International Journal of Economics, Commerce and Management
United Kingdom ISSN 2348 0386 Vol. IX, Issue 4, April 2021

http://ijecm.co.uk/

A THEORETICAL FRAMEWORK FOR A DATA COLLECTION PROCESS

Blerina Metanj (Subashi)
IDRA Research and Consulting, Albania
subashiblerina@yahoo.com

Abstract

This paper aims to provide a general theoretical framework for a fruitful data collection process, in which quality is considered during all the phases of its implementation. It is well known that data collection for the sake of producing analysis within a certain field of interest, includes a huge preparatory work, efforts, and many financial and human resources. During the preparatory work, every person involved should maintain the focus that the main objective is to provide high-quality data to the users for supporting them in evidence-based decision making. This is the reason why to take into account and not underestimate the quality dimensions since the beginning of the implementation of data collecting activities. The analysis is done in two directions. On one side a strategy that evidences specific activities to address quality dimensions will be proposed together with some specific indicators to measure their impact. On the other hand, the identification of some main phases during the data collection process will be done in line with a direct link on what quality dimension these specific phases could improve. This framework could be extended to every data collection process, quantitative or qualitative, which aim to have in focus the quality.

Keywords: Data collection, data dissemination, quality, management plan, quality indicators

INTRODUCTION

In general, there are three main identified data sources for users. They include administration data from institutions, data from surveys, and Census data. Census data are unique since they could provide information for a very specific area, and fill the gaps where other data sources might lack. Unfortunately, Census is a very expensive activity and does not provide periodic information since it is carried out every five or ten years.

On the other hand, sometimes administrative data do not provide the needed level of disaggregation and are not very easy and user-friendly to access and use by users.

In this perspective, a lot of institutions and agencies often implement their data collection with specific surveys in line with their information needs. Sometimes they may lack capacities and might hire other institutions more specialized to do this task. The final information provided is crucial to them since affects directly their project objectives. Due to the importance of such a process, the size of operation, complexity, and the number of people involved in a data collection process might be subject to many sources of errors. Therefore, a comprehensive system of quality assurance has to be designed and implemented together with the team involved. Consequently, a management plan that takes care of quality is needed to provide an overview of the methodologies and standards that will be adopted in managing the whole project and in the production of the outputs.

The definition and management of quality of data were discussed in several papers presented at the 1995 International Conference on Survey Measurement and Process Quality (Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin 1997, de Leeuw and Collins 1997, Dippo 1997, Morganstein and Marker 1997, Colledge and March 1997) and more recently in Collins and Sykes (1999). From a data analyst or a statistician's point of view, they might understand the calculation of different indicators such as variance, standard errors, bias, hypothesis testing, all measures of quality of a specific indicator they are measuring. Bracktone (1999) points out that the quality concept has been overused to some extend and questioned because of its vagueness.

Quality in a statistical context usually has been referred to the accuracy of the statistical product. This may be applicable sometimes, but quality encompasses a wider set of attributes such as relevance, accuracy, timeliness, accessibility, etc. In terms of an institution, such as research institutions, the terms quality is broader and it is part of the management and philosophy of the whole business. Under these terms, the extended view of quality is “the totality of features and characteristics of a product or service that bear on its ability to satisfy a given need” (ISO 8402 from 1986).

METHODOLOGY

Under the purpose of this paper, quality will be defined in general in terms of eight dimensions, which include Relevance, Accuracy, Timeliness/ Punctuality, Accessibility, Comparability, Coherence, Completeness, Transparency. The proposed framework will identify some detailed activities that address and impact a specific dimension of quality. These activities should be part of a management plan for the data collection process. It is quite obvious that the dimensions conflict with each other, as discussed by Holt and Jones (1998). For instance, timeliness conflicts with accuracy since good accuracy generally takes time to achieve. Consequently, the various dimensions cannot be treated as if they were independent. Also, the aspect of confidentiality will be considered due to its high importance in data collection. Moreover, in each quality dimension, a set of indicators are proposed to be taken into account as a way of control and monitoring the whole process.

Achieving an acceptable level of quality is the result of addressing, managing, and balancing over time the various dimensions of quality, with the use of the information, costs, respondent burden, and other factors that may affect data quality or user expectations. Actions taken to address one dimension of quality may affect other dimensions, often in ways that cannot be fully predicted. Decisions and actions aimed at achieving an appropriate balance of quality dimensions and other factors are based on knowledge, experience, reviews, feedback, consultation, and, inevitably, the judgment of the experts.

On the other hand, the management plan of data collection is composed of a set of logical steps. During the implementation of each step, the quality dimensions that this specific stage should take into account will be identified.

PHASES OF IMPLEMENTATION OF A QUALITATIVE DATA COLLECTION

The implementation of a successful data collection management plan that takes into consideration quality dimensions, begins with the identification of user needs. The user-producer dialogue should be regular and ongoing, and the information collected by such dialogues should be reflected in management activities of the data collection process. The objective is to meet the major needs of the users, and the acceptance criteria will be the approval of the instrument of data collection by the group of stakeholders.

Since in the beginning the key users of data should be identified. Then periodic consultations with them should be organized to clearly understand their data needs. It is recommended to conduct desk research to identify if new requirements have been developed in line with the kind of data that is being collected. During all the time it is recommended to uphold continuous meetings with stakeholders. It would be a good strategy to develop a user satisfaction index and monitor it over time.

Undertaking the following activities will ensure that the relevant quality dimension will be meet. The *relevance* reflects the degree that the current and potential needs of users are met. It is concerned whether the available information is in line with the most important issues to users. The challenge is to balance the conflicting between the most important needs within given resource constraints.

The other element of quality which should be taken into account during the implementation of a data collection management plan is *accuracy.* It refers to the degree to which the phenomena it was designed to measure are correctly described and measured. The accuracy of statistical estimates is usually quantified by the evaluation of different sources of error, where the magnitude of an error represents the degree of difference between the estimate and the true value. Different indicators might be calculated under this dimension such as coefficient of variations, response and non-response rate, over coverage rate, etc. It is important that the management plan clearly shows all the indicators that will be used and how they will be used.

Managing accuracy means investing in solid methods of research programs to increases the capacities. More specifically addressing the accuracy dimension under the proposed framework involves consultation with the experts about the data collection instrument design to ensure correct conceptualization of the data to be collected. The development of manuals will improve the common understanding of information that will be collected and adequately respond to data-collection challenges. The development and use of detailed and advanced training materials for field staff (data collectors) in order to improve the understanding of information to be collected and to adequately respond to data-collection challenges. Conducting thorough testing of questionnaires, forms, and procedures in a field cognitive testing and pilot the procedures. Taking notes about observation and checking of field staff will improve if any problem with the understanding of the data collection instrument. A development application for monitoring and managing operation fieldwork will help in addressing the problem in real-time. The development and application of a large set of editing and imputation rules to identify and solve inconsistencies during the data analysis phase with a strong collaboration with the IT staff and the expert in the field of interest.

Of course, the users of the data might have different requirements about timeliness. The development of technology has contributed to higher expectations regarding time. Sometimes *timeliness* is often based on trade-offs with accuracy – and cost. More specifically, we can refer to *timeliness* as the speed of dissemination of outputs. The proposed steps to take into account the timeliness of the dissemination of data results include a Well agreed dissemination plan of final results. Training of a pool of reserve field staff to avoid delays in enumeration due to drop-out. The use of modern technology to collect data in electronic format will ensure the shrink of time during the data collection. The application of an effective information system for monitoring data collection progress in the field in addressing the problem in real-time. Making available a reasonable calendar of data release to the users in advance.

Once data are producing the users should be given access to the information as part of the whole process. Data should be easily *accessible* to users and presented in a clear and understandable format, and accompanied by relevant supporting metadata. Metadata accompanying data dissemination improves the clarity and interpretability of the results. The indicators proposed to monitor accessibility may be proposed the use of media used for the dissemination of data and the way these media are used to access information. Having a good monitoring tool for how the information is accessed and downloaded by users is very useful.

The strategy related to accessibility and clarity of the outputs includes the documentation of the concepts and definitions, the classification that has been used during the data collection. Moreover, documentation of the procedures during the data collection phases is necessary to increase accessibility and clarity. Produce corresponding metadata for the data file presented in a format that facilitates proper interpretation, meaningful comparisons for the users.

The concepts of comparability refer to the measurement of the impact of differences in applied concepts, measurement instruments, and procedures, where statistics are compared between geographical regions, sectorial domain fields, or periods. One way to monitor the comparability is to use the times series, which means the use of previous data, or other countries that implement it preferably in the same time frame.

The proposed activities concerning the comparability include taking care of national or international recommendations on definitions, classifications about the data being collected. Maintaining overall comparability with the previous data collection procedures within the same project will contribute to better monitoring the data comparability. Considering other data sources and maintain where possible comparability with international context will be an added value to the whole process.

The coherence includes coherence between different data items about the same point in time, coherence between the same data items for different points in time, and international coherence. The confrontation of data from different sources, and their subsequent reconciliation or explanation of differences, is an activity that is often needed as part of pre-release review or certification of data to be published. All this process per se may include the production of a set of indicators that may have come up during the confrontation phase. Feedback from external users and analysts of data that point out coherence problems with current data is also an important component of coherence analysis. Some incoherence issues only become apparent with time and may lead to data revisions. The use of standard concepts, definitions, and classifications promotes consistency. Taking into consideration other sources in case of the same type of information is being collected would increase the output coherence.

Completeness refers to the degree to which data cope with the needs of data users as completely as possible, taking limited resources into account. During this phase is very important to take into serious consideration *confidentiality* and its protection through the implementation of statistical disclosure control methods, to preserve the re-identification of the personal and private information given to the respondents.

The completeness could be addressed through the organization of users' consultations to assess their needs. Giving the users the possibility to develop their customized self-tabulation, or having direct access to microdata are activities that promote completeness. And as mention, all this should be carried out in parallel with the development of clear steps on implementing statistical disclosure control techniques.

Transparency is the right of respondents to have information on a legal basis, the purpose for which the data is required, and the protective measures adopted. One indicator for monitoring the transparency might include the amount of data being released. The objective is to make accessible all the documentation of the process to the users and be transparent on the conduct of all its phases.

The management plan might consider the use of new technologies and innovative methods to share knowledge between producers and users of data to promote transparency. Once metadata are produced, they should be made available and accompany disseminated data. The documentation of all phases of the data collection implementation plan and make it available to the users will increase the transparency process and the reliability.

Table 1 identifies all the above activities as part of the whole management plan of a data collection process.

Table 1. Activities within the framework of the management plan

List of activities	Monitoring Indicators	Quality
• Identify the key users • Consult key users that will use the produced data and statistics • Conduct desk research to identify new requirements that have been developed in line with the data being collected • Uphold continuous meeting with stakeholder	Develop a user satisfaction index and monitor it over time	Relevance
• Consultations with experts on the data collection instrument design to ensure correct conceptualization of the data to be collected; • Development of manuals to improve common understanding of information that will be collected and to adequately respond to data-collection challenges; • Development and use of detailed and advanced training materials for field staff (data collectors) as to improve understanding of information to be collected and to adequately respond to data-collection challenges; • Thorough testing of questionnaires, forms, and procedures in a field cognitive testing and pilot;	• Coefficient of variations of the variables • Response and non-response rate • Over coverage rate	Accuracy

• Observation and checking of field staff; • A development application for monitoring and managing operation fieldwork; • Development and application of a large set of editing and imputation rules to identify and solve inconsistencies in the database;		
• Well agreed on timeless dissemination of final results; • Training of a pool of reserve field staff to avoid delays in enumeration due to drop-out; • Use of modern technology to collect data in electronic format • Application of effective information system for monitoring data collection progress in the field; • Publish a reasonable calendar of data release;	The potential difference between the real and planned data delivery	Timeliness
• Document the concepts and definitions, the classification that has been used during the data collection • Document the procedures during the data collection • Produce corresponding metadata for the data file presented in a format that facilitates proper interpretation, meaningful comparisons	• What kind of media is being used for the dissemination of data • Is this media being used to access information by users? To what extent?	Accessibility
• Compliance with national or international recommendations on definitions, classifications, and procedures; • Maintain overall comparability with the previous data collection procedures within the same project; • Maintaining comparability in the conceptualization with other data sources; • Maintain comparability in an international context where possible;	Differences between time series	Comparability
• Mention other sources in case of identification that the same type of information is being collected	Feedback from external users	Coherence
• Organize users' consultations to assess their needs • Develop the possibility for users to develop their customized self-tabulation • Offer access to microdata files • Develop clear steps on implementing statistical disclosure control techniques	User need	Completeness
• Use new technologies and innovative methods to share knowledge between producers and users of data; • Make available all relevant metadata to accompany disseminated data • Document all phases of the data collection implementation plan and make it available;	Amount of data being release	Transparency

After identifying the quality dimension and addressing each of them with specific tasks carried within the data collection process, the main phases that the data collection process are identified. In this analysis, the main phases of the implementation are considered the development of the instrument for the data collection, the preparation of the staff that will work in the field, the delivery of the materials, provision of the training, data analysis, and at the end the data dissemination.

Since the analysis has identified the activities that improve the specific quality dimensions, it is possible to link the phases of the data collection and the quality dimension that they should take into account. This is possible because every activity is part of a/some specific phases of the data collection.

A real situation might evidence more activities or some activities might be considered as sub-activity. Moreover, it is clear that in every step different human and financial resources will be engaged. In this perspective, the link between these steps and the quality dimension makes it more evident that the whole work will contribute to the final goal of providing a high-quality process concerning all the quality dimensions.

Table 2. The link between the phases of data collection and quality dimension

The phase of data collection	Quality dimension to address	
Design of the data collection instrument	Relevance Comparability Completeness	Accuracy Coherence Transparency
Planification of human resources	Accuracy Timeliness Transparency	
Preparation of the fieldwork (logistic)	Timeliness Punctuality	
Shipment and collection of material	Accuracy Timeliness Punctuality	
Organization of the training for the field staff	Accuracy Timeliness Punctuality	
Filed work staff training	Accuracy Timeliness	
Data analysis	Accuracy Comparability Coherence Transparency	Timeliness Confidentiality Completeness
Dissemination of final data	Relevance Punctuality Accessibility	Timeliness Completeness

CONCLUSIONS

A real study can provide high-quality data if the quality is considered since the beginning of the implementation of the data collection process. This requires firstly deciding upon what dimensions of quality will be taken into account during a specific data collection process. Afterward, for each of the agreed dimensions, a set of activities within the management plan should be identified. Moreover, this paper identifies some potential indicators that could be used during the monitoring of addressing each of the dimensions and to take the necessary decisions in case of deviations. Finally, the main phases of the data collection implementation are identified with the specific quality dimensions to be taken into consideration during their implementation. The set of proposed activities to address the quality dimension could be considered as a good start for each data collection procedure. And at the same time, other activities might be identified. The important thing is that each activity should be thought on how it will contribute to the overall quality of the process.

REFERENCES

Bradburn, N. M. “Respondent Burden”, Paper presented at the 138th Annual Meetings of the American Statistical Association, San Diego, CA., 1978

ESTAT/02/Quality/2005/9/Quality Indicators,

Eurostat 2014, ESS handbook for quality reports, Retrieved on 15 March 2021 from https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Fannie Cobben, Statistics Netherlands, Nonresponse in Sample Surveys, 2009

L. Lyberg et al. – Survey measurement and Survey Quality, Wiley, 1997

Lyberg et al., Wiley. In Survey measurement and process quality (p. 415-435)

Methodological documents, Definition of Quality in Statistics: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003

Methodological documents, Handbook “How to make a quality report: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003

Methodological documents, Standard Report: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003

Standard Quality Indicators - Producer oriented: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003