain Hrynaszkiewicz – Director, Open Research Solutions, PLoS

Dr Ian Bruno – Director of Data Initiatives, Cambridge Crystallographic Data Centre

Dr Ingrid Dillo – Council Co Chair, Research Data Alliance; Deputy Director, DANS; Coordinator, FairsFair

Jennifer Gibson – Managing Director, Dryad

Dr Lars Moller – Data Manager, PANGEA

Dr Mark Hahnel – CEO, Figshare

Maaike Verburg – Research Data Management Specialist, Data Archiving and Networked Services (DANS)

Marta Teperek – Head of Research Data Services, TU Delft

Matt Cannon – Head of Open Research, Taylor & Francis

Dr Robert Huber – Data Publisher for Earth & Environmental Science, PANGEA

Yan Wang – Data Stewardship Coordinator, TU Delft

We are particularly grateful to Dr Daniel Hook, Dr Helena Cousijn, Dr Ian Bruno, Matt Cannon, and Yan Wang for their detailed comments on draft versions of this report.

Conducted primarily by Great North Wood Consulting, the project was led and coordinated by Dr Joris van Rossom, STM Director of Research Integrity, and Eefke Smit, STM Director of Standards and Technology.

### 

## Data quality

In this section we set out a framework of characteristics of research data quality. It considers the concept and characteristics of data quality as articulated in academic literature, and how these might be applied in a research context. It also considers the role metadata must play in ensuring data quality.

### The concept of data quality

There is substantial general academic literature on data quality; it is a well-defined if multifaceted concept. Work dates back several decades and is most developed in management and business fields, where consideration of quality management arose in manufacturing and industry settings, as expressed by principles such as Total Quality Management (TQM) or Identify, Quantify, Implement, Perfect (IQIP)[[2]](#footnote-3).

This literature is rich in various data quality frameworks, which articulate distinct characteristics or dimensions of data quality (see Figure 1 for a prominent example). These foundational data concepts define aspects that are the hallmarks of quality and can help to identify ways to improve that quality[[3]](#footnote-4).

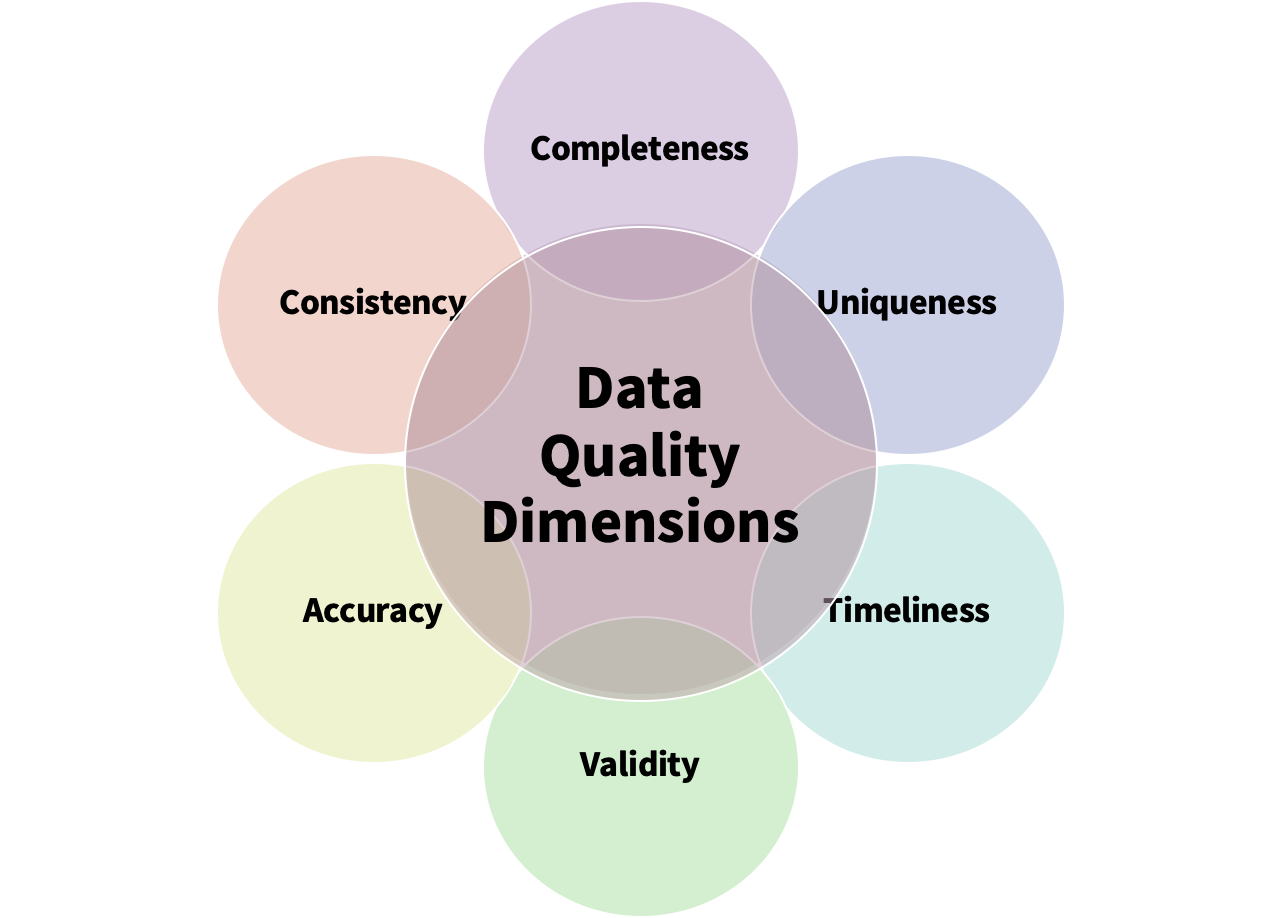


Figure 1: DAMA-UK (2013). The six primary dimensions for data quality assessment. October 2013.

Surveys[[4]](#footnote-5) and systematic reviews that identify and synthesize aspects of data quality have routinely identified more than one hundred distinctive characteristics. There is considerable overlap and different frameworks emphasize different elements[[5]](#footnote-6). Some focus on the development of data versus their use. Others individually consider the various elements that go into data construction, separating the data model, specific data values and alternative representations of data. The need for data of high intrinsic quality have always been a key element, incorporating such aspects as accuracy and completeness in relation to well-defined, specified cases of use.

The need for specialist processes and frameworks to manage Big Data, and the volumes of data they create, have developed the concept extensively. The advent of Big Data and the challenges it entails has brought greater focus5-[[6]](#footnote-7) on how data are represented, accessed and the contexts in which they are used.

### Considering data quality in research

While there is a broad and varied deal of literature on the use of data in academic settings, there is little systematic consideration of *data quality* in the research setting.

Table 1 outlines a framework utilizing data characteristics from the general literature[[7]](#footnote-8) and accommodates aspects of data quality unique to scientific research. It aims to incorporate the diverse ways in which researchers and other agents use data quality in research, while remaining conceptual. Greater understanding of the characteristics is best achieved through practical examples and by exploring how the quality of research data is assured; that is our focus in the next section.

|  |  |  |
| --- | --- | --- |
| Reliability | Accuracy | The degree to which data represent real world or described events. |
| Consistency | The degree to which data attributes are coherent and free of contradiction. |
| Completeness | The degree to which all required data are present. |
| Integrity | The absence of data value loss or corruption. |
| Relevance | Fitness for purpose | The degree to which data suit the intended purpose; the degree of [fit between data, methods, materials available and research questions](https://datascience.codata.org/articles/10.5334/dsj-2017-032/)[[8]](#footnote-9). |
| Utilization[[9]](#footnote-10) | The degree to which researchers have used data, either in the original context or to support novel empirical claims. |
| Presentation | Readability | The degree to which data are clear, simple to understand, interpretable and comprehensible to users, both human and machine. |
| Structure | The degree to which organization and format of data are appropriate and consistent. |
| Availability | Accessibility | The ease with which data can be discovered, consulted, or retrieved. |
| Timeliness | The degree to which the timing of data creation and the availability of the dataset are appropriate. |
| Authorization | The degree to which data are accessible and interpretable only by  suitably authorized users. |
| Usability | Definition/Documentation | The degree to which data values are in accordance with their definition, format specification. |
| Provenance | The degree to which the sources of data can be identified and validated, and how data were generated, transformed, and have traveled. |

Table 1: A framework for data quality in research

The following points were important considerations in developing the framework:

#### Intrinsic quality of data

Intrinsic quality is often mentioned in connection to research. It is taken to mean that data have quality in their own right and that the data values of accuracy, completeness and integrity conform to real world values.

However, data are always constructed. They result from the choices made by researchers on how and what to measure, and at what level of precision. Similarly, judgments on data quality are always made relative to the expertise, commitments, discipline, values, and goals[[10]](#footnote-11) of the person forming the judgment. Data quality is always context dependent.

Additionally, using data to support epistemic claims is a characteristic of quality itself. This becomes more relevant with the greater availability of data, meaning that data are not only used in varying contexts and unknown users[[11]](#footnote-12) can put data to unknown uses but also that some data remain unused.

All of this leads us to reject using *Intrinsic* as a specific characteristic of data quality, but instead to capture it within two separate qualities.

* *Reliability* captures the accuracy, completeness and integrity of data, the importance that data reflect the real world.
* *Relevance* captures the use of data to make claims and the fitness of data for claims that are made, whether that be the original context in which data was generated, or novel unanticipated contexts.

This reflects the way these terms[[12]](#footnote-13) are used in the literature on data quality and avoids the complications inherent in the term ‘Intrinsic data.’ The split into two separate terms does not mean that we cannot acknowledge the relationship between *Relevance* and *Reliability*.

#### Reproducibility & reproduction of research

Data quality is often mentioned in relation to the reproducibility crisis, whereby, for example, 50% of researchers cannot reproduce their own work, or 70% cannot reproduce the work of other researchers[[13]](#footnote-14).

We consider the reproduction of data as a mechanism for data quality assurance, specifically their *Reliability* and *Relevance*. We will return to this topic in depth in the next section (as well as the important distinction between reproduction and replication).

#### FAIR principles & Openness

The *FAIR* principles[[14]](#footnote-15) (*Findable, Accessible, Interoperable, Reusable*) are important characteristics of high-quality data, but making data FAIR does not equate to ensuring they are of high quality.

Surveys by researchers[[15]](#footnote-16) and our own consultation with experts from across the research lifecycle highlights the importance of data quality characteristics not covered by the FAIR principles, and the value of emphasizing the distinction. Go-FAIR, an initiative that aims to implement the FAIR data principles, articulates the crucial distinction[[16]](#footnote-17) between other elements of data quality or ethics.

Similarly, data openness is an important principle, distinct from the FAIR principles. It is correlated to many of the characteristics we have highlighted, and it is also essential to the validation of quality.

We wish to separate debates around the need for data to be open and FAIR from critical reflection on how data quality is evaluated, and the extent to which that evaluation requires a localized assessment of the needs, means and goals of each research environment.

As such, we have tried to articulate the characteristics of data quality that can be considered independent of the *FAIR* principles and openness.

#### Metadata quality

Metadata describe and give information about other data. Fundamentally, the only difference between metadata and data are the modes of use[[17]](#footnote-18). We do not specify *metadata quality* as a separate characteristic in this framework. We consider it fundamental to many of the characteristics listed rather than a separable element in its own right.

Purpose-orientated metadata[[18]](#footnote-19) drive utilization, but different users of data require distinct types and levels of metadata. Metadata are not just for description and discovery, but also for contextualization and for coupling users, software, and computing resources to data. Metadata do not fully determine the destinations to which the data will travel, and so diverse levels of de-contextualization or re-contextualization[[19]](#footnote-20) are required to allow data to be reused in unpredictable circumstances.

Indeed, different subdisciplines in the same field may have substantially different ontological metadata requirements for the same data, to say nothing of different disciplines and fields. This raises the possibility of the need for:

* a single dataset to be available in different formats with different metadata.
* multiple sets of metadata to be available for the same dataset.
* standard transformations of data into new formats to be readily available.

We will return to this topic below, when considering the curation of data.

Awareness of the importance of metadata has grown, and metadata quality is the target of several quality control processes and mechanisms that we will go on to consider in the next section. Yet expectations of metadata are still under-emphasized, and the need for metadata to be multifaceted is underestimated.

Metadata are not sufficient in and of themselves to ensure high-quality data, but they are the determinant of data quality in terms of *Availability* and *Usability*. Additionally, the quality of metadata alongside the *Presentation* of the data determines our ability to judge whether data are *Relevant* or *Reliable*.

Metadata are the target of several of the quality control processes and mechanisms that we will consider in the next section.

## Processes and mechanisms for improving data quality

Processes and mechanisms are designed to ensure and assess the quality of data operate throughout the research lifecycle.

We have grouped these processes under separate headings to explain how they contribute to various aspects of data quality, drawing on interviews and existing research[[20]](#footnote-21). While the boundaries between processes are fuzzy, they capture real and distinct ways in which researchers and organizations segment processes.

We have not set out to capture a linear process. Different processes may fall to different organizations depending on where they sit in the research cycle, (a natural and appropriate ‘divvying up of the roles’[[21]](#footnote-22)) there are multiple ways and places in which processes operate and interact.

Quality characteristics are established at different points by different mechanisms. Some mechanisms depend on degrees of one quality characteristic to improve or assess another.

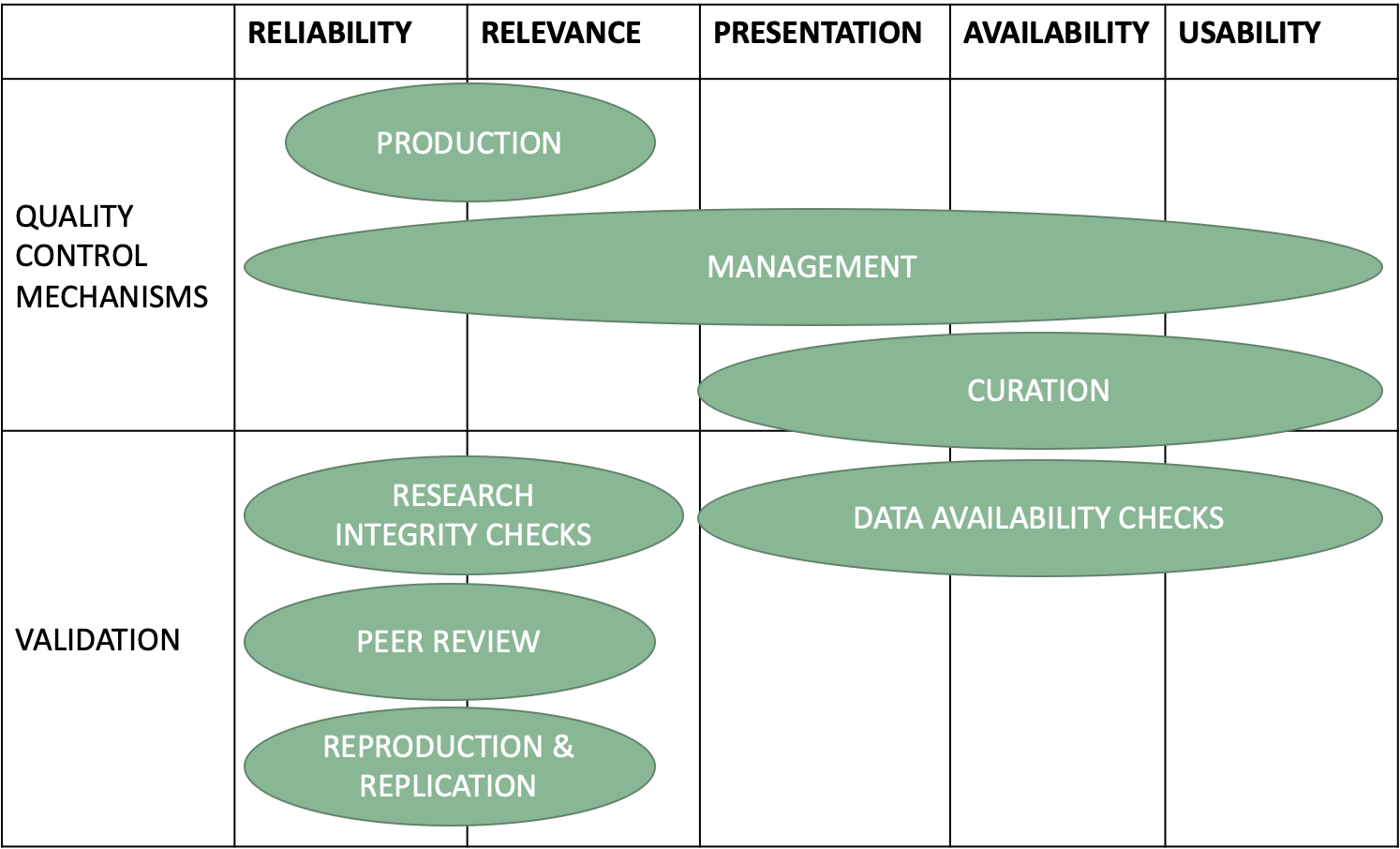


Figure 2: Quality control and validation processes and their relationship to quality characteristics

### Production

*Quality control mechanisms* are vital to generating and ensuring high-quality data. They are the responsibility of researchers, who define the research questions being asked, generate the data in question and determine the methods, materials and instruments used[[22]](#footnote-23). Controls are done by researchers themselves and often involve peers in the same or other research groups or faculties.

Such processes are primarily designed to avoid bias and error in data collection and production. There are common elements to processes that seek to ensure accuracy, completeness, and integrity of the data, but there will be a unique set of processes in operation for every individual experiment or data collection exercise, and there are enormous differences across subject areas and subdisciplines.

Examples of such processes include:

* calibration of measuring equipment
* refinement of applied algorithms
* review of experiment design by ethics committees or peers
* survey design and testing.

Researchers make choices about the type and depth of checks performed to ensure their data are of sufficiently high quality. Such choices are always resource dependent – more checks can always be performed – and researchers must make complex judgment calls to resolve tensions between different quality characteristics.

For example, longitudinal studies on patient health may seek to exclude data which are not reliable*[[23]](#footnote-24)* – incomplete or inconsistent – but may then inadvertently have an unrepresentative sample since data relating to poorly served or disadvantaged groups are more likely to be excluded.

In our framework, data production controls focus on data *Reliability* as well as *Relevance*, as far as the latter is concerned with the fitness of purpose to a defined research question or case of use.

In large-scale research projects collecting enormous quantities of data, such as projects in particle physics, meteorology, astronomy, and organism sequencing[[24]](#footnote-25), data are generated in the absence of a single research question or a necessarily well-defined case of use. Instead, anticipating data will be useful in multiple contexts.

### Management

*Data management* may be defined as the effective handling of information that is created during research[[25]](#footnote-26). It is taken to mean activities that might be described as ‘good housekeeping,’ such as version control, file storage and access requirements.

It also considers the documentation and full details of how data will be presented, made usable and available. Documentation is as essential as data; it effectively determines the quality of the data[[26]](#footnote-27) and the ability of reviewers and users to assess quality in other phases. Indeed, data management must operate throughout and beyond the production phase, but if the documentation is not written during or immediately after the production phase, it usually never gets done[[27]](#footnote-28).

Data managementtends to operate in research institutions since it involves the provision of support directly to researchers. Data management plans are the most familiar innovation developed to help researchers ensure quality. These plans help ensure data are managed efficiently, effectively, and securely, and that risks to data quality can be mitigated. A less familiar and commonly offered innovation regarding Data Stewards was pioneered in recent years by TU Delft in the Netherlands.

|  |
| --- |
| Case study – TU Delft  TU Delft initially began appointing Data Stewards in 2017 as a three-year experiment. The Executive Board of TU Delft recognized the importance of FAIR data and the fact that to put FAIR data into practice, data management experts must professionally support researchers. In the pilot, a dedicated Data Steward was appointed to each of TU Delft’s eight faculties and a Data Stewardship Coordinator was appointed to the library.  Data Stewards are data management professionals, who understand the research conducted in the faculty. They usually have a PhD degree, specialist data skills and specific experience in a particular subject area. At TU Delft, they are centrally coordinated, but are embedded in and employed by individual faculties. The research topics conducted at TU Delft are in diverse disciplines in science, engineering, and design. There are also many interdisciplinary research activities across engineering, social science, and humanities. The Data Stewards have a good understanding of the research conducted by their faculties, but they stay out of the actual design and operation of experiments, that is, they are fully removed from data production processes.  Demand for their services is so high, they do not have the capacity to provide firsthand assistance to researchers. Instead, they offer support at the faculty or research group level.  There are commonalities in the way the Data Stewards operate, but also differences. Some work in fields where personal data is commonly utilized and have expertise in confidentiality and access. Others work in fields heavily dependent on software to either generate data or as the main output of research, such as Mathematics or Electrical Engineering, and require more technical awareness.  The Data Stewardship pilot has been hugely successful; all the Data Stewards have been appointed permanent members of staff. Various other institutions and library consortia are planning to deploy similar models of data stewardship. |

Data management, as we define it, is distinct from both *Production* (see above) and *Curation* (see below). The boundaries are fuzzy, not fixed, but there are differences that make the distinction worthwhile, and they correspond closely to distinctions drawn by researchers and the teams supporting them.

Data production and data managementdiffer in how they deliver data quality. The former is about the *Reliability* and *Relevance* of data, while the latter ensures consideration is given to the *Presentation, Usability* and *Availability*. Data production is the preserve and responsibility of researchers, while data management may be the responsibility of a wider team, including data managers. Data management should involve the documentation of quality controls in operation during data production so that the data can be fully understood and assessed by others.

It is important to adopt good data management practices as early as possible in research design. It is much harder to fill in gaps in metadata later in the research workflow, whether by machine, author, or publisher, if data are missing from the beginning.

### Curation

*Data curation* involves the selection and appraisal of data, but also a wider range of functions[[28]](#footnote-29). It includes data storage, consideration of how data are organized and the creation and capture of metadata. This process ensures the accessibility and long-term preservation of data. Curation is not a ‘once and done’ process but an ongoing activity responding to changing technology, formats, and user needs.

Curation is naturally the preserve of repositories, which may be general or specialist according to the subject areas and disciplines they serve. They may be independent or hosted by an institution or other stakeholder, non-commercial or for profit. Different repositories offer a slightly different emphasis in the curation process, and we may distinguish between various levels of service.

*Hosting* – Some repositories offer a hosting service with no curation support beyond the submission process. Figshare or Zenodo offer this, as do some repositories based at research institutions. However, even the process of putting data into a repository can be seen as a form of quality control because the process will result in some metadata being automatically or manually added, for example, the assignment of a unique identifier. It can be seen as a guarantee of persistence.

*Curation* – This involves a review of the metadata, possibly including some checks of data reliability(e.g., looking for negative values or errors), and a review of the accessibility and structure of the data. Dryad and Figshare are general repositories that offer this type of curation. Institutional repositories offer it to varying degrees.

*Enhanced curation* – This encapsulates a range of technical or content-based[[29]](#footnote-30) enhancements operated by repositories. There may be an iterative process between researchers and a specialist data curator. It may involve ingestion of data into a larger dataset, facilitating findability and potential reuse, such as at PANGEA.

In terms of the data quality characteristics in our framework, curation is not responsible for ensuring that data are *Reliable,* but it facilitates assessments of the *Reliability* of datasets. Curationis directly responsible for increasing the degree of data quality in terms of *Presentation, Availability & Usability*.

|  |
| --- |
| Case study – PANGAEA  PANGAEA[[30]](#footnote-31) has an almost 30-year history as an open-access library for archiving, publishing, and disseminating georeferenced data from the Earth, Environmental and Biodiversity sciences. Originally evolving from a database for sediment cores, it is operated as a joint facility by the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI) and the Centre for Marine Environmental Sciences (MARUM) at the University of Bremen.  Each published dataset hosted by PANGAEA can be cited and holds a DOI. Metadata provided includes details of measurements, the methods and devices used for sampling and analysis as well as references to related literature and external resources. All data and metadata are compiled in close collaboration between the scientists and trained domain experts acting as data editors. Both data and metadata are checked for completeness and plausibility in compliance with the FAIR data principles. Semantic interoperability during data curation is ensured through terminologies following the international protocols and standards (i.e., terms of the World Register of Marine Species (WoRMS[[31]](#footnote-32)) and the Integrated Taxonomic Information System (ITIS[[32]](#footnote-33))). The review is complemented by inviting authors to proofread the datasets before the final publication.  PANGAEA utilizes sector-wide persistent identifiers (PIDs), such as ORCID iD, ROR, and Crossref Funder ID, as well as subject-related PIDs, such as the International Geo Sample Number (IGSN).  Grown from its origins as a paleoclimate archive, PANGAEA has specialized in tabular data that can be incorporated into its relational database system. In concert with the semantic harmonization, this improves the interoperability and reusability of data content and enables, among other things, meta-studies based on integrative access to identical parameters across multiple datasets.  The well-developed interoperability framework[[33]](#footnote-34) of PANGAEA allows most effective dissemination of metadata and data to all major internet search-engine registries, library catalogs, data portals, and other service providers, and ensures the optimal findability of data hosted by PANGAEA.  PANGAEA is a collaboration partner in numerous initiatives, projects, and consortia. It acts, to name only two, as an infrastructure partner for the German National Research Data Infrastructure initiative (NFDI) and the European Marine Observation and Data Network (EMODnet). In commitment to particular partnerships, PANGAEA provides specifically adapted data products for other repositories (such as Darwin core archives (DWCA) for GBIF and the OBIS network). PANGAEA has good relationships with publishers, with some recommending PANGAEA as the data repository. This has increased visibility, resulted in new features on both sides, and direct links to PANGAEA datasets. The successful cooperation between PANGAEA and the publishing industry along with the corresponding technical implementation enables the cross-referencing of scientific publications and datasets archived as supplements to these publications. It even provides the means to synchronize both editorial processes by allowing reviewers of articles to concurrently review respective data supplements. |

#### Specialist & general curation

Curation takes resources. It may be seen as the process that makes data accessible to research groups with potentially hugely different epistemic cultures[[34]](#footnote-35).

Specialist datasets are often focused on curating data for specific, known purposes. They may host only one or a very limited number of file types. Their curation will provide detailed checks on the file and its data and focus on detailed metadata requirements and descriptors of use within their specific subject area.

Generalist repositories have a broader focus, preparing data for potentially unlimited cases of use which cannot be anticipated. They facilitate the curation of a much wider set of file types and focus on more general metadata and checks to ensure data can be accessed and reused widely.

The different focuses create opportunities to enhance quality in distinct but complementary ways. Data may be hosted in different repositories, and undergo different curation processes at each, resulting in different metadata depending on the anticipated research use. Or data may be hosted in one repository and translated into a new format before being hosted at another. PANGEA operates partnerships that facilitate these types of processes.

Different layers of metadata and descriptors may be hosted at a single, general repository, facilitating not only general, unanticipatable reuse but also specific, defined ones. This is the focus of the Data Curation Primers designed by the Data Curation Network[[35]](#footnote-36). These are target-specific subjects, disciplinary areas or curation tasks that can be used as references to curate research data and ensure more specialist metadata requirements can be met by generalist staff or staff at generalist repositories.

The focus of specialist and general repositories may be understood in this context. In terms of our data quality framework, they attempt to make data fit for different purposes and ensure data can be used in diverse ways. The distinct types of repositories and their different curation processes create different opportunities to facilitate reuse of data.

### Validation

As curated data become available to a wider pool of stakeholders and (potential) users, a range of processes are undertaken to validate data quality. Our framework distinguishes several types of validation processes.

#### Validation of *Presentation, Availability & Usability:* Data availability checks

At a basic level, *Data availability checks* can be undertaken by non-subject-matter specialists and involve an assessment as to whether a minimum threshold has been met. They are conducted by publishers and funders who have set the expectations on minimum thresholds via their policies on data availability. These checks focus on whether data can be found and accessed. It might be categorized as assessing the FA in FAIR, or whether a dataset is FAIR for a human user.

Examples include the review of data availability statements in operation at publishers. One might also consider the automatic deposit of datasets offered by publishers with general repositories, like Figshare or Zenodo, as part of their workflow to fall into this category.

More detailed technical quality assessments of this type of data might require a subject matter specialist or even a specialist in aspects of data or statistics. These checks will focus on the quality of metadata, documentation and how well the data are described, sometimes including detailed review of software, code, and limitations of the data. These types of checks are focused on the IR in FAIRprinciples, or whether data can be considered FAIR for non-human use via artificial intelligence or machine learning.

Tools like DataSeer[[36]](#footnote-37) offer automation of this type of check. The F-UJI automated FAIR Data Assessment tool is an ambitious attempt to automate the assessment of whether datasets held within repositories meet the metrics associated with FAIR principles.

|  |
| --- |
| Case study – F-UJI tool  F-UJI[[37]](#footnote-38) is an automated assessment tool (‘F’ stands for fair, ‘UJI’ means test in Malay) that assesses the FAIRness of data objects. The tool was developed by Anusuriya Devaraju & Robert Huber from PANGAEA in the context of the FAIRsFAIR project. The tool evaluates datasets with several practical tests derived from 17 data object assessment metrics.  These 17 metrics are based on indicators proposed by the RDA FAIR Data Maturity Model Working Group, on the WDS/RDA Assessment of Data Fitness for Use checklist, and on prior work conducted by project partners such as FAIRdat and FAIREnough. The metrics were developed iteratively with consistent monitoring or mapping to the original FAIR data principles.  One focus point of the FAIRsFAIR project has been to support the assessment of (FAIR) data in (trustworthy) repositories. The work came about through the proposed data object assessment metrics and the development and piloting of assessment tools. It addressed three of the recommendations made by the European Commission in the *Turning FAIR into Reality* report:   * Rec. 8 (Facilitate automated processing) * Rec. 12: (Develop metrics for FAIR Digital Objects) * Rec. 25 (Implement FAIR metrics to monitor uptake)   FAIR-Aware is an online self-assessment tool to foster general awareness of FAIR practices in researchers, Data Stewards, and other data professionals.  F-UJI has been extensively and iteratively tested in six CoreTrustSeal-certified data repositories, one non-certified repository and in personal consultancy work. This was essential to assign meaning to outcomes of the automated assessment, and to gather feedback for further development and improvement. Failing a certain practical test does not always indicate an absence of evidence for the test, but it can indicate a failure of the tool to find or access the evidence. Solutions are not always unambiguous to identify.  F-UJI is accessible in several ways:   * Source code for the REST API implementation is available via Github at <https://github.com/pangaea-data-publisher/fuji> from where interested parties may download, install and run a F-UJI instance on their own servers, or in a Docker container. * A simple web demo client is available at <https://www.f-uji.net> which can be used for quick FAIR checks on individual data sets. This produces visualizations of the evaluations with details on the scores earned in a JSON file.   The F-UJI web client currently provides an implementation of a badging scheme consisting of the following levels:  0 – undefined  1 – initial  2 – moderate  3 – advanced  These scores are assigned for each practical test and then aggregated to one label per FAIR principle. HTML code snippets provide a means to embed the F-UJI badges in any website. Further steps need to be taken to refine this and other badging schemes and the governance around decision making, and there is a possibility that a network of FAIR-enabling Trustworthy Digital Repositories will play a role in assigning badges to datasets. |

#### Validation of *Reliability & Relevance: Research integrity checks* *& peer review*

While it is not possible to give a dataset a certification that says it is ‘scientifically sound’[[38]](#footnote-39), one can undertake assessments of data reliability via automated checks, computational assessments or by relying on the judgment of subject matter experts and peers.

Examples include *Research integrity* checklists run by a variety of publishers, statistical checks that look for suspicious patterns in data or which check for image manipulation. Some publishers utilize dedicated data scientists or specialist in-house experts to engage with specialist data sets and associated repositories.

Checks of this type consider whether the epistemic claims made by researchers match the data they are using. They consider the collection, analysis, and interpretation of data by researchers. This involves an assessment by specialists as to whether research is correctly contextualized and whether researchers have demonstrated appropriate awareness of prior studies. It is about whether a particular claim is valid, and whether the researchers have gone about trying to prove their claim in an appropriate manner. The data in themselves do not make any claims. This type of check is done via the review associated with a paper or preprint. As commonly understood, peer review of literature fits into this category.

Two types of validation need examining in more detail.

#### Peer review

Publishers’ editorial processes that support *article* quality may well or not reflect all the characteristics of *data* quality as we consider it in our framework.

The development of peer review in the past few decades has been well documented, particularly the increase in the volume of articles requiring review and growing burden on researchers. This limits the ability of peer reviewers to undertake checks on data that would fall into the first or second categories. The level of interaction that reviewers have with the data varies by discipline and journal.

Peer reviewers think primarily in terms of epistemic claims made in papers, but here there is an element of ‘the chicken and the egg’ as publishers and researchers do not expect to have to check data which, in the event, are not always available during the editorial process.

In addition, peer reviewers do not always undertake checks that fall into the first two categories (data reliability and integrity). A variety of editorial and automated checks concerned with both do and will continue to reduce the burden on peer reviewers, allowing them to focus on the *relevance* of data.

The term ‘data peer review’ is used in diversely in different contexts. It can refer to curation, elements of peer review or to the specific review of data done by peers in data journals (e.g., *Scientific Data*) which focus on the Presentation, Availability and Usability of data. Considering this ambiguity, we focus on the use of other terms.

|  |
| --- |
| Case study – T&F  In 2018, under the direction of their Open Research team Taylor & Francis (T&F) launched a data-sharing policy framework incorporating [multiple levels](https://authorservices.taylorandfrancis.com/wp-content/uploads/2019/04/Author-Services-Data-sharing-policies.pdf) varying according to the identifiers, licenses and approaches to data citation adopted, as well as the availability of data availability statements. Having a framework with multiple levels was necessary to reflect the varied nature of the subjects covered by the portfolio and the various levels of resources available to authors. The number of journals adopting a policy has increased by some 30–40% since 2018, and the team are currently working toward all relevant journals having a data-sharing policy of some kind.  After discussion and collaboration between T&F, their society partners, and relevant parts of the research community, some 10% of journals have moved to higher policy tiers.  In 2019, T&F introduced a range of support for journal editors adopting the higher tiers and more open data-sharing policies: Share Upon Request or higher. This includes checks on submission of the required data availability statements (DAS). For journals where data must be stored in a repository, workflows have been prepared to assist administrators run basic checks that the provided DAS meet the policy requirements.  The Open Research and Research Integrity teams at T&F have also begun reviewing the mechanics of making data available to editors and reviewers where ultimately this data will not be made open. The aim of this effort is to ensure that research papers complete the peer review process as the reviewers should have increased confidence in the research because they have access to the data created or collected by the author. This could be facilitated by some specialist and generalist repositories for datasets that will eventually be made openly available, but there is no general solution available at present. T&F are exploring the workflows and partnerships required to develop such functionality.  T&F have also adopted [Registered Reports](https://www.cos.io/initiatives/registered-reports), a rapidly growing initiative that has introduced another mechanism check for improving on data quality. As researchers split their research activity into two parts, with a round of peer review in the middle, it removes the temptation for authors to act unethically and increases the focus on the research undertaken rather than the findings. Given that authors are considering how they will manage their data before they start collecting, it should increase quality. Registered Reports published in journals with open data policies has the potential to dramatically introduce transparency and reproducibility.  [Eighteen T&F journals](https://authorservices.taylorandfrancis.com/peer-review/registered-reports) currently publish Registered Reports, with one requiring authors to follow the relevant process for all research articles. Most journals that have adopted this article type are in Psychology or aligned fields, reflecting the willingness to adopt new measures to address the reproducibility crisis in the field. T&F are actively working with research institutions and teams to develop the cultural shift in expectations that is required to ensure more researchers publish Registered Reports in future. |

#### Validation of *Reliability & Relevance: Reproduction & Replication* Reproducibility is a commonly discussed principle, particularly in terms of the management and curation of data and ensuring that research can, in principle, be reproduced through documentation and availability.

Reproduction is the more familiar term, used more consistently in relation to validation of results in research than replication. The two terms are used differently or interchangeably in different circumstances but nowadays replication is gaining more attention[[39]](#footnote-40). The Turing Way[[40]](#footnote-41) is working to delineate the terms more clearly, and we adopt their definitions here:

*Reproduction:* When the same results and epistemic claims are based on the same data and the same analysis.

*Replication*: When the same results and epistemic claims are made from different data and the same analysis.

Our purpose here is not to discuss whether data can *in principle* be reproduced or replicated, which is the intention of many initiatives concerned with data quality, but the actual validation of *Reliability* gained in respect of a particular question when research is actually reproduced or replicated. The importance of these processes is well understood but it is not sufficiently well embedded in connection with data quality.

Specialist repositories, sometimes in partnership with publishers, subject data to automated checks of *Reliability*, such as data analysis or computational interpretation. Sometimes this referred to as ‘running the data,’ as in the processes the Cambridge Crystallographic Data Centre (CCDC) operates when they make a dataset available. These automated checks represent data reproduction.

|  |
| --- |
| Case study – CCDC  The Cambridge Structural Database (CSD) was established in 1965 and is the international data repository for high-quality curated crystal structure data. It is managed and maintained by the Cambridge Crystallographic Data Centre (CCDC), a UK independent non-profit. More than 70,000 datasets are deposited each year, many associated with articles published in scientific journals.  Originally data was manually retyped from the literature or supporting information and subject to quality control procedures. Now, data are directly deposited by researchers via a web-based deposition process.  The deposition process ensures files are in a recognized standard format (CIF), that files are syntactically correct, and provides opportunities for additional metadata to be provided.  Importantly, the depositor can generate a validation report which assesses the completeness and consistency of data in a CIF and generates alerts of varying severity. This report is generated by ‘checkCIF’ and informs reviewers and future consumers of the data about quality aspects of the structure determination. The service is hosted by the IUCr and sponsored by the community, including publishers.  Following publication of a structure, the data are subjected to closer inspection by the CCDC before the structure is added to the CSD. The aim here is to enrich the deposited data so that they can be readily reused by chemists, materials scientists, structural biologists and others in the discovery and validation of new science.  The CCDC employs a data-integrity expert who develops and runs additional checks to identify inconsistencies that might deserve closer scrutiny, be they deliberate or unintentional.  Multiple connection points between publisher and CCDC occur throughout the workflow:   * Most major journals require crystal structures to be deposited with the CCDC prior to submission of a manuscript. * Many journals require a checkCIF to be run for structures associated with submitted articles and some insist that more severe alerts are addressed or explained. * Prior to publication, structures are made available confidentially to journal editors and referees so data can be assessed and used to validate claims in the article. * Links are established between article and dataset through Scholix and other means to enable access to data underpinning research. * Publication of a deposited structure is often triggered by publication of an associated article. When this occurs, datasets, including any checkCIF reports, become freely available at the CCDC for anyone to download. * CCDC’s curation and scrutiny processes uncover inconsistencies which bring into doubt the quality of a published dataset or article. Such situations are rare, but when encountered they are raised with the publisher and may lead to a correction or retraction. |

## Areas of focus

The *Presentation, Availability and Usability* ofdata have advanced a great deal through initiatives to enshrine FAIR principles and make data openly available. Yet more still needs to be done and STM is continuing to work on sharing, linking, and citing data through the Research Data Program.

Researchers have always been focused on the *Reliability* and *Relevance* of their data. Now more is required on the industry and system levels to consider these aspects of quality and how they intersect with *Availability, Presentation and Usability*. This situation reflects the ‘start of our journey’ of transition from narrative-based to data-dependent science.

The distributed nature of infrastructures that exist to produce, share, assess and improve high-quality data make it impossible for any one group to retain responsibility for data quality[[41]](#footnote-42). Specific data quality characteristics are not the responsibility of specific agents at specific points in the research lifecycle.

Table 2 plots the various activities concerned with research data quality related to key participants in the research ecosystem as we understand them from our consultations and case studies.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Involved in stages of the research cycle* | *Involved in processes related to data quality* | *Tools/resources applied* |
| *Researchers* | proposal writing, conduct research process (simulate, experiment, observe, manage data, analyze data, share data), publication | production, reproduction & replication, management | DMPs (data management plans), data software tools, repositories, publishing platforms |
| *Funders* | proposal writing | management | DMPs, repositories |
| *Institutions* | research process (simulate, experiment, observe, manage data, analyze data, share data) | production, management | DMPs, data software tools, institutional repositories, Data Stewards |
| *Repositories* | research process (analyze data, share data) | curation, peer review, data availability checks | curation tools, data checking tools, DataCite |
| *Publishers* | publication | research integrity checks, peer review, data availability checks | repositories, Scholix, DataCite, data availability statements, journal data policies |
| *Reviewers* | publication | peer review, reproduction & replication | repositories, publishing platforms |

*Table 2: Data quality research activities related to key participants in the research ecosystem. Research cycle taken from Tenopir, 2011[[42]](#footnote-43).*

Different entities have and will innovate with a focus on different quality control mechanisms and characteristics depending on where they operate in the research lifecycle. Data quality characteristics are interconnected and so joint initiatives, cross-sectoral partnerships and the critical roles played by organizations and initiatives such as the RDA, CODATA, FAIRsFAIR, and umbrella bodies like STM will be at the heart of drives to improve data quality.

The following are potential areas of focus and future development in the light of our consideration of data quality. Priorities and specific directions will emerge from the work of relevant stakeholders.

Simpler processes

* Continued investment in automated tools by publishers and repositories will help assess the *Reliability* of data and aid peer reviewers in their assessments. More partnerships and more co-development will surely emerge to facilitate this.
* More sophisticated workflows are required to facilitate all forms of editorial checking and peer review. Integration with such repositories as PANGEA or Dryad allow open review and interaction with data. It will ensure that data are accessible early in the editorial processes and that all forms of peer review can be respected. This requires considerable coordination between multiple parties in scholarly communications.
* More routine and sophisticated checks by editorial teams on the *Availability*, *Usability* and *Presentation* of data will reduce the burden on all parties by implementing more automated checks like DataSeer and F-UJI.

More support

* Enriched support for researchers put in place early in the research process will reduce the administrative burden researchers feel and accelerate cultural change. More institutions and national bodies are exploring and will introduce initiatives like the Delft model. Innovations in data management and support for researchers will arise from opportunities for networking and shared best practice – e.g., initiatives like the Data Curation Network[[43]](#footnote-44) and direct interaction between researchers, specialist curators and publishers such as those developed by PANGEA and CCDC.
* More dedicated funds will be made available by research funders enabling researchers to pay for third party curation and data management support where necessary.

Better tracking of use and appropriate assigning of credit

* More consideration will be given to how to embed and balance data and data quality in the reform of research assessment. FAIR and open data are a part of this discussion, but wider considerations of quality are critical.
* Fuller implementation of policies ensuring datasets are linked to and cited by relevant research will ensure more and better credit for researchers as these datasets can be counted as citations. Publishers still do not implement these practices consistently, despite the work of such organizations as Make Data Count, DataCite, STM and RDA[[44]](#footnote-45). The work of Make Data Count to develop metrics on research data will be valuable, and the consistent capture of reuse, replication & reproduction in metadata will represent a crucial step.
* More partnerships between entities curating data to ensure smooth transformation and appropriate credit tracking across distributed infrastructure. This will include more partnerships such as those between specialist repositories and/or generalist repositories of the type employed by PANGEA to enable reuse via co-hosting of data, multiple metadata schema for the same data, or routine transformation of data for new purposes.

More research and discussion

* More partnerships and collaborative research between repositories, institutions and publishers will generate better insight into researcher needs and priorities. Publishers have large pools of direct relationships with researchers which they can survey. Researchers may lack the capacity[[45]](#footnote-46) to provide feedback to curators and data managers on how well systems serve their research or the quality of datasets available to them.
* The more discussion and broader debate, and the more organizations centralizing data quality in their thinking to ensure high-quality data, the more adapted to data-driven research the scientific enterprise will be. More formal consideration of data quality as a concept or data quality characteristics in relation to research will emerge.

We hope this Research Paper offers a valuable perspective on the ongoing transition to data-dependent science. We also hope that our framework and the highlighted initiatives will inform current debate on how to ensure data quality.

## 

**References**

Assante, M., Candela, L., Castelli, D. and Tani, A., 2016. Are scientific data repositories coping with research data publishing? *Data science journal*, *15*.

<https://datascience.codata.org/articles/10.5334/dsj-2016-006/>

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI:

<http://doi.org/10.5334/dsj-2015-002>

Callaghan, S., 2019. Research Data Publication: moving beyond the metaphor. *Data Science Journal*, *18*(1).

<https://datascience.codata.org/articles/10.5334/dsj-2019-039/>

Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M.S. and Whyte, A., 2013. Processes and procedures for data publication: A case study in the geosciences. *International Journal of Digital Curation*, *8*(1), pp.193-203.

<http://www.ijdc.net/article/view/8.1.193>

Currie, A. and Kilbride, W., 2021. FAIR Forever? Accountabilities and Responsibilities in the Preservation of Research Data. *International Journal of Digital Curation*

​​DOI: <https://doi.org/10.2218/ijdc.v16i1.768>

Data Management Association (DAMA), 2013. The Six Primary Dimensions for Data Quality Assessment. Defining Data Quality Dimensions *Data Management Association (DAMA) UK Working Group on “Data Quality Dimensions”*

Ford, E., Rooney, P., Hurley, P., Oliver, S., Bremner, S. and Cassell, J., 2020. Can the use of Bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? Development of a novel method using simulated and real-life clinical data. *Frontiers in public health*, *8*, p.54.

<https://www.frontiersin.org/articles/10.3389/fpubh.2020.00054/full>

Hrynaszkiewicz, I., 2019. Publishers’ responsibilities in promoting data quality and reproducibility. In *Good research practice in non-clinical pharmacology and biomedicine* (pp. 319-348). Springer, Cham.

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. and Goudie, S., 2020. Developing a research data policy framework for all journals and publishers. *Data Science Journal*, *19*(1).

<https://datascience.codata.org/article/10.5334/dsj-2020-005/>

Hrynaszkiewicz, I., Harney, J. and Cadwallader, L., 2021. A survey of researchers’ needs and priorities for data sharing. *Data Science Journal*, *20*(1).

<https://datascience.codata.org/article/10.5334/dsj-2021-031/>

Ingram, C., 2019 How and why you should manage your research data: a guide for researchers (updated). *JISC*.

<https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data>

Jeffrey, K and Koskela, R, 2014. RDA Metadata Principles and their Use. *Research Data Alliance (RDA) Metadata Standards Dictionary* <https://www.rd-alliance.org/system/files/documents/RDA%20Metadata%20Principles%20and%20their%20Use%2020141114.docx>

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., McGeary, T.M. and Hull, E., 2018. Data curation network: A cross-institutional staffing model for curating research data. *International Journal of Digital Curation*

<http://www.ijdc.net/article/view/616>

Johnston, L., Hannah, H. and Blake, M., 2020. Extending the Research Data Toolkit: Data Curation Primers. *International Journal of Digital Curation*

<http://www.ijdc.net/article/view/713>

Sebastian-Coleman, L. 2013 Section 2. DQAF Concepts and Measurement Types, Editor(s): Sebastian-Coleman, L. In *MK Series on Business Intelligence, Measuring Data Quality for Ongoing Improvement*, Kaufmann, M. Pages 55-56, ISBN 9780123970336, <https://doi.org/10.1016/B978-0-12-397033-6.00036-5>.

Lee, C.A. and Tibbo, H.R., 2007. Digital curation and trusted repositories: steps toward success. *Journal of Digital Information*, *8*(2).

<https://jodi-ojs-tdl.tdl.org/jodi/index.php/jodi/article/view/229>

Leonelli, S., 2016. Data-centric biology. In *Data-Centric Biology*. University of Chicago Press

Leonelli, S., 2017. Global Data Quality Assessment and the Situated Nature of" Best" Research Practices in Biology. *Data Science Journal*

<https://datascience.codata.org/articles/10.5334/dsj-2019-039/>

Leonelli, S. 2020. Scientific Research and Big Data, *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.) <https://plato.stanford.edu/archives/sum2020/entries/science-big-data/>

Lowenberg, D., Lammey, R., Jones, M.B., Chodacki, J. and Fenner, M. 2020. Data Citation: Let’s Choose Adoption Over Perfection. *Make Data Count Blog*. 19 April 2021

DOI: <https://doi.org/10.5281/zenodo.4701079>

Puebla, Iratxe, Lowenberg, Daniella and FORCE11 Research Data Publishing Ethics Working Group. (2021) Joint FORCE11 & COPE Research Data Publishing Ethics Working Group Recommendations for the handling of ethical concerns relating to the publication of research data. *Joint FORCE11 & COPE Research Data Publishing Ethics Working Group*, 31pp.

DOI: <https://doi.org/10.5281/zenodo.5391293>

Ramasamy, A. and Chowdhury, S., 2020. Big data quality dimensions: a systematic literature review. *JISTEM-Journal of Information Systems and Technology Management*, *17*

<https://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/?lang=en>

Seo, S. and Kim, J., 2020. Data journals: types of peer review, review criteria, and editorial committee members’ positions. *science editing*, *7*(2), pp.130-135.

<https://www.escienceediting.org/upload/kcse-207.pdf>

Shankar, K., Jeng, W., Thomer, A., Weber, N. and Yoon, A., 2021. Data curation as collective action during COVID-19. *Journal of the Association for Information Science and Technology*, *72*(3), pp.280-284.

<https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.24406>

The Turing Way Community, Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R. Whitaker, K. 2019. *The Turing Way: A Handbook for Reproducible Data Science (Version v0.0.4).* Zenodo.

<http://doi.org/10.5281/zenodo.3233986>

University of California Curation Center (UC3), 2014. Fifteen ideas about data validation (and peer review) *University of California Curation Center (UC3) blog*, 8 May 2014 <https://uc3.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review/> (Accessed 29 Jun 2022)

Van Nederpelt, P. & Black, A., 2020. Dimensions of Data Quality (DDQ) - A Research Paper, *Data Management Association International*

<http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf>

Waaijers, L. and van der Graaf, M., 2011. Quality of research data, an operational approach. *D-Lib magazine*, *17*(1/2).

doi:10.1045/january2011-waaijers

1. Throughout, we take ‘scientific’ to refer to the areas covered by STM’s members, across scientific and technical, medical, and social sciences and humanities fields. [↑](#footnote-ref-2)
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, p.2. DOI:

   <http://doi.org/10.5334/dsj-2015-002> [↑](#footnote-ref-3)
3. See the following references:

   Data Management Association (DAMA), 2013. The Six Primary Dimensions for Data Quality Assessment. Defining Data Quality Dimensions *Data Management Association (DAMA) UK Working Group on “Data Quality Dimensions”*

   Van Nederpelt, P. & Black, A., 2020. Dimensions of Data Quality (DDQ) - A Research Paper, *Data Management Association International*

   <http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf>

   Ramasamy, A. and Chowdhury, S., 2020. Big data quality dimensions: a systematic literature review. *JISTEM-Journal of Information Systems and Technology Management*, *17*

   <https://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/?lang=en>

   Laura Sebastian-Coleman, 2013 Section 2. DQAF Concepts and Measurement Types, Editor(s): Laura Sebastian-Coleman, In MK Series on Business Intelligence, Measuring Data Quality for Ongoing Improvement, Morgan Kaufmann, Pages 55-56, ISBN 9780123970336, <https://doi.org/10.1016/B978-0-12-397033-6.00036-5>. [↑](#footnote-ref-4)
4. Van Nederpelt, P. & Black, A., 2020. Dimensions of Data Quality (DDQ) - A Research Paper, *Data Management Association International*

   http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf [↑](#footnote-ref-5)
5. Ramasamy, A. and Chowdhury, S., 2020. Big data quality dimensions: a systematic literature review. *JISTEM-Journal of Information Systems and Technology Management*, *17*

   <https://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/?lang=en> [↑](#footnote-ref-6)
6. Ramasamy, A. and Chowdhury, S., 2020. Big data quality dimensions: a systematic literature review. *JISTEM-Journal of Information Systems and Technology Management*, *17*

   <https://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/?lang=en> [↑](#footnote-ref-7)
7. See the following references:

   Data Management Association (DAMA), 2013. The Six Primary Dimensions for Data Quality Assessment. Defining Data Quality Dimensions *Data Management Association (DAMA) UK Working Group on “Data Quality Dimensions”*

   Van Nederpelt, P. & Black, A., 2020. Dimensions of Data Quality (DDQ) - A Research Paper, *Data Management Association International*

   <http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf>

   Ramasamy, A. and Chowdhury, S., 2020. Big data quality dimensions: a systematic literature review. *JISTEM-Journal of Information Systems and Technology Management*, *17*

   <https://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/?lang=en>

   Laura Sebastian-Coleman, 2013 Section 2. DQAF Concepts and Measurement Types, Editor(s): Laura Sebastian-Coleman, In MK Series on Business Intelligence, Measuring Data Quality for Ongoing Improvement, Morgan Kaufmann, Pages 55-56, ISBN 9780123970336, <https://doi.org/10.1016/B978-0-12-397033-6.00036-5>. [↑](#footnote-ref-8)
8. Leonelli, S., 2017. Global Data Quality Assessment and the Situated Nature of" Best" Research Practices in Biology. *Data Science Journal*

   <https://datascience.codata.org/articles/10.5334/dsj-2019-039/> [↑](#footnote-ref-9)
9. The Turing Way Community, Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R. Whitaker, K. 2019. *The Turing Way: A Handbook for Reproducible Data Science (Version v0.0.4).* Zenodo.

   <http://doi.org/10.5281/zenodo.3233986> [↑](#footnote-ref-10)
10. Leonelli, S., 2017. Global Data Quality Assessment and the Situated Nature of" Best" Research Practices in Biology. *Data Science Journal*

    <https://datascience.codata.org/articles/10.5334/dsj-2019-039/> [↑](#footnote-ref-11)
11. University of California Curation Center (UC3), 2014. Fifteen ideas about data validation (and peer review) *University of California Curation Center (UC3) blog*, 8 May 2014 <https://uc3.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review/> (Accessed 29 Jun 2022) [↑](#footnote-ref-12)
12. Van Nederpelt, P. & Black, A., 2020. Dimensions of Data Quality (DDQ) - A Research Paper, *Data Management Association International*

    <http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf> [↑](#footnote-ref-13)
13. Hrynaszkiewicz, I., 2019. Publishers’ responsibilities in promoting data quality and reproducibility. In *Good research practice in non-clinical pharmacology and biomedicine* (pp. 319-348). Springer, Cham. [↑](#footnote-ref-14)
14. <https://www.go-fair.org/fair-principles/> [↑](#footnote-ref-15)
15. Waaijers, L. and van der Graaf, M., 2011. Quality of research data, an operational approach. *D-Lib magazine*, *17*(1/2).

    doi:10.1045/january2011-waaijers [↑](#footnote-ref-16)
16. <https://www.go-fair.org/resources/faq/what-fair-is-not/> [↑](#footnote-ref-17)
17. Jeffrey, K and Koskela, R, 2014. RDA Metadata Principles and their Use. *Research Data Alliance (RDA) Metadata Standards Dictionary* <https://www.rd-alliance.org/system/files/documents/RDA%20Metadata%20Principles%20and%20their%20Use%2020141114.docx> [↑](#footnote-ref-18)
18. Assante, M., Candela, L., Castelli, D. and Tani, A., 2016. Are scientific data repositories coping with research data publishing? *Data science journal*, *15*.

    https://datascience.codata.org/articles/10.5334/dsj-2016-006/ [↑](#footnote-ref-19)
19. Leonelli, S., 2016. Data-centric biology. In *Data-Centric Biology*. University of Chicago Press. P25 [↑](#footnote-ref-20)
20. Waaijers, L. and van der Graaf, M., 2011. Quality of research data, an operational approach. *D-Lib magazine*, *17*(1/2). doi:10.1045/january2011-waaijers [↑](#footnote-ref-21)
21. University of California Curation Center (UC3), 2014. Fifteen ideas about data validation (and peer review) University of California Curation Center (UC3) blog, 8 May 2014 <https://uc3.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review/> (Accessed 29 Jun 2022) [↑](#footnote-ref-22)
22. Leonelli, S., 2017. Global Data Quality Assessment and the Situated Nature of" Best" Research Practices in Biology. *Data Science Journal*

    <https://datascience.codata.org/articles/10.5334/dsj-2019-039/> [↑](#footnote-ref-23)
23. Ford, E., Rooney, P., Hurley, P., Oliver, S., Bremner, S. and Cassell, J., 2020. Can the use of Bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? Development of a novel method using simulated and real-life clinical data. *Frontiers in public health*, *8*, p.54.

    <https://www.frontiersin.org/articles/10.3389/fpubh.2020.00054/full> [↑](#footnote-ref-24)
24. Leonelli, S., 2016. Data-centric biology. In *Data-Centric Biology*. University of Chicago Press. P18 [↑](#footnote-ref-25)
25. Ingram, C., 2019 How and why you should manage your research data: a guide for researchers (updated). *JISC*.

    <https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data> [↑](#footnote-ref-26)
26. Waaijers, L. and van der Graaf, M., 2011. Quality of research data, an operational approach. *D-Lib magazine*, *17*(1/2).

    doi:10.1045/january2011-waaijers [↑](#footnote-ref-27)
27. Waaijers, L. and van der Graaf, M., 2011. Quality of research data, an operational approach. *D-Lib magazine*, *17*(1/2).

    doi:10.1045/january2011-waaijers [↑](#footnote-ref-28)
28. Lee, C.A. and Tibbo, H.R., 2007. Digital curation and trusted repositories: steps toward success. *Journal of Digital Information*, *8*(2).

    <https://jodi-ojs-tdl.tdl.org/jodi/index.php/jodi/article/view/229> [↑](#footnote-ref-29)
29. Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M.S. and Whyte, A., 2013. Processes and procedures for data publication: A case study in the geosciences. *International Journal of Digital Curation*, *8*(1), pp.193-203.

    <http://www.ijdc.net/article/view/8.1.193> [↑](#footnote-ref-30)
30. <https://www.pangaea.de/> [↑](#footnote-ref-31)
31. <https://www.marinespecies.org/> [↑](#footnote-ref-32)
32. <https://www.itis.gov/> [↑](#footnote-ref-33)
33. <https://www.pangaea.de/about/services.php> [↑](#footnote-ref-34)
34. Leonelli, S., 2016. Data-centric biology. In *Data-Centric Biology*. University of Chicago Press. P16 [↑](#footnote-ref-35)
35. <https://datacurationnetwork.org/outputs/data-curation-primers/> [↑](#footnote-ref-36)
36. <https://dataseer.ai/how-it-works/> [↑](#footnote-ref-37)
37. <https://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool> [↑](#footnote-ref-38)
38. Assante, M., Candela, L., Castelli, D. and Tani, A., 2016. Are scientific data repositories coping with research data publishing? *Data science journal*, *15*.

    <https://datascience.codata.org/articles/10.5334/dsj-2016-006/> [↑](#footnote-ref-39)
39. https://i4replication.org/index.html [↑](#footnote-ref-40)
40. https://the-turing-way.netlify.app/welcome.html [↑](#footnote-ref-41)
41. Leonelli, S. 2020. Scientific Research and Big Data, *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.) <https://plato.stanford.edu/archives/sum2020/entries/science-big-data/> [↑](#footnote-ref-42)
42. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101 [↑](#footnote-ref-43)
43. <https://datacurationnetwork.org/> [↑](#footnote-ref-44)
44. https://www.rd-alliance.org/groups/data-citation-wg.html&sa=D&source=docs&ust=1656513186172669&usg=AOvVaw2b6HUVEVsN15OOBAH1O2Rc [↑](#footnote-ref-45)
45. Leonelli, S., 2016. Data-centric biology. In *Data-Centric Biology*. University of Chicago Press. P37 [↑](#footnote-ref-46)