Data Quality Management

Peer-reviewed author version

VANCAUWENBERGH, Sadia (2019) Data Quality Management. In: Suad Kunosic; Enver Zerem (Ed.). Scientometrics Recent Advances, IntechOpen, p. 1-15.

DOI: 10.5772/intechopen.86819
Handle: http://hdl.handle.net/1942/28871

Author Queries

[AQ1]	Please provide publisher location for Refs. [7, 8, 19, 20, 22].
[AQ2]	Please provide page range for Ref. [23].
[AQ3]	Please provide volume number and page range for Ref. [24].

Chapter

Data Quality Management

Sadia Vancauwenbergh

Abstract

Data quality is crucial in measuring and analyzing science, technology and innovation adequately, which allows for the proper monitoring of research efficiency, productivity and even strategic decision making. In this chapter, the concept of data quality will be defined in terms of the different dimensions that together determine the quality of data. Next, methods will be discussed to measure these dimensions using objective and subjective methods. Specific attention will be paid to the management of data quality through the discussion of critical success factors in operational, managerial and governance processes including training that affect data quality. The chapter will be concluded with a section on data quality improvement, which examines data quality issues and provides roadmaps in order to improve and follow-up on data quality, in order to obtain data that can be used as a reliable source for quantitative and qualitative measurements of research.

Keywords: data quality, data quality measurement, data quality management, data quality improvement

1. Introduction

Over the past decades, research organizations, administrations and researchers have been collecting data that describe both the input as well as the output side of research. This has resulted in an enormous pile of data on publications, projects, patents, … researchers and their organizations that are collected within database systems or current research information systems (CRIS). Such data systems are created according to specific goals and use purposes of individual organizations, which reflects their specific nature and the surrounding context in which they operate. However, over time these data systems, institutions as well as the research ecosystem at large have evolved, thereby potentially threatening the quality of the collected data and the resulting data analyses, particularly if no formal data quality management policy is being implemented. This chapter introduces the readers into the concept of data quality and provides methods to assess and improve data quality, in order to obtain data that can be used as a reliable source for quantitative and qualitative measurements of research.

2. Definition of data quality

In general, data can be considered of high quality if the data is fit to serve a purpose in a given context, for example, in operations, decision making and/or planning [1]. Although this definition of data quality seems to be straightforward, many other definitions exist that differ in terms of the qualitative or quantitative approach towards defining the concept of data quality.

2.1 Qualitative approach

In the qualitative approach, specific attention is drawn to defining data quality in terms of the different aspects, also termed dimensions. In 1996, Wang and Strong developed a data quality framework based on a two-stage survey on data quality aspects important to data consumers, and captured these dimensions in a hierarchical manner [2]. This model clusters 20 different data quality dimensions into four major categories: that is, intrinsic, contextual, representational and access data quality. Although the basis of this model still stands, some minor changes have been made over the years resulting in the model depicted in **Table 1** [3].

In brief, the *intrinsic* category comprises dimensions that define the accuracy of the data, that is, the extent to which data is certified, error-free, and reliable, as well as the objectivity of the data based on facts and impartial, and their reputation based on its sources or content. The *contextual* data quality category comprises dimensions that must be considered within the context of a specific objective for which one holds the data, that is, the data should be relevant, up to date, of an appropriate amount, yet complete, and ready for use for the stated objective. The *representational* category contains dimensions that reflect how the data are presented within a data system. Dimensions concerning the format of the data, that is, concise and consistent representation, as well as their compatibility, their interpretability and whether they are easy to understand, are considered. The last category is focused on the *accessibility* category that also defines aspects of data quality. Although this category is not always considered in the literature [4], this is an important aspect of overall data quality. The related dimensions include the accessibility of the data in terms of their availability or easily retrievable character, the security measures taken to restrict data appropriately and the traceability of the data to its source.

Category	DQ dimension
Intrinsic	Accuracy
	Objectivity
	Reputation
Contextual	Completeness
	Appropriate amount
	Value added
	Relevance
	Timeliness
	Actionable
Representational	Interpretable
	Easily understandable
	Consistent
	Concisely represented
	Alignment
Access	Accessibility
	Security
	Traceability

Table 1.
Data quality dimensions.

These dimensions can also be grouped into an internal and external group of dimensions. The internal group contains the dimensions that can be measured purely in terms of the data, and are generally more objective. Examples of these include the accuracy of the data, which can be examined by calculating a score on the magnitude of errors in the data or the data correctness, which can be measured through the number of errors in the data. On the other hand, the external group of dimensions evaluates how the data are related to their environment, and hence are somewhat more subjective in nature. Examples include the relevancy of data with regards to a stated objective, or their ease of understanding by the consumers of the data.

2.2 Quantitative approach

In the quantitative approach, data quality has been defined by J. M. Juran as the fitness of the data to serve a purpose in a given context, that is, in operations, decision making and/or planning as perceived by its users [1]. This concept is denoted as 'fitness for use' and is based on Juran's five principles: that is, who uses the data, how are the data used, is there a danger for human safety, what are the economic resources of the producers and users of the data and what are the characteristics taken into account by users when determining the fitness for use. This definition is widely accepted in both academic and industrial settings. However, in practice the fitness for use is a rather subjective measure as this highly depends on the users' judgement over the degree of conformity of the data to their intended use.

For example, consider the score of a student on an exam. If scores are rounded to integers, this can potentially influence the final grade that a student receives. Therefore, the rounding procedure might be accurate enough for the professors, but by rounding numbers, the students might miss out on obtaining a final grade and thus might be not accurate enough from the perspective of the student.

On the other hand, it might well be that not all uses of the data are known, neither its potential future use purposes. Therefore, DQ might be hard to evaluate using this definition.

Some definitions of data quality use the notion of zero defects, which aims to reduce defects by motivating people to prevent making mistakes by developing a constant, conscious desire to do the job right from the first time [5]. This zero-defect concept has been incorporated by P. Crosby in its Absolutes of Quality Management [6]. According to Crosby's Absolutes, data quality should conform to its requirements and prevention should be used as a manner to guarantee zero defects, which sets the performance standard. Consequently, data quality can be measured as the price of nonconformance. Although this zero-defect concept is not widely used in the data quality literature, it does emphasize again the necessity to measure data quality.

3. Measuring data quality

Based on the definitions of data quality, several DQ measurement methods have been developed, that can generally be divided into objective and subjective methods. While objective methods tend to evaluate data quality rather from the perspective of the data producer based on hard criteria, subjective methods rather take the user's perspectives and beliefs into account.

3.1 Objective DQ measurement methods

Measurements of data quality are generally intended to assess the dimensions of data quality as defined in the previous section. As a first step, a framework must be set up with the indicators that one wants to assess. Next, a proper reference for verification of the data within the data systems must be determined.

Ideally, the data are compared using real world data, which allows for validation and, if required immediate corrective actions. This method is termed ***data auditing*** and is the only way of measuring the quality level of dimensions like accuracy, completeness. Furthermore, by going through the data itself, one can discover data quality issues that were unexpected and therefore are of great value for taking corrective measures to improve data quality. However, data auditing comes at a high cost as it is very time consuming and the need of experts in the respective field is required. Furthermore, data auditing can be also very labor-intensive and requires that data controllers have access to the actual data.

For example, consider the metadata of publications that are contained in publication databases. If a data controller validates the content of the metadata fields with the metadata as indicated on the publications, inaccuracies can be detected. These can contain expected flaws like spelling errors but can also provide valuable information on unexpected errors that also might be highly relevant in the context of bibliometric analyses.

If the conditions for data auditing are not met, data controllers can ***use rule-based checking*** in order to determine data quality. This method heavily relies on business rules that are drafted based upon the domain knowledge and experience that the data controllers have with regards to the data. Consequently, these rules can only check for flaws that were anticipated by the data controllers. However, rule-based checking also offers important advantages, especially as they can be automated after conversion to validation rules, which allows for the identification of the errors (or possibly correct outliers!) via data mining techniques. Nevertheless, the presumed errors still need to be corrected, which remains labor-intensive.

3.2 Subjective DQ measurement methods

Some dimensions, however, cannot be measured objectively because of their intrinsic properties. For example, the dimension relevancy pertains to the extent to which data is applicable and helpful for the stated objective. Obviously, this dimension can only be evaluated using the ***perception of the users***. Although this results in a subjective scoring, user evaluations are the only way to measure dimensions that describe external data quality attributes. Internal data quality dimensions on the contrary are preferably measured using objective DQ measurement methods as described above.

Regardless of which methodology is chosen to measure data quality, it is always important to provide information about the measurement method and parameters in addition to the dimension under evaluation, in order that the measurement results can be interpreted correctly by everyone. Furthermore, although a lot of attention always goes to correcting errors, it is important to stress that eliminating the root cause should always be the ultimate goal [7].

4. Data quality management

4.1 Data quality frameworks

As data are extremely valuable resources in today's society, a plethora of data quality management frameworks have been published in the last decades that all strive to preserve the quality of data and to make it accessible for future use. The most popular models are listed below, however more DQM frameworks can be found throughout the literature that show slight differences.

- DAMA DMBOK's Data governance model [8]
- EWSolutions' EIM Maturity Model [9]
- Oracle's Data Quality Management Process [10]

All frameworks are basically centered around three basic elements, that is, the metadata associated with the data, the processes involved in the registration, organization and (re)use of the data, and the organizational context in relation to the data (**Figure 1**). The quality of each individual element, as well as the interplay in between them, ultimately determines the quality and thus the true value of an organization's data heritage. Ideally, an organization uses metadata standards that are understandable throughout the organization and aligned with the organization's processes, business strategies and goals. Rather than describing all popular frameworks, we will describe critical success factors that are useful for developing effective DQ management strategies, and that can be found in all DQ frameworks.

4.2 Critical success factors

Critical success factors, also termed CSFs, have been defined by Milosevic and Patanakul as '*characteristics, conditions, or variables that can have a significant impact on the success of i.e., a company or a project when properly sustained, maintained, or managed*' [11]. In 2014, Baskarada described 11 CSFs in the field of information quality management that provide valuable means for developing effective DQ management strategies [12]. These CSFs can be clustered into four major groups, that is, training, governance, management and operational processes, that have inter-dependencies with each other.

4.2.1 Operational processes

The first group of critical success factors deals with the operational processes involved in the collection, storage, analysis and security of the data, which are all highly interdependent. As data is a valuable good, its quality should be managed throughout its entire lifecycle. In practice this comes down to taking measures that maximize, whenever possible, the **automated capture** of data in **real-time**, directly from its **original source**. This minimizes the risk of errors introduced by manual data entry, which can result in typo's, inaccuracies, missing values, erroneous data

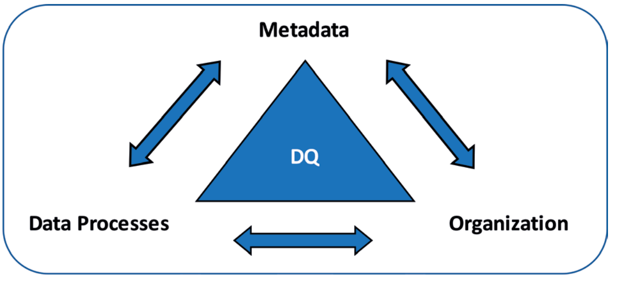

Figure 1.
The cornerstone of data quality frameworks.

due to misinterpretations, multiple copies of the same data entry. Such errors have been identified in almost all existing research and innovation databases, but have a significant impact on the resulting scientometric analyses. Suppose a highly cited paper is included in the Web of Science with typo's in the author's name. This can erroneously lead to the omission of this paper in the bibliometric analyses performed on this author, which on its turn can have a major impact on this researcher career perspectives in terms of chances of success in obtaining grants, promotion.

In addition, these errors can be due to a lack of the use of common **standards** for the concepts contained within the databases and a uniform interpretation thereof by both information providers as well consumers throughout the entire organization. Nevertheless, such standards are available, that is, the Common European Research Information Format (CERIF) is a well-known standard for exchanging research information created by the EuroCRIS organization and is widely used throughout Europe [13], the CASRAI dictionary is a standard created by the organization on Consortia Advancing Standards in Research Administration Information (CASRAI) and was created in Canada [14]. Although both communities work closely together to align the concepts and meanings described in the standards, some differences remain which might cause difficulties in exchanging information in between CRIS systems. Furthermore, the inclusion of a standard in the information model of a data system does not safeguard that all data providers use the standard similarly, nor that the data users grasp the information as intended. Next to using standards for aligning the concepts and meanings of research-related data, the formats of the data fields should be standardized as well. A well-known example here includes the various formats in which a (publication) date is recorded. By means of standardizing this format in a data system, important gains can be obtained in terms of ease of interpretation of the data, leading to more accurate analyses. However as described above, efforts should also be made to clarify what the concept of (publication) date means. For instance, it could point to the creation date, submission date, the published online date, the publication date for in print papers, the date on which the material was made available.

Furthermore, when storing research-related data, it is highly recommended to provide **traceability** to the raw data, which ensures that the data quality can always be controlled. Most bibliometric databases, including the Web of Science and Scopus, comply to this rule by providing a link to the journal article. Research data repositories mostly refer to the creator of the datasets involved. However, over time, researchers can switch positions and thus institutions and as the data are stored in institutional repositories, it would be more meaningful to refer to the research institution in question. In addition, **versioning** should be included when storing research data, as this can be very helpful to understand and potentially (re)use data. Although this is frequently observed in research data repositories, bibliometric and patent databases usually do not show version control. Finally, **back-up** and **data recovery** processes should be ensured when storing research-related data, which is mostly realized via back-up servers at various physical places.

The access to research information should be managed using an **information security management** plan in order to safeguard the intellectual property rights of the researchers that created the information, including their respective institutions. Although large data repositories on bibliometric, innovation and research data control accessibility rights, researchers themselves do not always closely follow the measures taken to control access. Particularly when it comes down to research data that may contain sensitive data [15], strict follow-up of information security measures is needed as emphasized by the EU Regulation 2016/679, also known as the General Data Protection Regulation (GDPR) that protects natural persons with regards to the processing of personal data and on the free movement of such data.

DOI: http://dx.doi.org/10.5772/intechopen.86819

Although the GDPR regulation only applies to personal data *in se*, it nicely underpins some elements present in information security management plans.

These information security management plans indeed not only entail the accessibility rights of individuals, including user authentication and a regular update of their access rights, but also include the secure storage, archival, transmission, and if required, destruction of the information. In case of research data on natural persons, this can be achieved via pseudonymization, for example, through encryption, or via anonymization of the research information residing in data systems or on data carriers. Obviously, when transmitting research information, the proper legal agreements should be put in place, for example, non-disclosure agreements with third parties are well-known examples used to secure research information. Finally, information security management plans should also contain audit trails in order to constantly monitor and adjust the security of research-related information.

4.2.2 Management processes

A second group of CSFs encompasses the managerial processes that are imposed on these operational processes, and which are primarily aimed at the alignment of the data quality with the organization's goals with regards to the data and the resulting data analyses. Consider for example, the information requirement of a university that wants to monitor the research funds obtained via researchers. In order to answer this question, the concepts of research funds and researchers should be clear and uniform between information providers and users. Although this might seem straightforward, it could well be that the interpretation of 'researcher' is different in between stakeholders, that is, while some might include PhD students, other might omit this group. Furthermore, it could well be that the university does not have a specific label for clustering funds as belonging to the 'research' category, or that the information is only partially provided by the researchers. These examples clearly illustrate that the lack of management of operational DQ processes, has a devastating effect on the data analyses and the conclusions based thereon.

Managerial processes of data quality essentially focus on four sequential processes, that is, the determination of the information quality requirements, the assessment of the risks associated with DQ issues, the assessment or monitoring of DQ and the continuous improvement of the related DQ processes [16]. First, the **information quality requirements** should be determined of the collected data, considering all stakeholders. Next, a conceptual information model should be drafted using high-level data constructs, generally described in non-technical terms in order to be understandable by executives and managers. This model should then be translated into a logical data model that uses entities, attributes and relationships that are customized towards the organization's use of the data, in terms of the organization's terminology, semantics as well as the prevailing business rules. Finally, the logical model should be transferred to developers that can derive a physical data model in line with this logical model including validation rules, based upon the business rules, that are useful for automating data quality control. Obviously, the constructed models must consider the importance of the data within the organization. For example, certain data will be more important than others, and poor DQ of those data might have a larger negative impact in terms of loss of reputation, financial loss. of the organization. The explicit **management of these DQ risks** is a must as a manner to guarantee data quality. As stated by Baskarada '*using gut feeling will result in inefficiency and an ineffective use of resources*' [16].

Next, a framework of key performance DQ indicators needs to be set up in line with the organization's goals, in order to **assess the DQ performance**. This assessment must be performed on a regular basis in order to allow for the **continuous**

improvement of data quality in terms of analyzing the root cause of the errors as well as cleansing erroneous data.

The application of such DQ managerial processes has already been implemented to some extent in CRIS systems that contain research information. For example, the Flanders Research Information Space, also termed FRIS, is a research information portal sustained by the Department of Economy, Science and Innovation in Flanders, Belgium that collects research information from a wide range of Flemish stakeholders in the research field, that is, research universities, higher education colleges, strategic research centers and research institutions (www.researchportal.be) [17]. Underlying the FRIS architecture, a conceptual metamodel was developed in order to model all concepts, attributes and relationships that are contained within FRIS. This conceptual model is based on the CERIF standard, but customized to the Flemish context. In addition, in line with the use purposes of this CRIS system, business rules were drafted to safeguard the quality of the contained information. These business rules were translated to validation rules that are used for the automated quality control of the research information received. If non-compliances to these rules are detected, the research information is rejected, and the information providers receive a notification thereby allowing for immediate data cleansing. Furthermore, the Flemish government also performs manual quality checks on a regular basis in order to validate the research information contained as validation rules in general are not well suited for detecting unpredicted errors. Such errors generally provide valuable input for root cause analyses that can identify important underlying problems which can be caused by human, process, organizational or technological factors.

4.2.3 Governance process

A third group of CSFs encompasses the governance processes associated with DQ management. These processes can be largely summarized as the **commitment of an organization's top management** to set DQ management as a priority and to stimulate a culture change throughout the entire organization in this respect. In the field of information governance, Gartner Research defined information governance as '*the specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archival and deletion of information*' [18]. In practice, information governance basically comes down to allocating budget and resources to the process of DQ management by defining roles and responsibilities, making agreements on related concepts, terms and associated DQ processes, including the monitoring, control and improvement thereof. The FRIS-system as indicated above has included data governance in order to ensure proper DQ management [17].

4.2.4 Training

Although an organization might have all operational, managerial and governance processes perfectly in place, a complete implementation of DQ management also requires the investment in training throughout the organization. A first and foremost important goal is to inform people on the importance of qualitative data to the organization. Secondly, people should receive training via training programs, course series, mentorships. on the rules as set out in the operational, managerial and governance processes in order to ensure a systematic implementation of DQ throughout the entire organization. Finally, a continuous follow-up is also needed which allows for swift adjustments in case of unpredicted errors, adjustment of business rules, etc.

5. Data quality improvement

In order to safeguard the continuous monitoring of data quality and the adoption of measures to improve data quality, a DQ improvement workflow needs to be established. This workflow essentially comprises a repetitive workflow of five consecutive phases, that is, the definition, measurement, analyze, improvement and control phase as depicted in **Figure 2**. A best practice is to formalize this data quality improvement process, in terms of properly documenting all related processes and activities in each phase, as this allows for the tracking of progress throughout the entire DQ improvement workflow.

5.1 Definition of DQ project

The DQ improvement workflow starts with defining the scope of the DQ improvement project. This includes the selection of a dataset relevant to a specific business goal, and the determination of the data attributes required. When collecting this information, it is very important to discuss the meaning of the metadata required with all stakeholders in order to be able to identify any discrepancies in interpretation of the required data attributes versus the meaning of the existing metadata, as this prevents erroneous data collection, analysis and interpretation. All obtained information should be documented using domain modeling techniques that include information on the data and the associated operations on the data [19]. Examples of such techniques include Business Process Model Notation (BPMN) diagrams [20], data flow diagrams. of which the resulting information should be contained in data governance tools together with the accompanying semantics. In addition, data quality dimensions important to the specific use purposes of the data should be determined, and if possible, these are preferably defined in a measurable manner which facilitates further steps in the DQ improvement process.

For example, consider the use of bibliometric data as part of a researcher's evaluation in the context of career-wise promotions. In order to provide an adequate, qualitative data-analysis, a clear framework should be defined by an organization's management comprising what should be evaluated, that is, which publications (books, journals.), validation criteria (peer reviewed, group author.) are to be used as well as the accompanying processes. This information should be discussed with all stakeholders, that is, researchers, librarians, data analysts and IT-staff in order to harmonize the data flow, the accompanying semantics, procedures and models in accordance with the management's goals. Next, the *As Is* situation should be evaluated with regards to these intentions and according to the relevant data dimensions. In bibliometric analyses, accuracy, completeness, timeliness, relevance,

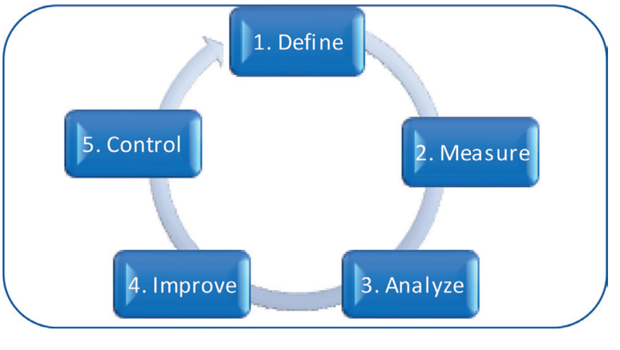

Figure 2.
Data quality improvement workflow.

accessibility, traceability of the data are all relevant dimensions, of which the accurate and complete collection and analysis of a researcher's published works are the foremost ones.

5.2 Measurement DQ

In order to determine the quality level of the current data in relation to the organization's objectives, the quality dimensions need to be expressed in a measurable manner. While the internal dimensions can be scored in a quantitative manner by means of expressing the errors in the data set in terms of magnitude, number of errors or missing records., the external dimensions are measured in a qualitative manner based on the context of the data's use purposes. Independent of the dimension under analysis, measurements must always be relevant for the purpose for which the data will be used and according to the task's requirements. Although in most cases, common sense will be used to identify task requirements, in other cases specific techniques like sensitivity analysis might be used which allows for identifying critical factors and errors in data models [21, 22]. Furthermore, data profiling is another technique frequently used in DQ assessment as a method to discover the true content, structure and quality of data by means of rule-based checking [23]. Obviously, this technique does not find all inaccurate data, as it can only identify violations to the predefined rules, and hence expected errors. For instance, data profiling can identify invalid data values (i.e., using column property analysis), invalid data combinations (i.e., through structure analysis), inaccurate data (i.e., through value rule analysis). Importantly, data profiling also provides metrics on the data inaccuracies in a dataset, that is, the number of violations, the frequency of invalid data values, etc. Such metrics can be useful as a means to communicate to stakeholders on the (in)accuracy of a data set, and the follow-up of the progression in subsequent DQ improvement programs.

In our bibliometric example, the accuracy and completeness of the bibliometric records for a given author, collected in a university's database system should be compared to a publication list provided by the author. By manually auditing the registered data found within the database system, one could indeed record the completeness of information. Furthermore, the accuracy can be tested using a manual auditing procedure. This allows for the identification of spelling errors, erroneous exchange of an author's last versus first name, etc. In addition, manual auditing also allows for identification of rather unexpected data entries, like changes in the author's first or last name over time. The latter example of a DQ inaccuracy, can however not be detected through data profiling as rule-based checking is unable to test for unexpected errors. Nevertheless, data profiling has an important role in DQ measurement as it allows for automated and thus efficient screening of DQ.

5.3 Analyzing DQ issues

Once DQ inaccuracies have been detected, these should be analyzed in order to screen for the potential existence of (groups of) common underlying root causes. For example, author names can have various problems like misspelling, last names mistaken for first names, etc. The grouping of such errors that show similar patterns, also called error cluster analysis, allows for the identification of common causes and is often more efficient in terms of time and resources as compared to handling all inaccuracies in a stand-alone way. In addition, a data event analysis can be performed which evaluates the time points when data are created and updated in order to facilitate the identification of the root causes of problems. For example, the manual entry of author names in a database system might result in misspelling, the lack of automated verification in the recording process, the lack of domain specific knowledge of the persons responsible for recording the data, … might affect the occurrence of DQ inaccuracies.

Commonly used techniques to identify root causes include the auditing of the data, the surveying of the user perceptions and the evaluation of the data process. The identified causes can then be depicted in cause and effect diagrams, also termed Ishikawa or fishbone diagrams [24]. These diagrams cluster causes together in groups which is instrumental in identifying, classifying and prioritizing the impact of root causes to a problem. In our example root cause analysis could result in the identification of the field 'author name', as a string datatype, that is, completed according to the data provider's interpretation and accuracy. Because the datatype is set as a string, multiple inaccuracies can occur during the registration process.

5.4 DQ improvement trajectories

In the next phase, the focus resides on finding solutions to eliminate the root cause of the problem. These solutions, also termed remedies, are in fact changes to data systems or processes in order to prevent data inaccuracies from happening including the swift detection upon their occurrence. While some solutions might be oriented towards improving the data registration, others might focus on the implementation of validation rules or periodic data profiling. In addition, re-engineering of associated data processes and even training of the data provider and user community on data quality aspects, should be considered. Data cleansing might be applied as well, however this mostly is not a solution to eliminate the root cause itself.

Although solutions might be found using common sense, in most cases more efforts are needed. A frequently used method encompasses the organization of topic-oriented brainstorm sessions in the presence of all stakeholders. This approach has the benefit to tackle the problem for multiple viewpoints and at the same time enables a higher engagement of the stakeholders. Importantly, all relevant solutions to the problem should be listed and effects of the proposed solutions should be investigated carefully. In general, continuous, short-term improvements are to be preferred as these might result in quick wins which can result in additional business benefits (as DQ improvement is mostly not a goal in itself).

In our example many solutions can be found that focus on improving the correct registration of the author name. However, if an author ID would be registered and coupled to an author name, the specific focus on registering the name perfectly in a wide variety of bibliometric sources diminishes. Although this seems an easy solution at first glance, this strategy also includes the re-engineering of business processes, that is, the authentication of research publications by an author using its author ID. In order to investigate the effect of this proposed solution, one could investigate the number of publications that can be attributed to a group of authors that has registered and authenticated their research publications versus a group of authors that have no author ID (i.e., the control group) in an experimental setting. By measuring the DQ of both groups in terms of accuracy and completeness, one can see the effect of the proposed solution.

5.5 DQ control and follow-up

Based on all DQ solutions tested, the most appropriate solution(s) should be selected for implementation. It is important to note here that the success of implementation is dependent on the guidance foreseen to all stakeholders. In essence, this comes down to providing information on the solution and its effectuation on all (related) business processes to everybody involved. In addition, business rules, definitions, roles and responsibilities must be defined in consultation with all stakeholders.

Obviously, a close monitoring is needed in order to follow-up on the effectiveness of the implemented DQ solution in the real-world setting as a means to validate the (positive) impact of the proposed DQ solution. At the same time, it allows for the detection of unexpected errors that were unanticipated in the experimental test phase, and the swift adoption of corrective measure in case required. Specific monitoring tools that can be used here include control charts, also known as Shewhart charts, cause and effect diagrams, check sheets, histograms, Pareto charts, scatter diagrams, … [25].

With regards to the author disambiguation example described, it will be required to install business process that allow for the coupling of a unique author ID with corresponding research publications. This includes the close cooperation of the authors, research administrators, data analysts and data system/IT-staff on the definitions, business rules and responsibilities of each stakeholder. For instance, it might well be that authors are obliged to enter a unique author ID in a database system in fixed format, rather than a free text field. A business rule could be that for each author, an author ID of a given type (i.e., ORCID, Researcher ID, Scopus ID, Research Gate ID.) should be kept in a data system, which translates to a value of a given format, that is, an integer, in terms of a derived validation rule. This author ID field might be used to search large bibliometric databases such as Web of Science, Scopus, … for publications that might be coupled to this author ID, which could be added to the bibliometric profile of a researcher. Furthermore, publications might also be retrieved using an author name search that are not yet coupled to this author ID. Therefore, an authentication step is required here in which the author has a critical responsibility to validate these publications. Research administrators and data analysts should be informed on the process of authentication in order to use the information in a correct manner. Although this might seem a perfect solution, the reality demonstrates that a continuous follow-up is required as practice demonstrates that authors sometimes use several author IDs of the same type. Therefore, a corrective action could be to adapt the business rules in order to allow for only one author ID of a give type within the data system as well as the notification to the author to take corrective measures in this respect and the follow-up thereof.

It is clear from the example described above, that data quality improvement is a process that requires continuous monitoring due to internal and external factors that might effectuate data quality and its related processes. Therefore, the systematic and continuous retaking of the DQ improvement workflow will be the only manner to constantly have qualitative data instrumental for high quality data analyses.

6. Conclusion

Research organizations worldwide are using data on research input and output, that is, publications, patents, research data nowadays for a wide variety of use purposes, such as evaluation, reporting and visualization of a researcher' or research organization's expertise. This places high demands on the quality of the data gathered for these purposes, which have—in most cases—largely outgrown the initial intentions when the data systems were constructed. Moreover, the research world has evolved in a global, dynamic in which research data are increasingly being used in order to monitor the efficiency of research processes, the research productivity and even strategic decision making. In order to safeguard correct data analysis, research-related data must be assessed on all relevant quality dimensions, and inaccuracies must be addressed using data quality improvement trajectories as discussed in this chapter. The integration of a data quality management policy, is the only way to ensure the fitness for use of research-related data for various applications and business processes across the research world as the impact of inaccurate date can have tremendous effects on a researcher's or research organization's future prospects.

Acknowledgements

This work is carried out for the Expertise Centre for Research and Development Monitoring (ECOOM) in Flanders, which is supported by the Department of Economy, Science and Innovation, Flanders.

A. Abbreviations

BPMN	Business Process Model Notation
CASRAI	Consortia Advancing Standards in Research Administration Information
CERIF	Common European Research Information Format
CRIS	current research information systems
FRIS	Flanders Research Information Space
DQ	data quality
DQM	data quality management

Author details

Sadia Vancauwenbergh
ECOOM-Hasselt and Hasselt University, Hasselt, Belgium

*Address all correspondence to: vancauwenbergh@uhasselt.be

References

[1] Juran JM, Blanton Godfrey A. Juran's Quality Handbook. 7th ed. Europe: McGraw-Hill Education; 2016. p. 992. ISBN-10: 9781259643613

[2] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems. 1996;**12**(4):5-33. DOI: 10.1080/07421222.1996.11518099

[3] Moges H-T, Dejaeger K, Lemahieu W, Baesens B. A total data quality management for credit risk: New insights and challenges. International Journal of Information Quality. 2012;**3**(1):1-27. DOI: 10.1504/IJIQ.2012.050036

[4] Culnan M. The dimensions of accessibility to online information: Implications for implementing office information systems. ACM Transactions on Office Information Systems. 1984;**2**(2):141-150. DOI: 10.1145/521.523

[5] Halpin JF. Zero Defects: A New Dimension in Quality Assurance. New York City: McGraw-Hill; 1966. p. 228. OCLC 567983091

[6] Crosby PB. 8: Quality Improvement Program. Quality Is Free: The Art of Making Quality Certain. New York City: McGraw-Hill; 1979. pp. 127-139. ISBN 9780070145122. OCLC 3843884

AQ1 [7] Redman TC. Data Quality: The Field Guide. Digital Press; 2001. p. 256. ISBN-10 1555582516

[8] DAMA International. The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK). Technics Publications, LLC; 2009. p. 430. ISBN-10: 0977140083

[9] EWSolutions, Foundations of Enterprise Data Management [Internet]. 2013. Available from: https://www.ewsolutions.com/foundations-enterprise-data-management/ [Accessed: 18 April 2019]

[10] Oracle, Oracle Warehouse Builder Users Guide 10g Release 2(10.2.0.2) [Internet]. 2009. Available from: https://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223/concept_data_quality.htm [Accessed: 18 April 2019]

[11] Milosevic D, Patanakul P. Standardized project management may increase development projects success. International Journal of Project Management. 2005;**23**:181-192. DOI: 10.1016/j.ijproman.2004.11.002

[12] Baskarada S, Koronios A. A critical success factor framework for information quality management. Information Systems Management. 2014;**31**(4):276-295. DOI: 10.1080/10580530.2014.958023

[13] EuroCRIS, CERIF: Main features of CERIF. Available from: https://www.eurocris.org/cerif/main-features-cerif [Accessed: 18 April 2019]

[14] CASRAI: CASRAI dictionary. Available from: https://dictionary.casrai.org/Main_Page [Accessed: 19 April 2019]

[15] European Commission: Sensitive data. Available from: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en [Accessed: 19 April 2019]

[16] European Commission: Data protection in the EU. Available from: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en [Accessed: 19 April 2019]

[17] Vancauwenbergh S, De Leenheer P, Van Grootel G. On research information and classification governance in an

inter-organizational context: The Flanders Research Information Space. Scientometrics. 2016;**108**(1):425-439. DOI: 10.1007/s11192-016-1912-7

[18] Bell T, Logan D, Friedman T. Key issues for establishin information governance policies, processes and organization. Gartner Research; 2008

[19] Fowler M. Patterns of Enterprise Application Architecture. Addison Wesley; 2003. p. 116. ISBN-10: 0321127420

[20] White SA. Process Modeling Notations and Workflow Patterns. IBM Corporation; 2006

[21] Saltelli A. Sensitivity analysis for importance assessment. Risk Analysis. 2002;**22**(3):1-12. DOI: 10.1111/0272-4332.00040

[22] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global Sensitivity Analysis: The Primer. John Wiley & Sons; 2008. ISBN-10: 0470059974

AQ2 [23] Woodall P, Oberhofer M, Borek A. A classification of data quality assessment and improvement methods. International Journal of Information Quality. 2014;**3**(4). DOI: 10.1504/ijiq.2014.068656

AQ3 [24] Ishikawa K. Guide to quality control. Asian Productivity Organization. 1976. ISBN: 92-833-1036-5

[25] Tague NR. The Quality Toolbox. Milwaukee, Wisconsin: American Society for Quality; 2005. p. 15. ISBN-10: 0873896394