Trustworthy or not? Research data on COVID-19 in data repositories

18

Otmane Azeroual[a] *and Joachim Schöpfel*[b]
[a]German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany, [b]GERiiCO Laboratory, University of Lille, Lille, France

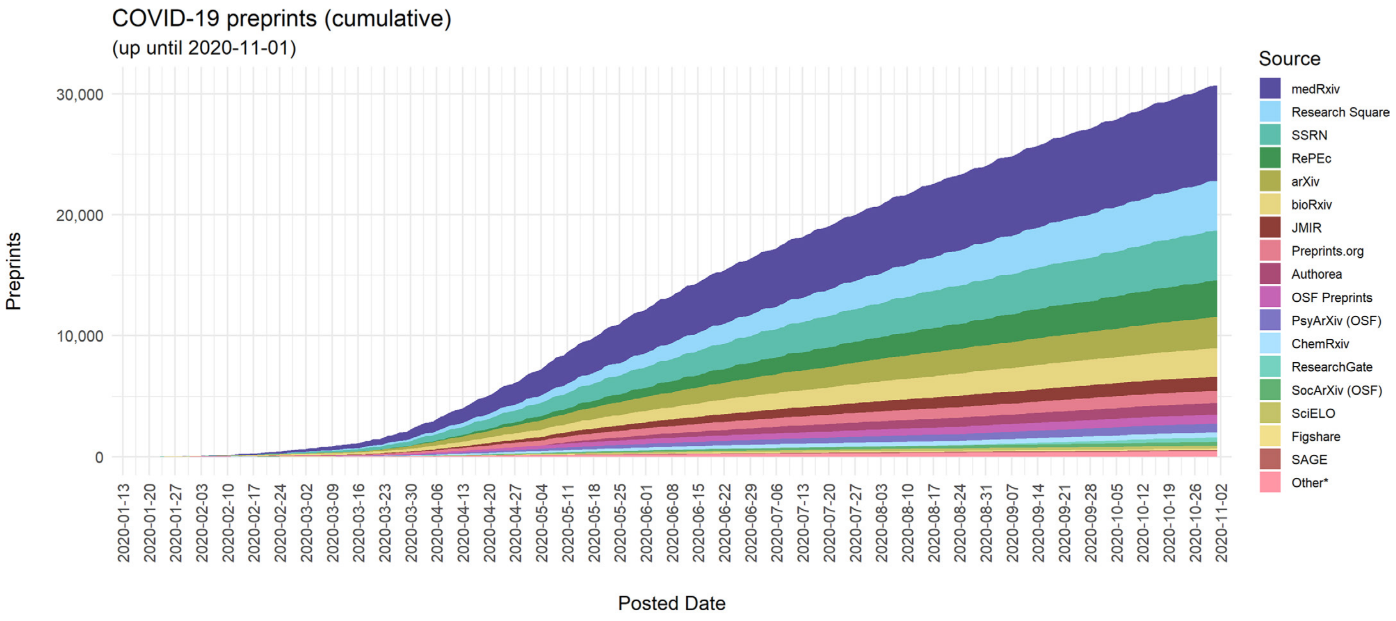
1 Acceleration, quality, and trust

The first case of the COVID-19 (or coronavirus) pandemic was identified in Wuhan, China, in December 2019. Ten months later, when starting to write this chapter (October 22, 2020), the World Health Organization (WHO) dashboard announces 41,104,946 confirmed cases and 1,128,325 deaths. The pandemic has become a major economic, social, and health policy priority in many countries, and the research is soon expected to provide insights and results relating to the development and innovation of new treatment protocols, drugs, and vaccines. Referenced by Google Scholar, about 108,000 papers have already been published on COVID-19, and likely 10 times more on related topics. The US COVID-19 Open Research Dataset (CORD-19), created by the Allen Institute for AI in partnership with Microsoft, IBM, the National Library of Medicine, and others, in coordination with The White House Office of Science and Technology Policy, contains over 100,000 research papers with full text about COVID-19, SARS-CoV-2, and related coronaviruses, provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights.

Most of the papers on COVID-19 are open access (Aristovnik et al., 2020; Arrizabalaga et al., 2020). Additionally, an important number of preprints—preliminary reports that have not been peer-reviewed—have been posted on preprint servers such as medRxiv, with nearly 8000 articles, and bioRxiv, with more than 2000 articles (Fig. 18.1). Acceleration, speeding up research and innovation, is one purpose of open science (Haider, 2018). The Budapest Open Access Initiative declared in 2001 that "removing access barriers to [peer-reviewed journal literature] will accelerate research… make this literature as useful as it can be and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge" (BOAI, 2001).

The trouble with acceleration is selection and quality. After noting that the server is receiving many new papers on coronavirus SARS-CoV-2, bioRxiv reminds (and warns) that these preprints "should not be regarded as conclusive, guide clinical practice/health-related behaviour, or be reported in news media as established information" (BioRxiv, 2020). Bypassing the "lengthy process" of peer-reviewing and the "reduction of quality control can lead to the spreading of misinformation creating

Libraries, Digital Information, and COVID. https://doi.org/10.1016/B978-0-323-88493-8.00027-6



* 'Other' refers to preprint repositories containing <50 total relevant preprints. These include: AfricArXiv (OSF), AgriXiv (OSF), BioHackrXiv (OSF), Cambridge University Press, Copernicus GmbH, EcoEvoRxiv (OSF), EdArXiv (OSF), engrXiv (OSF), ESSOAR, Frenxiv (OSF), INA-Rxiv (OSF), IndiaRxiv (OSF), LawArXiv (OSF), MediArXiv (OSF), MetaArXiV (OSF), NutriXiv (OSF), ScienceOpen, SportRxiv (OSF), Techrxiv (IEEE), WHO, Zenodo.

Fig. 18.1 COVID-19 preprints (cumulative).
Source: Nicholas Fraser, COVID-19 preprints https://github.com/nicholasmfraser/covid19_preprints.

additional problems that could originally be addressed during the peer-reviewing procedure" (Rios et al., 2020). The issue is, as mentioned above, the absence of peer review or some other form of quality control and selection.

The same observation can be made regarding the second pillar of open science: the open access to research data, "defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings" (OECD, 2007). Since the start of the coronavirus pandemic, a growing number of data is shared on different kinds of platforms. The COVID-19 data repository hosted by Figshare contains 2109 items, with 877 datasets. Elsevier's Mendeley Data search engine retrieves 11,314 resources on COVID-19 in data repositories, like Zenodo (2141), Mendeley Data repository (323), Harvard Dataverse (191), Apollo Cambridge (178), or the Robert Koch Institut repository at the Humboldt University, Berlin (69). The real number of COVID-19 datasets is hard to estimate; among the retrieved items are software, posters, and other research materials, and much data are not available on data repository because the research is still in progress or because of legal, industrial, or other reasons.

The problem with research data sharing is that data repositories most often do a superficial check of deposits; they try to ensure that the content is research data and not "rubbish" but do not provide a deeper assessment and moderation of the intrinsic, scientific quality and value of the data. The case of the controversy over hydroxychloroquine (HCQ) and the Surgisphere papers (Piller, 2020) shows that uncontrolled sharing of unreliable, early, and nonvalidated findings is not only an academic problem but also, at least potentially, a societal issue and a matter of passionate and engaged debate in the media and by politicians. Belief in misinformation about COVID-19 poses a potential risk to public health; therefore, scientists play a key role as disseminators of factual and reliable information (Roozenbeek et al., 2020).

Yet, on the other hand, Besançon et al. (2020) highlight that the lack of data sharing or third-party reviewing has led to the retraction of four major papers and had a direct impact on the study design and conduct of international trials. A review of 689 clinical trials regrets a lack of quality, coordination, and research synergies and, moreover, the lack of systematic sharing of trial data for the generation of evidence and metaanalyses (Janiaud et al., 2020). The issue is not too much data versus not enough data. The underlying question is about trust: How to ensure the quality of freely available research results? What is or could be done to increase the data's trustworthiness?

2 Control and assessment of data deposits

The quality of datasets is generally addressed using the dimensions of accuracy, currency, completeness, and consistency (Batini et al., 2009). These dimensions are not specific for research data. What is specific and special with research data has been described as their "contextual quality" (Stausberg et al., 2019, following Wang and Strong, 1996), because of the particular and often dynamic nature of datasets in given discipline environments, communities, and infrastructures; especially in research fields that produce unstructured and semistructured data, manual data quality checks are considered an important safeguard against fraud (Konkiel, 2020).

Research data infrastructures and, in particular, research data repositories play an essential role in this issue, insofar as their main function is preservation and dissemination. In the context of open science, data repositories are a key element in the deposit and sharing of datasets. Today, there are many and very different data repositories, disciplinary, institutional, governmental, and other platforms, some covering a large spectrum of research fields, while others focused on a particular topic, community, equipment, or material. According to international directory re3data, 1438 research data repositories provide some kind of quality management—about 55% of all registered platforms. But only 79 repositories (3%) declare certified procedures to ensure data quality.

Often, as mentioned above, ingestion control is light. Mendeley Data, for instance, provides manual checking for all posted datasets to ensure the content constitutes research data (raw or processed experimental or observational data), is scientific in nature, and does not only contain a previously published research article. Spam or nonresearch data are rejected but there is no validation or curation of the contents of valid research datasets (Haak et al., n.d., forthcoming). Another example is the Inter-university Consortium for Political and Social Research repository (ICPSR) which performs manual data and documentation (metadata) quality checks as part of the data deposit process, rejecting deposits with inadequate documentation and/or of poor quality.

Academic journals have the potential to contribute to the quality check of datasets. At *Cell Press*, for instance, the peer review process includes looking at the data. If an author publishes an article in *Cell Press*, the associated data are then also asked for, and the editors and reviewers look at the data and check for the data quality. The journal *Cell* has published hundreds of papers on the COVID-19, one part of them (like Schulte-Schrepping et al., 2020) along with reviewed and validated datasets.

This seems to be an exception, however. Following the "Transparency and Openness Promotion" initiative of the Center for Open Science, less than 5% of the already evaluated and registered journals mention peer-reviewing of deposited datasets. An editorial of the *International Journal of Cardiovascular Sciences* states that "most of the time, reviewers do not examine the raw data of the studies they review," adding that "one of the multiple benefits of Open Science is that research data can be checked by anyone who accesses the data repository, thereby reducing the likelihood of scientific misconduct" (Mesquita, 2020). In other words, academic journals, at least, should ask for (if not require) the deposit of datasets.

Data journals and, more generally, data papers are a new way to publish data and information about data and to ensure a certain level of peer review of deposited and shared data. Quality control of data papers—some kind of peer review—always implies an evaluation of the datasets themselves and their respective repositories. But for the moment, this new way of academic publishing represents a very small and marginal part of the overall research output (Schöpfel et al., 2019).

3 What makes data trustworthy?

The purpose of data sharing is reuse. Reuse requires some kind of guarantee of the data's integrity, authenticity, and quality. When this guarantee is missing, in the absence of evaluation and quality control, how does one trust research data? What makes research data trustworthy?

The crucial but not the only variable of trustworthiness is the quality of infrastructure, the data repository; the two other variables are the quality of the underlying research and, of course, of the data itself (Fig. 18.2).

The quality of the underlying research process can be described in terms of research ethics, as "doing good science in a good manner" (DuBois and Antes, 2018). Good science means research conducted according to common standards of excellence, while good manners include, among others, appropriate data storage, management of conflicts of interest, protection of human participants and animal subjects, honest reporting of findings, and proper citation of sources. In the field of COVID-19, the suspicion of conflict of interest is one of the major concerns in the debate concerning the credibility and trustworthiness of research outcomes. Other studies mention a large variety of ethical principles applying to scientific values, such as honesty, objectivity, integrity, carefulness, openness, trust, accountability, respect for colleagues and for intellectual property, confidentiality, fairness, efficiency, human subject protection, animal care, and so forth.

Scientific integrity has been described in terms of individual behaviour, covering scientific misconduct such as falsifying research data, ignoring or circumventing major aspects of human-subject requirements, not properly disclosing conflict of interest, changing the design, methodology or results of a study in response to pressure from a funding source, inappropriately assigning authorship credit, and so on (Martinson et al., 2005). The reputation of the research team, of the individual author and of the affiliated institutions, will contribute to the perceived quality of research, without being a guarantee. The application of open science principles like transparency and openness is designed to foster the "doing good science in a good manner."

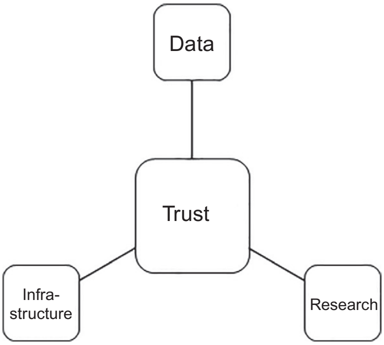


Fig. 18.2 Three variables of data trustworthiness.

With respect to data repositories, two types of use should be distinguished: the use of repositories to store, preserve, and share research data (deposit); and the use of repositories to verify published results, merge datasets, perform reanalyses, and so on (reuse). In the first case, the focus will be on the device, the reliability and security of the system, the promise of service (e.g., long-term preservation), the ease of deposit, the number of deposits by other researchers, and so on. In the second case, trustworthiness is also, and above all, conditioned by the content, the resources stored, and this will involve quality variables (the quality of the data and the quality and richness of the metadata), the right to reuse (licences), and interoperability with other operating systems.

This relationship between user confidence in preservation devices and confidence in the digital content of those devices has been modeled for digital archives in general (Donaldson, 2019) and formalized as an ISO standard (Open Archival Information System, ISO 14721:2012). Empirical studies, such as those by Yakel et al. (2013) or Yoon (2014), have led to a better understanding of some of the key factors of trust or mistrust in data repositories. Among these factors, three appear to be particularly important: the transparency of the system, the guarantee (promise) of long-term preservation (sustainability), and the reputation of the institution that manages and/or hosts the system. In addition to these factors, there are other criteria, such as the perception and experience of the functionalities and services and the quality of the data and the measures implemented to control, guarantee, and improve this quality, in terms of data sources and selection upstream and cleansing further downstream. Section 4 will deal with some of these issues.

The institution plays a separate role. Part of the trust placed in a platform of this type is linked to the characteristics of the institution responsible for and in charge of the platform. What is its reputation? Is it a reference in this field? What is its field of activity, does it have authority in this field? Who does it work with? What are its own references?

Recently, Science Europe presented a list of minimum criteria for a trustworthy repository, organized around four major themes: the assignment of unique and durable identifiers, the use of traceable and community-based metadata (standards), data access and licensing, and preservation, including the guarantee of data integrity and authenticity (Science Europe, 2019). Lin et al. (2020) summarized five principles under the acronym "TRUST"—that make data repositories trustworthy:

- **Transparency**: The repository must provide transparent, honest, and verifiable evidence of practices and procedures to convince users that it is able to guarantee the integrity, authenticity, accuracy, reliability, and accessibility of data over a long period of time. Transparency also means providing accurate information about the scope, target user community, mission and policy, and technical capabilities of the repository (including conditions of use).
- **Responsibility**: Repositories should take responsibility for managing their data holdings and serving the user community, through adherence to the designated community's metadata and preservation standards and the management of the deposited datasets, such as technical validation, documentation, quality control, protection of authenticity, long-term preservation, IP management, protection of sensitive information assets, and the security of the system and its content.

- **User focus**: Data repositories must provide services that correspond to the practices and needs of the target user community, which may vary from one community to another; they should be integrated into the data management practices of the target user communities and can therefore respond to the evolving needs of the community.
- **Sustainability**: Repositories must guarantee uninterrupted access to data, through risk management and a sustainable governance and business model.
- **Technology**: Repositories should have reliable, high-performance technology, i.e., appropriate software, hardware, and technical services that are up to the challenge, including compliance with standards and the implementation of measures to ensure the security of the system and data.

Regular audits and certification is one way to go there and to show compliance with standards, principles, and quality criteria. In the field of research data repositories, a couple of certificates have been developed, including the World Data System (WDS) certificate supported by the International Council for Science (ICSU) and the Data Seal of Approval (DSA) developed by the Dutch organization DANS from 2008. In 2018, the two procedures converged into the CoreTrustSeal (CTS), which today is the most recognized international certificate. The CTS certificate contains 16 themes with specific requirements for the organizational infrastructure, for the digital object management and for the repositories' technology. If the compliance with all requirements is needed for certification and to ensure the repositories' trustworthiness, four themes in particular address research data quality (CTS 2019):

(1) **Confidentiality/ethics (CTS requirement 4)**: *The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms*. The main concern is disclosure risk, the risk that an individual who participated in a survey can be identified or that the precise location of an endangered species can be pinpointed. Expected good practice includes special guidance and procedures and the request of confirmation that data collection or creation was carried out in accordance with legal and ethical criteria.

(2) **Data integrity and authenticity (CTS requirement 7)**: *The repository guarantees the integrity and authenticity of the data*. This requirement covers the whole data lifecycle within the repository; good practice includes checks to verify that a digital object has not been altered or corrupted, documentation of the completeness of the data and metadata, a version control strategy, a strategy for data changes, maintaining provenance data and related audit trails, and maintaining links to metadata and to other datasets.

(3) **Appraisal (CTS requirement 8)**: *The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users*. Which are the criteria of data acquisition and metadata? How do repositories control and validate the deposit of data? Good practice includes a collection development policy to guide the selection of data for archiving, procedures to determine that the metadata required to interpret and use the data are provided, automated assessment of metadata adherence to relevant schemas, and a list of preferred formats and checks to ensure that data producers adhere to the preferred formats.

(4) **Data quality (CTS requirement 11)**: *The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations*. Repositories must be able to evaluate completeness and quality of the data and metadata, and they must ensure there is sufficient information about the data for the designated community to assess the quality of the data.

> Good practice includes quality control checks to ensure the completeness and understandability of data deposited, the ability of the scientists to comment on, and/or rate data and metadata, and the provision of citations to related works or links to citation indices.

Together, these recommended good practices of data management contribute to the quality and trustworthiness of research data repositories, as much for the deposit as for the reuse of data. But certification requires a considerable effort of self-assessment and, subsequently, of a long-term action, monitoring, maintenance, and constant improvement. At the time of writing, about 100 data repositories have obtained the CoreTrustSeal certificate—less than 5% of all registered data repositories. With respect to COVID-19, 51 research data repositories are registered in the re3data directory, mainly from the United States and the European Union; only one, the WHO's International Clinical Trials Registry Platform (ICTRP), has been certified.

A last approach to make research data repositories trustworthy has been presented as FAIR guiding principles for scientific data management and stewardship (Wilkinson et al., 2016). Their objective is to improve the infrastructures that support the reuse of research data, by making data findable, accessible, interoperable, and reusable, with a focus on improving the ability of machines to automatically find and use data, as a complement to reuse by researchers. The "FAIRization" of data repositories relies to a great extent on the standardization of metadata and protocols. It is important to keep in mind the vision of this approach, which aims to foster the coherent development of the global Internet of FAIR data and services, particularly within the framework of the European Research Area and the European Open Science Cloud (EOSC) infrastructures. While inclusion in a FAIR and/or certified repository can be thought of as an indicator of quality, on its own, the FAIR principles do not aim to this day to ensure the quality of research data itself.

4 How to improve data quality

One major problem with research data is the ever-increasing volume; another is the large variety of format, typology, structuration, size, description, context, origin, and more. The standards, recommendations, and certifications require action to assure data quality, such as quality checks and data cleansing procedures. We mentioned above the need for appraisal, for some basic validation and for selection based on "good research practice." What can and should data repositories do to ensure data quality of accepted datasets, beyond fast checking, quick reviewing, and request of confirmation that data collection or creation was carried out in accordance with legal and ethical criteria?

Authorities, business and research communities expect scientific institutions and infrastructures to have the quality of their research data under control. A high data quality is therefore essential for an organization to become or remain trustworthy with its target groups of users (Budianto, 2019). But quality needs to be measured and improved continuously, because of constant changes, quality deterioration, obsolescence, and so on (Mahanti, 2019). According to recent studies, many institutions are dissatisfied with the quality of the research data collected and processed, and rate it as low or rather low (Cai and Zhu, 2015; Logan et al., 2020). Often, they only improve their data quality occasionally, or do not take any action at all.

If institutions want to ensure consistently high quality, they should set up an ongoing data quality initiative. This does not have to be an all-or-nothing project, but can initially be carried out on a departmental basis and gradually expanded. Libraries, which of course rely on high-quality data, are particularly useful in this regard. The overriding goal of such an initiative should be that the research data in the respective area are uniform, complete, and up-to-date. Inconsistent and incorrect research data lead to wrong decisions, loss of trust, failure of projects, employee dissatisfaction, and more. For this reason, the following questions always arise, especially with exponentially increasing amounts of data: How can insufficient research data in institutions be identified? How can the quality of the data be improved?

In Computer Science, many papers have discussed methods for the data quality management in information systems from different domains, such as the first-time-right principle, the closed-loop principle, data catalogue, data profiling (Azeroual et al., 2018b), data cleansing (Azeroual et al., 2018a), data wrangling (Azeroual, 2020), data monitoring, data lakes (Mathis, 2017), data text mining (Azeroual, 2019), and machine learning (Duka and Hribar, 2010; Maali et al., 2010); these papers have also shown how the methods can be used in practice to ensure data quality. The methods of data cleaning and monitoring range from fully automated to mostly manual operations, which is closely related to the amount of knowledge required for each operation. However, before these methods can be used, a number of steps must be performed.

First of all, the importance of high data quality must be anchored in the awareness of employees. Only when management is aware that data repositories are not functioning properly without clean data can the strategic goal of clean data storage be successfully transferred to libraries. In order to make employees aware of the potential of a clean data repository, it helps to formulate specific goals. These can then be tracked with data quality management, for example, by improving transparency and decision-making in facilities or stabilizing customer relationships. The starting point for data cleansing is knowing the actual quality of the research data. Today, effective analysis tools and methods are available to record and map the current situation. Quality problems with manageable effort can be identified, and the error frequency can be given in an order of magnitude. Usually redundant author data, incomplete datasets, and incorrectly recorded data, as well as contradictions between different databases, occur. Before the actual data cleansing can begin, rules should be worked out that define clear standards for which data are relevant and what a clean dataset should look like. For example, is the author's metadata required or optional? Or is a dataset with affiliations of the author already considered complete? Based on these characteristics, it is possible to evaluate the data repository and determine what to do with incorrect data.

The data records can be cleaned up as soon as they are recorded in institutions or transferred to a data repository. With the systematic initial cleaning of incorrect data, a solid basis for a permanent data maintenance strategy can be created. It is important to determine which level of quality is to be achieved over what time period by a first cleanup run. However, data quality assurance must not remain a one-off problem. Keeping it clean is the key to long-term success. In order to guarantee perfect data quality in the long term, processes for regular quality control and data cleansing should be set up. The following steps of data cleansing of a data repository can be used here and can also be reused as continuous mechanisms for monitoring data integrity (Azeroual et al., 2018a):

(1) **Parsing** is the first critical component of data cleansing and helps the user understand and transform the attributes. Individual data elements are referenced according to the metadata. This process locates, identifies, and isolates individual data elements, as, for example, for names, addresses, zip code, and city. The parser profiler analyzes the attributes and generates a list of tokens from them, and with these tokens, the input data can be analyzed to create application-specific rules. The biggest problem here is different field formats, which must be recognized.

(2) **Correction and standardization** is necessary to check the parsed data for correctness, then to subsequently standardize it. Standardization is the prerequisite for successful matching, and there is no way around using a second reliable data source. For address data, a postal validation is recommended.

(3) **Enhancement** is the process that expands existing data with data from other sources. Here, additional data are added to close existing information gaps. Typical enrichment values are demographic, geographic, or address information.

(4) **Matching** There are different types of matching: for reduplicating, matching to different datasets, consolidating, or grouping. The adaptation enables the recognition of the same data. For example, redundancies can be detected and condensed for further information.

(5) **Consolidation (Merging)** Matching data items with contexts are recognized by bringing them together.

Many errors can be avoided by doing these data cleansing steps directly when entering data. According to Azeroual et al. (2018a), all of these steps are important in order to achieve and maintain maximum data quality in a database or information system. Quality errors in the acquisition and integration of several data sources in one system are eliminated by data cleansing. The manual effort for data cleansing can be minimized through a high degree of automation. During data cleansing, duplicates are eliminated, data types corrected, or incomplete data records completed. Finally, data monitoring checks the quality of the available data at regular intervals. If the data quality changes, the monitoring system provides information and enables new analysis or corrective measures to be initiated. Only if the data quality is continuously monitored and the results are communicated, can the quality of the data stocks be maintained over longer periods of time. If there is no monitoring, the quality level achieved will continuously decrease over time. The change in quality levels can be tracked via trend reports and alarms from data monitoring.

For a consistent and redundancy-free data repository, not only those who are responsible for the data quality in the specialist departments, but also for the entire facility should be named. Clear responsibilities—for data entry and error correction—avoid distributed responsibilities and specific data storage in parallel data repositories, which would have to be continuously synchronized and checked for redundancies. At the same time, those responsible for data maintenance need freedom of action and access to management in order to initiate and implement improvements. Even if employees are actively involved in data quality management, they should be trained regularly. Data quality remains present in day-to-day business; factors that cannot be regulated via automatic mechanisms should be consistently observed. There are numerous applications on the market to minimize the manual effort in facilities to ensure high data quality. They automate data analysis, data cleansing, and data monitoring processes using intelligent algorithms. The huge amounts of data in the big data environment can only be kept at a high-quality level with such supporting software.

Achieving and maintaining high data quality are in the best interests of any institution that wants to derive added value from its research data, but this costs money. However, poor data quality can cost even more in the medium term. For example, in addition to a multitude of challenges (such as many data sources, different data formats, and different updates) which make efforts to improve data quality more difficult, the research on COVID-19 is changing in front of our eyes and is more open, transparent, and collaborative within a few months or weeks become. Many relevant research data and publications on COVID-19 are currently freely accessible; many of the publications enable direct access to the original data on which they are based. Many scientific papers also appear as a so-called preprint at the moment of completion, without delay due to a peer review process and other adversities of the commercial publication process. This transparency is by no means a guarantee that the majority of COVID-19 studies are of high quality and of a high ethical standard. It is therefore necessary to identify and correct these problematic studies early on before storing this publication data in data repositories. If quality is taken into account during data acquisition, many metadata errors can be eliminated from the outset. This is more effective than laboriously looking for mistakes afterward. Problems with poor data quality should be addressed before the data is used—right at its place of origin. Since the real costs of incorrect, incomplete, and redundant data are difficult to quantify, many scientific institutions delay investments in targeted data quality management. Investing in data quality cannot be combined with a return on investment. However, this can be a prerequisite for the successful further development of the business model. Institutions should therefore make data quality a strategic goal. Looking at future developments, with the increasing networking inside and outside of institutions, it becomes clear that topics such as data repositories can only be successfully implemented if the challenge of managing the flood of data professionally is accepted at the same time. It is to be expected that the use of data repositories will increase significantly in the coming years and their application can only succeed if a clean amount of data is available.

5 Conclusion

On November 13, 2020, the WHO dashboard gave 51,848,261 confirmed cases and 1,180,868 deaths due to COVID-19. The pandemic is far from over. In the meanwhile, about 10,000 new papers have been published, and Mendeley Data indexes 7121 more datasets, an increase of 63%. More and more information of all kinds, volume, velocity, and variety: This is part of what can be called the academic big data. So, what about the fourth V: veracity?

Given the fact that not all published results are reliable and trustworthy, this chapter has provided insights into the control of data deposits, the factors that make data trustworthy, and the methods and procedures to improve quality of deposited data. Data repositories and their hosting institutions play a key role; essential requirements to ensure a minimal level of data quality and reliability have been listed and described.

When the European Union member states prepared their action plan for open science (Amsterdam Call for Action, 2016), the political concern focused on Ebola and Zika. The 2020 COVID-19 pandemic for the first time demonstrated the potential of open science, but also the issues, especially with early and nonvalidated results. The pandemic has triggered many open science initiatives (see Uribe-Tirado et al., 2020) and many calls for opening research to make research results available for everyone, as soon and as open as possible. The severity of the pandemic has increased academic scientists' willingness to share data and results, even if other factors like corporate ownership of data and increased competition can also generate obstacles "since disclosing crucial data and information can improve competitors' positions and reduce one's chances to succeed" (Younes et al., 2020). It is too early to assess if and how open science contributes not only to acceleration of research but also to innovation; if and how, for instance, the very short delay for the development of new vaccines (Pfizer, BioNTech) was at least partly conditioned by open and seamless access to research results or not. Data sharing has advantages and disadvantages, as well for the individual scientist as for the community and the scientific development as a whole (Rios et al., 2020). Other issues have been raised, in particular potential data misuse and abuse of published virus genome sequences; "traceability should be a key function for guaranteeing socially responsible and robust policies. Full access to the available data and the ability to trace it back to its origins assure data quality and processing legitimacy" (Minari et al., 2020).

Yet, if open science is partly the cause of some of the issues associated with data quality, it also provides the solution, through increased accessibility, transparency, and integrity of research data and the whole process. Open science is not the problem; the problem is inappropriate adoption of open science principles (Besançon et al., 2020). Not less but more open science is the way forward, with preregistered analyses, registered reports, open reviews of methodologies and data, data sharing by default via dedicated platforms, and so on. If data are open, it can be assessed, by usual peer-reviewing but also by "crowd-reviewing." Another perspective is artificial intelligence. Elsevier is looking into automated machine learning detection of scientific fraud; together with Humboldt University in Germany, they are working on mechanisms to check the integrity of the articles, the images, and the data (HEADT Centre). The next years will show if and how this AI approach will contribute to the quality, integrity, and trustworthiness of research data. Probably (hopefully), this will be after the COVID-19 pandemic.

References

Amsterdam Call for Action, 2016. Amsterdam Call for Action on Open Science, Amsterdam, April 4–5. https://www.government.nl/topics/science/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science.

Aristovnik, A., Ravšelj, D., Umek, L., 2020. A bibliometric analysis of COVID-19 across science and social science research landscape. Sustainability 12 (21), 9132. https://doi.org/10.3390/su12219132.

Arrizabalaga, O., et al., 2020. Open access of COVID-19-related publications in the first quarter of 2020: a preliminary study based in PubMed. F1000Research 9, 649. https://doi.org/10.12688/f1000research.24136.2.

Azeroual, O., 2019. Text and data quality mining in CRIS. Information 10, 374. https://doi.org/10.3390/info10120374.

Azeroual, O., 2020. Data wrangling in database systems: purging of dirty data. Data 5, 50. https://doi.org/10.3390/data5020050.

Azeroual, O., Saake, G., Abuosba, M., 2018a. Data quality measures and data cleansing for research information systems. J. Digit. Inf. Manag. 16 (1), 12–21. https://arxiv.org/abs/1901.06208.

Azeroual, O., Saake, G., Schallehn, E., 2018b. Analyzing data quality issues in research information systems via data profiling. Int. J. Inf. Manag. 41, 50–56. https://doi.org/10.1016/j.ijinfomgt.2018.02.007.

Batini, C., et al., 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41 (3), 1–52. https://doi.org/10.1145/1541880.1541883.

Besançon, L., et al., 2020. Open science saves lives: lessons from the COVID-19 pandemic. BioRxiv. https://doi.org/10.1101/2020.08.13.249847. 2020.08.13.249847.

BioRxiv, 2020. COVID-19 SARS-CoV-2 Preprints from medRxiv and bioRxiv. https://connect.biorxiv.org/relate/content/181.

BOAI, 2001. Budapest Open Access Initiative. https://www.budapestopenaccessinitiative.org/read.

Budianto, A., 2019. Customer loyalty: quality of service. J. Manag. Rev. 3 (1), 299–305. https://doi.org/10.25157/jmr.v3i1.1808.

Cai, L., Zhu, Y., 2015. The challenges of data quality and data quality assessment in the big data era. Data Sci. J. 14, 2. https://doi.org/10.5334/dsj-2015-002.

CoreTrustSeal Standards and Certification Board, 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022., https://doi.org/10.5281/ZENODO.3638211.

Donaldson, D.R., 2019. Trust in archives—trust in digital archival content framework. Archivaria 88 (Fall), 50–83. https://muse.jhu.edu/article/740193/summary.

DuBois, J.M., Antes, A.L., 2018. Five dimensions of research ethics: a stakeholder framework for creating a climate of research integrity. Acad. Med. 93 (4), 550–555. https://doi.org/10.1097/ACM.0000000000001966.

Duka, D., Hribar, L., 2010. Implementation of first time right practice in software development process. In: *The 33rd International Convention MIPRO*, Opatija, pp. 382–387.

Haak W. et al., Mendeley data. In Schöpfel J. and Rebouillat V. (Eds.), *Les entrepôts de données de recherche*, Forthcoming, ISTE Editions, London.

Haider, J., 2018. Openness as tool for acceleration and measurement: reflections on problem representations underpinning open access and open science. In: Herb, U., Schöpfel, J. (Eds.), Open Divide? Critical Studies on Open Access. 17–28, Litwin, Sacramento CA https://lup.lub.lu.se/search/publication/070c067e-5675-455e-a4b2-81f82b6c75a7.

Janiaud, P., et al., 2020. The worldwide clinical trial research response to the COVID-19 pandemic—the first 100 days. F1000Research 9, 1193. https://doi.org/10.12688/f1000research.26707.2.

Konkiel, S., 2020. Assessing the impact and quality of research data using Altmetrics and other indicators. Scholarly Assess. Rep. 2 (1). https://doi.org/10.29024/sar.13.

Lin, D., Crabtree, J., Dillo, I., et al., 2020. The TRUST Principles for digital repositories. Scientific Data 7 (1), 144. https://doi.org/10.1038/s41597-020-0486-7.

Logan, C., et al., 2020. Improving data quality in face-to-face survey research. PS: Polit. Sci. Polit. 53 (1), 46–50. https://doi.org/10.1017/S1049096519001161.

Maali, F., Cyganiak, R., Peristeras, V., 2010. Enabling interoperability of government data catalogues. In: Wimmer, M.A., Chappelet, J.L., Janssen, M., Scholl, H.J. (Eds.), Electronic Government. EGOV 2010. Lecture Notes in Computer Science, vol. 6228. Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-14799-9_29.

Mahanti, R., 2019. Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. ASQ Quality Press, Milwaukee, WI.

Martinson, B.C., Anderson, M.S., De Vries, R., 2005. Scientists behaving badly. Nature 435 (7043), 737–738. https://doi.org/10.1038/435737a.

Mathis, C., 2017. Data lakes. Datenbank Spektrum 17, 289–293. https://doi.org/10.1007/s13222-017-0272-7.

Mesquita, C.T., 2020. Open science and the role of cardiology journals in the COVID-19 pandemic. Int. J. Cardiovasc. Sci. 33 (4), 305–306. https://doi.org/10.36660/ijcs.20200191.

Minari, J., Yoshizawa, G., Shinomiya, N., 2020. COVID-19 and the boundaries of open science and innovation. EMBO Rep. 21 (11). https://doi.org/10.15252/embr.202051773.

OECD, 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding, Report. Paris https://www.oecd.org/sti/inno/38500813.pdf.

Piller, C., 2020. Who's to blame? These three scientists are at the heart of the Surgisphere COVID-19 scandal. Science, 20. https://doi.org/10.1126/science.abd2252. June 8.

Rios, R.S., Zheng, K.I., Zheng, M.H., 2020. Data sharing during COVID-19 pandemic: what to take away. Expert Rev. Gastroenterol. Hepatol. 14 (12), 1125–1130. https://doi.org/10.1080/17474124.2020.1815533.

Roozenbeek, J., et al., 2020. Susceptibility to misinformation about COVID-19 around the world. R. Soc. Open Sci. 7 (10), 201199. https://doi.org/10.1098/rsos.201199.

Schöpfel, J., et al., 2019. Data papers as a new form of knowledge organization in the field of research data. Knowl. Organ. 46 (8), 622–638. https://doi.org/10.5771/0943-7444-2019-8-622.

Schulte-Schrepping, J., et al., 2020. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. Cell 182 (6), 1419–1440.e23. https://doi.org/10.1016/j.cell.2020.08.001.

Science Europe, 2019. Practical Guide to the International Alignment of Research Data Management. Science Europe Working Group on Research Data, Brussels. https://www.scienceeurope.org/media/jezkhnoo/se_rdm_practical_guide_final.pdf.

Stausberg, J., et al., 2019. Indicators of data quality: review and requirements from the perspective of networked medical research. GMS Medizinische Informatik, Biometrie und Epidemiologie 15 (1). https://doi.org/10.3205/mibe000199. Doc05.

Uribe-Tirado, A., et al., 2020. Open science since COVID-19: open access + open data. SSRN Electron. J. https://doi.org/10.2139/ssrn.3621047. June 3.

Wang, R., Strong, D., 1996. Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. 12 (4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.

Wilkinson, M.D., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3 (1), 160018. https://doi.org/10.1038/sdata.2016.18.

Yakel, E., et al., 2013. Trust in digital repositories. Int. J. Digit. Curation 8 (1), 143–156. https://doi.org/10.2218/ijdc.v8i1.251.

Yoon, A., 2014. End users' trust in data repositories: definition and influences on trust development. Arch. Sci. 14 (1), 17–34. https://doi.org/10.1007/s10502-013-9207-8.

Younes, G.A., et al., 2020. COVID-19: Insights from innovation economists. Sci. Public Policy. https://doi.org/10.1093/scipol/scaa028. paaa036.