information

MDPI

Article

Quality of Open Research Data: Values, Convergences and Governance

Tibor Koltay

Institute of Learning Technologies, Eszterházy Károly University, HU-3300 Eger, Hungary; tkoltay@freemail.hu

Received: 30 January 2020; Accepted: 23 March 2020; Published: 25 March 2020

Abstract: This paper focuses on the characteristics of research data quality, and aims to cover the most important issues related to it, giving particular attention to its attributes and to data governance. The corporate word's considerable interest in the quality of data is obvious in several thoughts and issues reported in business-related publications, even if there are apparent differences between values and approaches to data in corporate and in academic (research) environments. The paper also takes into consideration that addressing data quality would be unimaginable without considering big data.

Keywords: data quality; research data; data governance

1. Introduction

Open science encourages cooperative work in research and making knowledge publicly accessible in digital formats with minimal or no restrictions [1]. Research data is no exception. As, for example, the Royal Society [2] stated, effective data ecology requires all stakeholders of scientific work to elaborate and implement transparent policies, paying attention to custodianship and data quality (DQ).

Indeed, a growing number of research funding organizations all over the world have begun to promote the sharing of scientific data. However, merely releasing or posting data is not enough. Achieving real benefits of sharing research data is possible only if data are reused by others [3]. This requirement involves being concerned with its quality.

Based on a non-exhaustive review of the literature, this paper aims to identify the most important (overwhelmingly theoretical) issues of research data quality and data governance. Although we will focus on the characteristics of research data quality, being aware of the corporate word's considerable interest in the quality of data, some business-related publications on these issues could not be neglected, even if there are apparent differences between values and approaches to data in corporate and in academic (research) environments.

The distinction between these two types of data is clear. Research data are the outputs of a systematic investigation that involves observation, experiments or the testing of a hypothesis [4], and consists of "heterogeneous objects and items used and contextualized, depending on the academic discipline of origin" [5] (p. 3). In contrast to this, data used in the corporate world are usually seen to be company assets, the latter being defined as a resource controlled by a business entity as a result of past events or a transaction, and from which "future economic benefits are expected to flow to the entity." [6].

Even if it is difficult to assess data, because its appraisal requires deep disciplinary knowledge and manually evaluating datasets is both time consuming and expensive while automated approaches are in their infancy [7], we can say that DQ is one of the cornerstones of the fourth, data-intensive paradigm of scientific research, also called Research 2.0 or eScience.

Let us add that researchers do not need more data, but require the right data. To serve as research data, small amounts of data can be just as valuable as big amounts, and—in many cases—there are no data, because relevant data cannot be found, are not available, or do not exist at all [3].

Negotiating a trade-off between the value and risks of data processing happens on different levels. In the case of research data, the upper tier is the macro (international, national, sectoral) level of the funders. The level of organizations, or institutions, such as repositories and data centers, is an intermediate level (often called the meso-level), which requires an important group at the micro (individual) level—support staff members [8,9]. By describing data quality concepts and their relationships with each other in general, this paper aims to inform first of all, stakeholders on these latter two levels.

Nevertheless, before examining data quality issues themselves, we need to consider the changes which the views on the nature of data have undergone.

2. Changing Views on the Nature of Data

Current views on DQ are profoundly influenced by the changing perspectives on data. Traditionally, data have been perceived as somewhat secondary to information, because they were regarded to occupy the bottom of the data–information–knowledge–wisdom pyramid [10]. The relationship between these fundamental concepts is no longer seen as simply any as it has been portrayed earlier [11]. Today we accept that data are anything recordable in a database in a semantically and pragmatically sound way. From this point of view, recordings should be understood as true or false statements. The pragmatics require that we favor recordings that seem to be concrete facts [12]. It is acknowledged that—from the ontological point of view—both data and information are signs, and thus are closer to each other than was perceived earlier [13]. Nonetheless, there are criticisms against data that they puts on "the garments information used to wear" [14]. As for research data in particular, they can be seen not just as the result of empirical work or the raw material for statistical analysis, but also as research objects in their own right [15].

None of the above views excludes perceiving data as primary intellectual assets that can be subjected to quality assessment [16]. Notwithstanding, we should understand that research data consist of "information artefacts that scholars draw upon for evidence in supporting research claims, and producing new knowledge" [17]. However, raw data are not the only form of evidence; thus, we need to build on reusing research data [3].

As we pointed out, there is a close relationship between research data and business data; it is worth noting that until the early 1990s, data were not treated as valuable business assets, but regarded to be by-products that lose most of their value beyond a given transaction [18].

Baskarada and Koronios [19] offer a somewhat independent perspective that allows looking at the differences between academic and corporate views on quality. As a part of dissolving the dichotomy between them, they direct attention to the semiotic point of view. This view, including syntactics and semantics and scan be explained by the following example. At the level of syntactics we find accuracy, concise representation, consistency and ease of manipulation, whereas semantics involves believability, understandability and interpretability. Pragmatics contains reputation, added value, completeness and relevance. If we take the example of sentences, we can say that their quality is high because they are characterized by correct letters, spelling and grammar. After all, we can perceive poor quality when a non-native speaker does not understand the given sentence, either partially or in its entirety. This implies that data quality problems are caused not only by engineering flaws (e.g., because of using improper software and/or hardware), but by human factors (brought about, e.g., by systematic spelling mistakes). This dictates that DQ should be assessed against predefined specifications for encoding, storage and communication of data. To do this, we can chose the level of empirics, which displays accessibility, timeliness, security and traceability.

3. The Stakeholders of Research Data Quality

There is a diverse range of relevant stakeholders who can play a number of roles and represent much value, including through their supporting of the evaluation and maintenance of quality.

The leading stakeholders of research data quality are the researchers themselves. However, we will not address their knowledge and skills, among others reasons, because "The increasing ease and speed with which researchers can collect large, complex datasets is outpacing their development of the knowledge and skills that are necessary to properly manage them" [20] (p. 382).

Varied research institutions can have an active role in maintaining DQ, among other things, giving attention to the technical and scientific quality of research data, to be described later.

Several repositories and data centers have developed quality assurance measures and offer a range of services with which to evaluate the technical quality of datasets [21]. Data centers with expertise in data curation can play an important role not only in enhancing data management skills of research library staff, but DQ. Data curation, exercised mainly in data centers is increasingly aimed at improving data quality, in order to guarantee authentication of digital resources throughout use and reuse [17].

To secure the reusability and long-term preservation of these data, funding organizations and publishers are urging researchers to deposit their data in certified and accredited repositories. In this context, accreditation or certification to appropriate standards of repositories is a key to ensuring the quality of data.

Scholarly publishers want to exploit the possibilities of research data by providing new products, such as publishing data journals, and offering services to enhance DQ. Data journals are community peer-reviewed open access platforms for publishing, sharing, and disseminating data that cover a wide range of disciplines. The papers published in data journals contain information on the acquisition, methods, and processing of specific datasets. The published papers are cross-linked with approved repositories, and they cite datasets that have been deposited in repositories or data centers. The publication of data papers maximizes the opportunities for data re-use. Publishers can also help to increase DQ by promoting transparency and openness in several key areas:

- Understanding researchers' problems and motivations;
- Raising awareness of issues and encouraging behavioral or cultural change;
- Improving the quality and objectivity of the peer-review process;
- Improving scholarly communication infrastructure with journals that publish scientifically sound research.

They also promote partnering with data repositories and are instrumental in making research communication more accessible with open-access publishing options [22].

Quality assurance is one of the main components of scholarly research data management (RDM), offered mainly by academic libraries [23]. By being a broad concept, RDM includes processes undertaken to create organized, documented, accessible, and reusable quality research data [24]. DQ and RDM are also connected to data literacy, which is a "human competence to locate, analyze, organize, present, and evaluate" research data [25] (p. 136). Evaluation takes place already when we collect data, and the importance of critically analyzing others' interpretations of data as accurate, and representing conclusions correctly is underlined. Data literacy is closely connected to information literacy that emphasizes critical thinking and the necessity to recognize message quality [26]. By being an implicit measurement of DQ, data citation is also an issue, promoted not only by libraries, but practiced to a substantial extent by them [27].

The fate of DQ depends considerably on the successful interventions of data professionals. Although it is extremely difficult to distinguish between their different groups because of their overlapping functions, we will concentrate on three of them: data managers, data librarians, and data scientists. Data managers come largely from the ranks of computer scientists and information technologists, and take responsibility for computing facilities, storage, continuing access, and preservation of data. To ensure DQ for reuse and long-term preservation, data managers must be equipped with a range of quality assurance and control strategies to enable them to establish quality assurance mechanisms. Their collaboration with funders, publishers, and journal editors is central to ensuring the enforcement of relevant policies [20].

Data librarians are professionals, trained and specialized in the curation, preservation, and archiving of data. More recently they have been involved in advocacy, research data management, and delivering training at universities, research institutes, and their libraries [28].

Data scientists undoubtedly represent a new breed of professionals [29]. In their work, the design for quality assessment takes place in the first phase of the data lifecycle, i.e., the theoretical and systematic construction of a problem to be solved [30]. This is in accordance with the opinion that it is fundamental to understand the principles of design and information architecture [31]. All in all, data scientists can play an important role in quality control [32].

The opinions about data science and data scientists are sometimes contradictory. According to Wang [33], data science often ignores DQ. He adds that—if this situation continues, the old saying of the 1980s "garbage in, garbage out" will remain true in the big data age. On the other hand, Cao [34] stated that data science—being creative, intelligent, exploratory, nonstandard, and personalized—inevitably involves various quality issues. Beyond these three groups of professionals, data stewards (already mentioned above and called data curators as well), are professionals involved in curation, cleansing, annotation, selection, and appraisal of data. They are required to be familiar with quality assurance standards [35]. Data stewards are not only instrumental in supporting data management but also actively promoting data governance among researchers

4. The Nature of Data Quality

As underlined by Foster et al. [8], there are several general uncertainties, related to data, but the quality of the data remains a key risk. Data quality is a multidimensional concept with context-independent properties of individual data values and others that are dependent on the context of use [36]. Besides this, legal and regulatory requirements [37], the differing nature of big data [38], and cloud computing [39] have an impact on DQ, as well.

4.1. Data Quality Attributes

In the literature on DQ we can find a wide variety of attributes, often grouped differently.

On the premise that data consumers have a broad data quality conceptualization, in the initial phase of their investigation, Wang and Strong [40] found 179 attributes.

Distinguishing between the intrinsic and extrinsic nature of data quality seems to be worth consideration. The reason for this is that data quality is defined to a certain extent by the fact that the value of data is based on its fitness for use, which is an extrinsic property [41], while digitized objects hold intrinsic value, regardless of whether they are ever utilized for any purpose [42]. This is the reason why the emphasis on intrinsic DQ appears in the categorization of Strong, Lee, and Wang [43], who group DQ dimensions into four categories. Intrinsic DQ includes accuracy, objectivity, believability, and reputation. Accessibility and access security pertain to accessibility DQ. Contextual DQ is defined by relevancy, value-addedness (value added nature), timeliness, completeness, and the amount of data. They also explain that difficulties in utilizing data can be led back to poor relevancy, which means that data is not relevant to the data consumers' tasks due to being incomplete for analysis and aggregation, either because there are changes in the data consumers' needs, or problems in operational data production occur.

Representational DQ falls into the dimensions of interpretability, ease of understanding, concise representation, and consistent representation.

Table 1 shows data quality attributes as categorized by Strong, Lee, and Wang [43].

Table 1. Data quality attributes.

Intrinsic DQ	Accessibility DQ	Contextual DQ	Representational DQ
accuracy	accessibility	the amount of data	concise representation
objectivity,	access security	completeness	consistent representation
believability		relevancy	ease of understanding
reputation		timelines	interpretability
		value-addedness	

Some of these attributes can be described in more detail. Objectivity, means the degree of being impartial. Reputation indicates that data and its source are kept in high consideration. The amount of data must be appropriate in the sense of being available in the suitable quantity. Completeness refers to the appropriate scope of the information in the data. Relevancy involves the degree of being applicable or interesting. Timeliness denotes the age of the data. The value added nature provides a competitive advantage to data. Concise representation is understood as compactness. Consistency means that data is continuously presented in the same format. Ease of understanding signifies being clear, readable, or understandable. Interpretability, stands for the extent to which the data meaning is explained [40,44].

Trust (trustworthiness) is one of the most notable DQ attributes. Similarly to many other attributes, it is complex in itself, being influenced by the given subject discipline, the data creators' reputation, and the biases of those who evaluate the data. On the other end of the data lifecycle, we see that one basic component of trust can be approached by answering the question of whether the given dataset has been reviewed by anyone other than its creator or not [45]. As pointed out by Wolski, Howard, and Richardson [46], trust related either to research data in general or data in research data repositories is of extraordinary importance. Their Trust Framework for Online Research Data Services contains six determinants. Competence serves for providing evidence to give the end-user confidence that the service provider can do what they promise by providing information about governance; evidence regarding sources of methods, data, and tools; and security, privacy, and other legal compliance, just to name a few. In this framework, culture characterizes the organization, which provides the given service. Let us add that research institutions and universities have a benevolent level of trust inbuilt because of their inherent mission, especially if they demonstrate openness and transparency. Commitment is also a key characteristic for building trusting relationships, which can be shown through the website of the service provider or can have its origin in the lived experience of the user. Commitment can be documented in terms of service and agreements between partners who provide the service. Conflicts may arise from different perceptions and expectations about goals or levels of services and should be resolved properly. Efficient communication between the service provider and the end-user is critical at all levels. Risks should be reduced and satisfaction building should be considered in developing a response to the trust framework, remembering that different perceptions of value propositions have to be managed.

From among the components of trust, understandability of data depends on providing sufficient context in the form of documentation, metadata, or information on provenance. Usability means that data are discoverable, accessible, and in a usable format. Proving data to be identical, at the bit level, to some prior accepted or verified state enables integrity, which is required for usability, understandability, and authenticity—in other words, overall quality.

Trust depends to a substantial extent on subjective judgements on authenticity, acceptability, or applicability of the data; and is also influenced by the given discipline (subject), the reputation of those responsible for the creation of the data, and the biases of the persons who are evaluating the data. It is also related to cognitive authority, which has two levels. At an operational level, it is the extent to which users think that they can trust the information [47]. On a more general level, cognitive authority refers to influences that a user would recognize as proper because the information therein is thought to be credible and worthy of belief [48]. The elements of trust also include the lineage, version, and error

rate of data, and the fact that they are understood and acceptable [49]. In the case of data repositories, the perception of trust is based largely on a "lack of deception" [50].

Figure 1 shows the relationships between elements of trust, where the numbed elements of the figure are the following ones:

1. Origin (originality);
2. Methods of collecting and processing;
3. Authenticity;
4. Acceptability;
5. Applicability;
6. Understandability;
7. Subject discipline;
8. Reputation of the creator(s);
9. Biases of the evaluators;
10. User's confidence;
11. "Lack of deception";
12. Independent review;
13. Reuse.

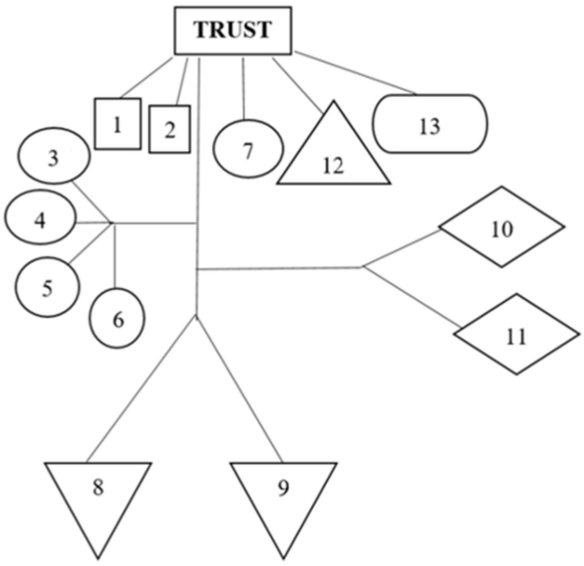

Figure 1. The relationships between elements of trust.

Authenticity is another prominent DQ attribute; it measures the extent to which the data are judged to represent the proper ways of conducting scientific research, including the reliability of the instruments used to gather the data, the soundness of underlying theoretical frameworks, the completeness, accuracy, and validity related to other dimensions, such as the accuracy, timeliness, completeness, and security of the data [51]. In order to evaluate them, the data must be understandable. The correct evaluation of understandability is allowed by context, which is in the form of documentation and metadata. To achieve this, data have to be usable. To make data usable, they have to be discoverable and accessible, and be in a usable file format. The individuals judging DQ need to have at their disposal an appropriate tool to access the data, which has to show sufficient integrity to be rendered. Integrity of data assumes that the data can be proven to be identical, at the bit level, to some previously accepted or verified state. Data integrity is required for usability, understandability, authenticity, and thus overall quality [47].

4.2. Technical and Scientific Quality

In a somewhat different dimension, we can speak about technical quality, which includes aspects such as the completeness and consistency of a dataset, and the use of appropriate standards and software [21].

As described by the World Data Center for Climate, the minimal set of technical quality criteria is the following:

- The number of datasets is correct and bigger than 0;
- The size of every dataset is also bigger than 0;
- The datasets and corresponding metadata are accessible;
- The data sizes are controlled and correct;
- The metadata is consistent to the data;
- The format is correct;
- Data and its descriptions are consistent [52].

Technical quality is complemented by scientific quality, which is assessed by the research community through pre- and post-publication peer-review. Its assessment is based on the assumptions that the peer reviewers do not normally work for economic benefit, but in order to exercise academic authority, and that they are driven by allegiance to the given scientific community [53].

4.3. Data Quality and Data Reuse

Reuse of data can be defined as the use of data by someone who did not collect them. It is a secondary use of data, which aims at addressing new problems [54], including the reproduction or replication of prior study results, which requires considering ethical questions of collecting and managing data, because—unlike scholarly outputs that have established peer review systems—there is no similar process for validating data [55].

Reuse must be enabled by the provision of high quality curated data resources [9], and data curation provides the means for assuring data integrity and authenticity; and enables reusers to access high-quality data they can trust [55]. The perceptions, personal knowledge, skills, and understanding of errors influence how reusers judge DQ [56]. Data quality attributes, such as accessibility, completeness, ease of operation, data credibility, and the quality of documentation also contribute to decisions about what data to reuse [57]. In this context, answering questions about the creator and the financial supporters of the dataset, as well as about bias, currency of maintenance, references about its past use, and updates, has to be emphasized. A similar set of useful questions can also be added, as follows:

- Who is in charge of checking for quality?
- What process do they use?
- How is missing data handled? [58]

Naming conventions, units used, and dates of creation and update also play roles in reuse; and data citation is an effective tool for identifying data reusers, as well [57]. Information about the producer is somewhat helpful in making decisions for reusing data [59].

The same views and quality attributes are advantageous for those who repurpose data. Somewhat differently from data reuse, that means using the data more than once for the same purpose; repurposing tends to mean the use of data for a completely different purpose than the original ones [55].

4.4. Other Quality Factors

Depending on the type of the given data standard and the aspects of data quality considered, data standards may or may not impact it. There are at least two world-views on approaching quality. One is conformity to specifications; the other is fitness for use. While they are not necessarily different, there may be some divergence between them [36].

Notwithstanding, the data quality model based on the ISO/IEC 25012 standard deserves attention. This model classifies data quality characteristics into three categories. Most characteristics are seen in specific contexts of use. The quality characteristics, described in this standard, are classified in three main categories:

- Inherent data quality;
- System dependent data quality;
- Inherent and system dependent data quality.

The first one refers to data itself in terms of data domain values and possible restrictions, relationships of data values, and metadata. The second category depends on the technological domain in which data are used and is the degree to which data quality is reached and preserved within a computer system. The third one is related to both categories.

Table 2 shows the distribution of the 15 data quality categories.

Table 2. ISO data literacy categories based on [60].

Inherent Data Quality	System-Dependent Data Quality	Inherent and System-Dependent Data Quality
Accuracy	Availability	Accessibility
Completeness	Portability	Compliance
Consistency	Recoverability	Confidentiality
Credibility		Efficiency
Currentness		Precision
		Traceability
		Understandability

From among the group of inherent categories, accuracy has two separate aspects that represent the correctness of data in regard to syntactic and semantic features, respectively (syntactic accuracy and semantic accuracy). Completeness signifies that data have values for all expected attributes and related entity instances. Consistency means being coherent with other data and free from contradiction. Credibility signifies being regarded as true and believable, and includes the concept of authenticity (the truthfulness of origins, attributions, and commitments). Currentness is the degree to which data have attributes that are of the right age.

The first of the three system-dependent aspects is availability; it is based on attributes that enable data to be retrieved by authorized users and/or applications. Portability enables data to be installed, replaced, or moved from one system to another, preserving its existing quality. Recoverability enables data to maintain and preserve a specified level of operations and quality, even in the event of failure. From the group featuring both inherent and system-dependent components, accessibility enables data to be accessed in varied, specific contexts of use. Compliance means adherence to standards, conventions, or regulations in force, related to data quality. Together with aspects of information, such as security, availability, and integrity, confidentiality ensures that data is only accessible to and interpretable by authorized users. Efficiency signifies that it is possibility to process data and provide the expected levels of performance. Precision measures the exactness of data attributes, while traceability is the degree to which data has attributes that provide an audit trail of access to the data. Understandability enables data to be read and interpreted by users. Some information about data understandability is usually provided by metadata [60]. Some inherent data quality attributes are described by Strong, Lee, and Wang as intrinsic DQ [43].

4.5. Big Data Quality and Smart Data

Most quality attributes generally applied to research data are apparently valid for big data as well. Nonetheless, in the case of big data there might be little understanding of what the data actually mean and in which context data are collected; therefore, decision-making quality may be jeopardized [38].

The 3Vs, traditionally attached to big data, i.e., volume, velocity, and variety, can be supplemented by value and veracity. The latter one is of distinguished importance to DQ because it includes consistency (or certainty) and trustworthiness, which in the case of big data is defined not only by the origin of data, but by the methods of its collection and processing. Consistency of big data can be defined by its statistical reliability. Big data veracity ensures that data is trusted, both by its provenance and on the storage side of its lifecycle. Veracity also means that data are authentic and protected from unauthorized access and modification. It also requires:

- The integrity of data;
- The identification of both data and source;
- The trustworthiness of computers and storage platforms;
- Availability and timeliness;
- Accountability;
- Reputation [61].

Figure 2 shows the relationship between the attributes of veracity, as enumerated by Demchenko et al. [61], and trust, as conceived by Buckland [49].

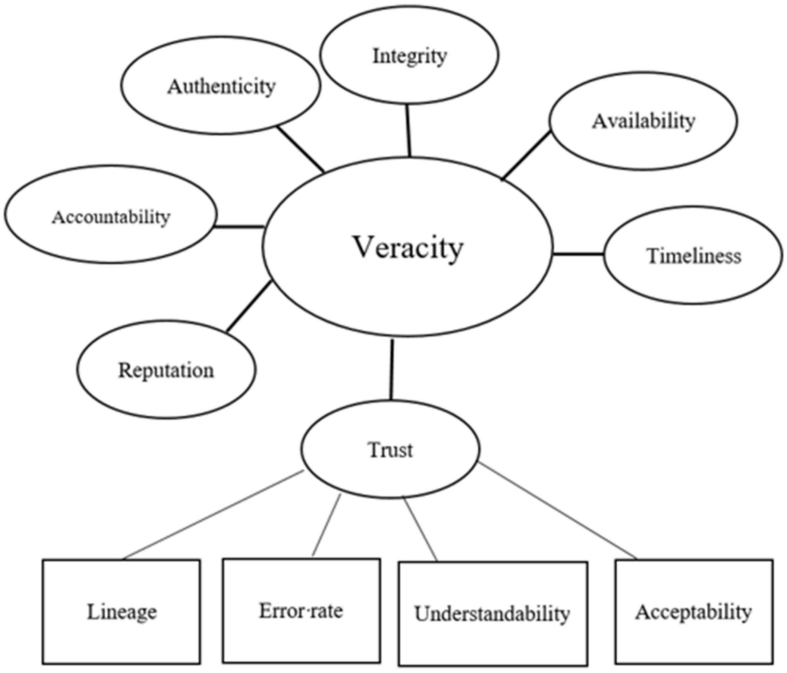

Figure 2. The relationship between veracity and trust.

In the above figure, the vertical line connecting veracity and trust does not mean that any of them would be subordinate to the other. They are rather co-equal.

To predict data quality in large unstructured datasets, it is mostly necessary to analyze the entire dataset, firstly of all because there is unpredictability of "correct" or "expected" data values. This fact also hampers data optimization at the same time as the data collection is performed [62].

Typically, vast amounts of heterogeneous (big) data are integrated by means of data lake architectures; i.e., systems or repositories of data, stored in its original, natural, or raw format. Data lakes often provide low quality contents from various data sources; thus, they need to be cleaned [63].

This reminds us of Frické's statement that "science needs more theories and less data" [64] (p. 660), implying what has been expressed by Halme, Komonen, and Huitu [65]; i.e., focusing on quality can be more useful than allowing quantity to dominate. One of the reasons for stating that, is that data processing in big data environments is not performed on clean datasets from well-known and limited sources, but big data may originate in many different sources, which are not always consistent and

verifiable [66]. Depending on the methodology and the purpose of reusers, even datasets that are from trustworthy sources need some form of data cleansing before data can be processed.

The digital humanities may offer smart data, which are defined less by size than by the fact that their creation involves human agency, meaning that such data are actively and purposefully created by people for interpretation. Prototypical smart data are scholarly digital editions, which are semi-structured, following a data model expressed in a schema that also enables flexibility. Smart data could become a type of data that may serve as an alternative to the big data–small data opposition. It also has the capability to reduce homogeneity, and thereby diminish the occurrence of incorrect conclusions [67]. This prospect might affect approaches to big data quality.

Due to the distinct challenges offered by big data, data governance (to be discussed below) is urgently important, especially in environments where data integrity may be at risk [68].

In general, the incomplete and often uncertain nature of big data negatively influences their quality, making them a source of risk. All in all, future research should determine what metrics of quality are adequate to it and how accurate big data need to be [37].

5. Data Governance

Although several definitions of data governance underline that it is exercising control over the data within corporate environments [69], it can be applied to research data management. This is an example of the reciprocal relationship between the corporate sector and academia, mentioned earlier [70].

Data governance specifies a cross-functional framework that formalizes data policies by describing standards and procedures, and monitoring compliance. It comprises the definition of roles, responsibilities, metrics, and management processes, and the continuous measurement of quality levels. The relationship between data governance and data quality strategy is based on a reciprocal whole–part relationship that we could call also a mutual dependence. Nonetheless, we have to emphasize that data governance has the potential to improve DQ, because it can increase the level of numerous quality factors. For instance—by creating risk-mitigating policies, proper data governance enables caring for the management of risks that may arise due to non-conformance with information policies or the lack of control [37].

The five main domains of data governance as identified by Khatri and Brown [71] are the following ones:

- Clarifying the role of data as an asset;
- Safeguarding DQ;
- Defining and providing metadata;
- Specifying access requirements;
- Determining the definition, production, retention, and retirement of data.

Data governance can refer to organizational bodies; organizational bodies' rules, policies, standards, decision rights, and accountabilities; and the methods of enforcement. It enables better decision-making and protects stakeholders' needs. Reducing operational friction and encouraging the adoption of common approaches to data issues are some of the roles fulfilled by data governance, which—governed by the principles of integrity, transparency, and auditability—also facilitates building repeatable, standard processes. It should increase effectiveness through the coordination of efforts and by enabling transparency of processes. It is based on agreed-upon models, describing who can take what actions, when and under what circumstances, using what methods [72]. It specifies a cross-functional framework that formalizes data policies by describing standards and procedures and monitoring compliance [37]. It is important to underline that data governance should not be optional, because it runs horizontally among others by defining the nature of data, and is use in the data management process. Practical definitions of the data and ways of using them fall in the domain of data management, which involves

integrating data into the organization and regulating them, defining responsibilities for overseeing the administration of data processes pertain to data governance [69]

Successful data governance must guarantee that data can be trusted and assures that responsibility will be taken by people who can be made accountable for poor quality [73]. This implies that the success of an information governance program depends on robust DQ that can be achieved if we reduce the proliferation of incorrect or inconsistent data by continuous analysis and monitoring [74].

By creating risk-mitigating policies, data governance should care for the management of risks that may arise due to non-conformance with information policies or the lack of control [37].

Correct and efficient governance depends as much on technology as on organizational culture. Obviously, this does not preclude making data transparent by good governance technology. On the contrary, technology helps in identifying areas where performance can be improved [75]. Notwithstanding accountabilities for managing data quality, it should not be exclusively assigned to information technology departments [76]. Neither should data governance be treated informally only and in ambiguous ways, as it was done in earlier times by the preponderance of information technology [18].

Data governance for traditional IT environments could be built upon six dimensions that control:

- The functions of data governance and its structure;
- Organization;
- Technical issues;
- Environmental factors;
- Measuring and monitoring tools.

However, the nature of data governance may be influenced by the growing importance of cloud computing. The most important differences between the two types is that all traditional data governance policies are handled in-house, while in the case of cloud-based governance it is up to a third party to ensure that guidelines are followed. On-site infrastructure characterizes the former, in contrast to the multi-site working of the latter. Governance, exercised by a third party, involves loss of control to a certain degree, if we compare it to protecting data in-house [18].

In big data environments, timeliness plays an important role. It is defined as the time delay within which meaningful analysis can take place; thus data should be available from data generation and acquisition to utilization as soon and possible. In other words, data must be prepared in a timely manner, without forgetting about the needs for consistency, reliability, trust in its source, and meaningfulness of the results [77].

Data stewardship, sometimes used interchangeably with data governance is also directed towards ensuring data quality. Data stewards take care of data assets that do not belong to the stewards themselves; therefore, they represent the concerns of others, and ensure that data-related work is performed in accordance with policies and practices as determined through governance. On the other hand, data governance is an overall process that brings together cross-functional teams to make interdependent rules or to resolve issues and to provide services to data stakeholders [78].

6. Quality Beyond Characteristics and Attributes

As stated at the beginning of this paper—good quality research data is a benefit to an ecology of scientific work that follows transparent policies. In the case of publicly funded research, following FAIR Data Principles and managing data quality coincide, because published data is "a function of its ability to be accurately and appropriately found, reused, and cited over time, by all stakeholders, both human and mechanical." [79].

If we take aside the level of trust resulting from the inherent, built-in openness and transparency, which are assumed as basic features of scholarly research [48], critical discussions around data reveal risks and pitfalls [14]. While quantitative measures point towards numerical equivalencies, these

numbers may prove unreliable without (unquantified) context and interpretation [80]. All these deliberations underpin the need for qualitative critical approaches to data quality.

7. Conclusions

The arguments of this is paper have been deliberately restricted to stakeholders of scholarly research on two levels: the intermediate and the micro levels. However, we have to acknowledge that knowledge on data quality might serve a wider audience of researchers. Moreover, there is interest beyond these communities; e.g., from those involved in citizen science. The interests of these communities show both diversity and convergence if compared to research data quality as treated in this paper. This means that a future discussion of data quality should not be limited to characteristics and attributes of research data. As a part of this thinking, we can say that the seemingly endless list of data quality attributes, having stemmed to a substantial extent in business studies, supported by proper data governance, needs to be supplemented by examining the critical issues outlined above.

Acknowledging that data is a crucial element of scholarly research profoundly influences the perspectives on data quality, the leading stakeholders of which are not only the researchers.

Funding: This research received no external funding,

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bueno de la Fuente, G. What is Open Science? Introduction. Available online: https://www.fosteropenscience.eu/content/what-open-science-introduction (accessed on 29 January 2020).
2. Royal Society. *Science as an Open Enterprise*; Royal Society Science Policy Centre: London, UK, 2012.
3. Borgman, C.L. *Big Data, Little Data, no Data: Scholarship in the Networked World*; MIT Press: Cambridge, MA, USA, 2015.
4. Pryor, G. *Managing Research Data*; Facet Publishing: London, UK, 2012.
5. Semeler, A.R.; Pinto, A.L.; Rozados, H.B.F. Data science in data librarianship: Core competencies of a data librarian. *J. Libr. Inf. Sci.* **2017**, *51*, 771–780. [CrossRef]
6. *The IFRS for SMEs*; International Accounting Standards Board: London, UK, 2015.
7. Ramírez, M.L. Opinion: Whose role is it anyway? A library practitioner's appraisal of the digital data deluge. *Bull. Am. Soc. Inf. Sci. Technol.* **2011**, *37*, 21–23. [CrossRef]
8. Foster, J.; McLeod, J.; Nolin, J.; Greifeneder, E. Data work in context: Value, risks, and governance. *J. Assoc. Inf. Sci. Technol.* **2018**, *69*, 1414–1427. [CrossRef]
9. Neylon, C.; Belsø, R.; Bijsterbosch, M.; Cordewener, B.; Foncel, J. *Open Scholarship and the Need for Collective Action*; Knowledge Exchange: Bristol, UK, 2019.
10. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [CrossRef]
11. Makani, J. Knowledge management, research data management, and university scholarship: Towards an integrated institutional research data management support-system framework. *Vine* **2015**, *45*, 344–359. [CrossRef]
12. Frické, M. The knowledge pyramid: A critique of the DIKW hierarchy. *J. Inf. Sci.* **2009**, *35*, 131–142. [CrossRef]
13. Yu, L. Back to the fundamentals again: A redefinition of information and associated LIS concepts following a deductive approach. *J. Doc.* **2015**, *71*, 795–816. [CrossRef]
14. Špiranec, S.; Kos, D.; George, M. Searching for critical dimensions in data literacy. *Inf. Res.* **2019**, *24*. Available online: http://InformationR.net/ir/24-4/colis/colis1922.html (accessed on 29 January 2020).
15. Golub, K.; Hansson, J. (Big) Data in Library and Information science: A brief overview of some important problem areas. *J. Univers. Comput. Sci.* **2017**, *23*, 1098–1108. [CrossRef]
16. Heidorn, P.B. The emerging role of libraries in data curation and e-science. *J. Libr. Admin.* **2011**, *51*, 662–672. [CrossRef]

17. Sans, S.; Night, W.S. Introduction. In HathiTrust: Large-Scale Data Repository in the Humanities. 2019. Available online: https://akthom.gitboooks.io/hkdcws/content/Introduction/Introduction.html (accessed on 29 January 2020).
18. Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. A systematic literature review of data governance and cloud data governance. *Pers. Ubiquitous Comput.* **2018**, *23*, 839–859. [CrossRef]
19. Baskarada, S.; Koronios, A. Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Aust. J. Inf. Syst.* **2013**, *18*, 5–24. [CrossRef]
20. Whitmire, A.L.; Boock, M.; Sutton, S.C. Variability in academic research data management practices: Implications for data services development from a faculty survey. *Program* **2015**, *49*, 382–407. [CrossRef]
21. RECODE Policy Recommendations for Open Access to Research Data. RECODE Project Consortium. Available online: https://trilateralresearch.co.uk/wp-content/uploads/2018/09/RECODE-D5.1-POLICY-RECOMMENDATIONS-FINAL.pdf (accessed on 29 January 2020).
22. Hrynaszkiewicz, I. Publishers' Responsibilities in Promoting Data Quality and Reproducibility. In *Handbook of Experimental Pharmacology*; Springer: Berlin/Heidelberg, Germany, 2019.
23. Kim, J. Who is teaching data: Meeting the demand for data professionals. *J. Edu. Libr. Inf. Sci.* **2016**, *57*, 161–173. [CrossRef]
24. Corti, L.; Van den Eynden, V.; Bishop, L.; Woollard, M. *Managing and Sharing Research Data: A Guide to Good Practice*; SAGE Publications Limited: Southend Oaks, CA, USA, 2019.
25. Schneider, R. Research data literacy. In *European Conference on Information Literacy*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 134–140.
26. Hobbs, R. Multiple visions of multimedia literacy: Emerging areas of synthesis. In *International Handbook of Literacy and Technology*; Routledge: London, UK, 2006; Volume 2, pp. 15–28.
27. Candela, L.; Castelli, D.; Manghi, P.; Tani, A. Data journals: A survey. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 1747–1762. [CrossRef]
28. Lyon, L.; Brenner, A. Bridging the data talent gap: Positioning the iSchool as an agent for change. *Int. J. Dig Curation* **2015**, *10*, 111–122. [CrossRef]
29. Swan, A.; Brown, S. The skills, role, and career structure of data scientists and curators: An assessment of current practice and future needs. 2008, Report to the JISC. Available online: http://eprints.soton.ac.uk/266675 (accessed on 29 January 2020).
30. Hayashi, C. What is data science? Fundamental concepts and a heuristic example. In *Data Science, Classification, and Related Methods*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 40–51.
31. Voulgaris, Z. *Data Scientist: The Definitive Guide to Becoming a Data Scientist*; Technics Publications: Basking Ridge, NJ, USA, 2014.
32. Virkus, S.; Garoufallou, E. Data science from a library and information science perspective. *Data Tech. Appl.* **2019**, *52*, 422–441. [CrossRef]
33. Wang, L. Twinning data science with information science in schools of library and information science. *J. Doc.* **2018**, *74*, 1243–1257. [CrossRef]
34. Cao, L. Data science: Nature and pitfalls. *IEEE Intell. Syst.* **2016**, *31*, 66–75. [CrossRef]
35. Madrid, M.M. A study of digital curator competences: A survey of experts. *Intern. Inf. Libr. Rev.* **2013**, *45*, 149–156. [CrossRef]
36. Nahm, M.; Hammond, W.E. Data standards≠ data quality. *Stud. Health Technol. Inf.* **2013**, *192*, 1208.
37. Abraham, R.; Schneider, J.; vom Brocke, J. Data governance: A conceptual framework, structured review, and research agenda. *Int. J. Inf. Manag.* **2019**, *49*, 424–438. [CrossRef]
38. Janssen, M.; van der Voort, H.; Wahyudi, A. Factors influencing big data decision-making quality. *J. Bus. Res.* **2017**, *70*, 338–345. [CrossRef]
39. Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. Data Governance Taxonomy: Cloud versus Non-Cloud. *Sustainability* **2018**, *10*, 95. [CrossRef]
40. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
41. Altman, M.; Marciano, G.; Lee, C.; Bowden, H. Mitigating threats to data quality throughout the curation lifecycle. In *Curating For Quality: Ensuring Data Quality to Enable New Science*; National Science Foundation: Alexandria, VA, USA, 2012; pp. 1–119.

42. Sposito, F.A. What do Data Curators Care About? Data Quality, User Trust, and the Data Reuse Plan. Paper Presented at IFLA WLIC 2017. Available online: http://library.ifla.org/1797/ (accessed on 29 January 2020).
43. Strong, D.M.; Lee, Y.W.; Wang, R.Y. Data quality in context. *Commun. ACM* **1997**, *40*, 103–110. [CrossRef]
44. Daraio, C.; Lenzerini, M.; Leporelli, C.; Naggar, P.; Bonaccorsi, A.; Bartolucci, A. The advantages of an Ontology-Based Data Management approach: Openness, interoperability and data quality. *Scientometrics* **2016**, *108*, 441–455. [CrossRef]
45. Laranjeiro, N.; Soydemir, S.N.; Bernardino, J. A survey on data quality: Classifying poor data. In Proceedings of the 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), Zhangjiajie, China, 18–20 November 2015; pp. 179–188.
46. Wolski, M.; Howard, L.; Richardson, J. A Trust Framework for Online Research Data Services. *Publications* **2017**, *5*, 14. [CrossRef]
47. Giarlo, M.J. Academic libraries as data quality hubs. *J. Libr. Sch. Commun.* **2013**, *1*, 1–10. [CrossRef]
48. Rieh, S.Y. Judgment of information quality and cognitive authority in the Web. *J. Am. Soc. Inf. Sci. Technol.* **2002**, *53*, 145–161. [CrossRef]
49. Buckland, M. Data management as bibliography. *Bull. Am. Soc. Inf. Sci. Technol.* **2011**, *37*, 34–37. [CrossRef]
50. Yoon, A. End users' trust in data repositories: Definition and influences on trust development. *Arch. Sci.* **2014**, *14*, 17–34. [CrossRef]
51. Miller, H. The multiple dimensions of information quality. *Inf. Syst. Manag.* **1996**, *13*, 79–82. [CrossRef]
52. Deutsche Klimarechenzentrum, Quality Assurance of Data. Available online: https://www.dkrz.de/up/services/data-distribution/data-publication/quality-assurance-of-data (accessed on 29 January 2020).
53. Colepicolo, E. Information reliability for academic research: Review and recommendations. *New Libr. World* **2015**, *116*, 646–660. [CrossRef]
54. King, G. Replication, replication. *Polit. Sci. Polit.* **1995**, *28*, 444–452. [CrossRef]
55. Yoon, A.; Lee, Y.Y. Factors of trust in data reuse. *Online Inf. Rev.* **2019**. [CrossRef]
56. Borgman, C.L. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*; MIT Press: Cambridge, MA, USA, 2010.
57. Faniel, I.M.; Kriesberg, A.; Yakel, E. Social scientists' satisfaction with data reuse. *J. Assoc. Inf. Sci. Technol.* **2015**, *67*, 1404–1416. [CrossRef]
58. Zilinski, L.D.; Nelson, M.S. Thinking critically about data consumption: Creating the data credibility checklist. *Proc. Am. Soc. Inf. Sci. Technol.* **2014**, *51*, 1–4. [CrossRef]
59. Zimmerman, A.S. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Sci. Technol. Hum. Values* **2008**, *33*, 631–652. [CrossRef]
60. ISO/IEC 25012. Available online: https://iso25000.com/index.php/en/iso-25000-standards/iso-25012 (accessed on 29 January 2020).
61. Demchenko, Y.; Grosso, P.; De Laat, C.; Membrey, P. In addressing big data issues in scientific data infrastructure. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 48–55.
62. Dresp, B.; Dresp-Langley, B.; Ekseth, O.K.; Fesl, J.; Gohshi, S.; Kurz, M.; Sehring, H.-W. Occam's Razor for Big Data? On detecting quality in large unstructured datasets. *Appl. Sci.* **2019**, *9*, 3065. [CrossRef]
63. Ceravolo, P.; Azzini, A.; Angelini, M.; Catarci, T.; Cudré-Mauroux, P.; Damiani, E.; Mazak, A.; Van Keulen, M.; Jarrar, M.; Santucci, G. Big data semantics. *J. Data Semant.* **2018**, *7*, 65–85. [CrossRef]
64. Frické, M. Big data and its epistemology. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 651–661. [CrossRef]
65. Halme, P.; Komonen, A.; Huitu, O. Solutions to replace quantity with quality in science. *Trends Ecol. Evol.* **2012**, *27*, 586–588. [CrossRef]
66. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Khan, S.U. The rise of "big data" on cloud computing: Review and open research issues. *Inform. Syst.* **2015**, *47*, 98–115. [CrossRef]
67. Schöch, C. Big? Smart? Clean? Messy? Data in the Humanities. *J. Dig. Hum.* **2013**, *2*, 2–13.
68. Breaking Big: When Big Data Goes Bad. In *The Importance of Data Quality Management in Big Data Environments*; Information Builders: New, York, NY, USA, 2014.
69. Smith, A.M. Data Governance Best Practices—The Beginning. *EIM Insight* **2013**, *1*. Available online: http://www.eiminstitute.org/library/eimi-archives/volume-1-issue-1-march-2007-edition/data-governance-best-practices-2013-the-beginning (accessed on 29 January 2020).

70. Willoughby, S. Open data and the environment. In *The State of Open Data: Histories and Horizons*; Davies, T., Walker, S., Rubinstein, M., Perini, F., Eds.; African Minds and International Development Research Centre: Cape Town, South Africa; Ottawa, ON, Canada, 2019; pp. 103–118.
71. Khatri, V.; Brown, C.V. Designing data governance. *Commun. ACM* **2010**, *53*, 148–152. [CrossRef]
72. DGI Definitions of Data Governance. Available online: http://www.datagovernance.com/adg_data_governance_definition/ (accessed on 29 January 2020).
73. Sarsfield, S. *The Data Governance Imperative: A Business Strategy for Corporate Data*; IT Governance Publishing: Ely, UK, 2009.
74. IBM. *Successful Information Governance through High-Quality Data*; IBM Corporation: Armonk, NY, USA, 2012.
75. ORACLE. *The Five Most Common Big Data Integration Mistakes to Avoid*; Oracle Corporation: Redwood City, CA, USA, 2015.
76. Weber, K.; Otto, B.; Österle, H. One size does not fit all—A contingency approach to data governance. *J. Data Inf. Qual.* **2009**, *1*, 4. [CrossRef]
77. Al-Badi, A.; Tarhini, A.; Khan, A.I. Exploring Big Data Governance Frameworks. *Procedia Comput. Sci.* **2018**, *141*, 271–277. [CrossRef]
78. Rosenbaum, S. Data governance and stewardship: Designing data stewardship entities and advancing data access. *Health Serv. Res.* **2010**, *45*, 1442–1455. [CrossRef]
79. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
80. Seaver, N. The nice thing about context is that everyone has it. *Media Cult. Soc.* **2015**, *37*, 1101–1109. [CrossRef]

© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).