DATA QUALITY ASSESSMENT METHODOLOGY: A FRAMEWORK

Eugene M. Burns, Purificacion O. MacDonald, and Amrut Champaneri
Bureau of Transportation Statistics, U.S. Dept. of Transportation, Washington, DC 20590

KEYWORDS: Data Quality; Quality Assessment; Quality Criteria; Metadata; Statistical Metadata; Quality Measurement; Performance Measures

Introduction

Ten years ago, Congress established the Bureau of Transportation Statistics (BTS) as the statistical agency within the U.S. Department of Transportation (DOT). Although BTS was not given responsibility for collecting and disseminating statistical data for the rest of DOT, BTS was designated as the lead agency within DOT for promoting data quality. The data quality programs include (1) providing guidelines for data quality and (2) reviewing the quality of DOT data systems. This paper focuses on the latter program. Data quality reviews focus on data systems that are used to measure DOT performance under Government Performance Results Act (GPRA) measure or that are related to DOT's strategic goals. The DOT strategic goals, Safety, Mobility, Economic Growth, Human and Natural Environment, and National Security, encompass a wide range of statistical data systems.

BTS has developed a data quality framework to guide the data quality review portion of BTS's mandate. The data quality review program consists of comprehensive assessments of DOT data systems, accompanied by recommendations and suggestions for data quality improvements. The program is intended to be collaborative, and to build working relationships with the other DOT operating administrations (e.g., the Federal Aviation Administration, the Federal Highway Administration).

The data quality review framework is built around three linked components:

- Quality criteria,
- Data system metadata, and
- Rating questions.

Quality criteria define the desirable characteristics for data systems to exhibit. Metadata are the information about the existing data and data processes that are needed to review the data systems. Rating questions (being developed) specify how the metadata (and the data) relate to the quality criteria.

The three components of the data quality assessment framework will be discussed in the following three sections of this paper. The final section discusses some possible extensions of the framework

What is Data Quality?

The answers to the question "What is data quality?" define the scope of the data quality assessments, as well as provide the criteria for assessing data systems. Different analysts and different agencies provide different answers (Brackstone 1999, Carson 2000, Pipino *et al*. 2002), but all agree that "data quality" is a multidimensional concept. The dimensions specified by statistical agencies tend to cover largely what Huang *et al.* (2002) would term "external" criteria. These are criteria defined from the perspective of the information user rather than from the perspective of the information system.

The BTS strategic goals (BTS 2000) exemplify a set of quality dimensions. Figure 1 compares Brackstone's (1999) data quality dimensions (adopted by Statistics Canada) with the BTS strategic goals. Both agencies define data quality as possessing six dimensions. Brackstone's Relevance dimension corresponds to two dimensions, Relevance and Completeness, in the BTS strategic goals. BTS's Utility dimension includes both the Brackstone Accessibility and Interpretability dimensions. Finally, while the BTS strategic goal of Quality may seem a strange dimension *of* quality, BTS defines it in terms of accuracy, reliability, and objectivity. Thus the BTS Quality goal corresponds to Brackstone's Accuracy dimension.

The views and opinions expressed in this paper are the those of the authors and are not necessarily those of the Bureau of Transportation Statistics.

Quality Dimensions (Brackstone 1999)	Strategic Goals (BTS 2000)
Relevance	Relevance
Accuracy	Quality
Timeliness	Timeliness
Accessibility	Utility
Interpretability	Completeness
Coherence	Comparability

Figure 1. Quality Dimensions (Brackstone 1999) Compared With the BTS Strategic Goals (BTS 2000)

Both the Brackstone and the BTS schemes would imply similar issues as being in-scope for a data quality review. Given that the quality dimensions seem to be describing a similar concept, the choice between competing schemes is somewhat arbitrary. However, having an organization's strategic goals as quality criteria lends the program more weight and, potentially, more organizational support. For the data quality assessment project, the strategic goals of BTS, as the DOT statistical agency, have been taken as the quality criteria applicable to all DOT statistical data.

It should be noted that a program of data quality reviews might not be the best tool for investigating all of these dimensions of data quality. A data quality review focuses most heavily in the Accuracy/Quality area, but has limited ability to address an area such as Completeness (having data in every area of interest). The assessment of Completeness requires a global view of both (1) an agency's suite of data collection programs and (2) the data needs of its customers. BTS has completed such an assessment (BTS 2002b) under a different project.

Metadata and the Data Quality Assessment Template

The data quality assessment template specifies the information that each reviewer should assemble for the target data system. The template organizes the statistical metadata and other information, by data system and assessment process. It also serves as an outline for the assessment report.

Figure 2 lists the topics covered in the data quality assessment template. Sections A through F, the statistical metadata, consist of information about the purpose and history of the data system and the processes involved in a data collection activity: sampling, data collection, data preparation, data dissemination, and evaluation. Section G contains the results of the data reviewer's analysis of the data, from ease of obtaining data to presence of outliers to relationship with other data systems. Sections H and I provide feedback to the data system sponsors through a quality assessment and recommendations for improving the data quality. Finally, besides its use as a starting point for follow-on assessments, the bibliography (Section J) can be useful for analysts in general.

Section	Topic Covered
A	Background
B	Frames and Sampling (if applicable)
C	Data Collection
D	Data Preparation
E	Data Dissemination
F	Sponsor Self-Evaluation
G	Data Analysis Results
H	Assessment
I	Recommendations and Suggestions
J	References

Figure 2. Contents of the Data Quality Assessment Template (Draft, 9/24/2001)

A. Background
1. Name of data system:
2. Sponsoring agency:
3. Legal authority: *Legislation, regulations*
4. When initiated:
5. Original purpose of data system:
6. Target population: *Events/objects/businesses/persons/etc. of interest, and rationale for choosing*
7. History of data system: *Significant changes in purpose, data uses, collection strategies, etc.*
8. Future plans: *Have any? How formulate?*

Figure 3. Section A of the BTS Data Quality Assessment Template (Draft, 9/24/2001)

Figure 3 is a sample section from the data quality assessment template. The analysts performing the assessments can enter the information as they find it and then use the template as the basis for a structured interview with the data system sponsor to capture the remaining information.

Section H of the template lists the BTS strategic goals (used as a quality criteria) and cross-references the goals to previous sections of the template. For example, for assessing timeliness, Section H directs the reviewer to consult information in Sections A, C, E, and F.

However, as the first few assessments were being conducted, it became apparent that the guidance in Section H was not sufficient. The problem was that the knowledge of how to use the metadata elements to assess the data systems was left to the reviewer's judgment, leading to excessively subjective evaluations. Particularly since the data systems reviewed are not necessarily BTS data collection systems, it was important to be able to show data system sponsors how the evaluations were made.

Rating Questions

The rating questions, organized by BTS strategic goal, provide explicit links between the metadata elements and the evaluation criteria. They are designed to (1) ensure that the assessments are consistent by providing guidance to reviewers and (2) make the data quality assessment transparent to the data system sponsors. Ultimately, DOT data system owners may be able to use the rating questions for self-assessment.

Figure 4 contains the rating questions for the Utility quality criterion. Utility is divided into its subcomponents, and each question cross-references the relevant metadata items in the template.

Different dimensions (strategic goals) have different numbers of rating questions, varying by the degree to which the data quality assessments target that dimension. For example the Quality dimension, the dimension most targeted by the assessments, is addressed in four subcomponents: coverage and sampling, data collection, data preparation, and data dissemination. Appropriately for data quality reviews, the four Quality subcomponents currently contain about half of all rating questions.

Ease of Access

- Are the data readily accessible to meet the needs of the data users? [A5, A7, A8, E1, E2, E3]
- How easy is it to access the data? [G1]
- How has user feedback been used to improve data accessibility? [A8]
- How clear and usable are reports and other publications of the data? [D7, D8 (?), E4, E5, E6, E7, F4, F5]

Ease of Understanding

- How complete is the documentation? [B5, C11, D9, F3, G2]
- Is the documentation for data element definitions, the data collection procedures and other meta-information adequate? [G2, C11, B5, D9]
- Do the statistical tables include presence of errors (e.g. standard errors), frequency of missing data, number of units on which percentages or rates are based? [E4, E6]
- Are the statistical tables that are released accompanied by explanations of sources and accuracy statements and data limitations? [E6, F3]

Ease of Use

- Were you able to reproduce published estimates? [G9]
- Are there problems in the file that make it difficult to work with? [D4, D5, D6, G3, G3A, G4, G5, G6, G7, G8, G11]

Figure 4. Rating Questions (Draft, 12/14/2001) for the Strategic Goal of Utility

Once developed, a satisfactory set of rating questions could become the basis for a measure of data quality performance. The literature on data quality measurement is still evolving. Wang et al. (1999) include an Information Quality Assessment (IQA) survey instrument, along with rating scores, developed after a decade of research on information and data quality at MIT. In the transportation area, Elvik (2002) presents ideas for developing a quality scoring system for road safety evaluation studies, tailored to the needs of researchers who conduct or use the results of such studies. He describes the characteristics of an ideal formal quality scoring system and includes a list of questions any formal quality scoring system should include.

BTS is just beginning to develop the BTS data quality rating questions, and is currently sponsoring a peer review of the data quality assessment framework methodology. The pioneering application of the data quality assessment framework at BTS will contribute to this field of research.

Results and Related Work

Results from the first two years of data quality assessments indicate that BTS and DOT data systems are weakest with respect to the Utility dimension. For many DOT data systems, data system documentation may be prepared by computer-oriented staff members. These staff members, naturally, are more focused on processing steps and data file characteristics than on the stages of planning and data collection that preceded initial data entry. Response rates and coverage may not be mentioned, and data quality self-assessments may be hard to find, making it difficult for potential users to decide whether the data meet their needs. The need to assemble or create metadata also hinders the progress of the data quality assessments.

While the data quality assessment project reviews one data system at a time, the statistical guidelines project provides guidance for all DOT systems at once. A recent Office of Management and Budget directive (OMB 2002) has helped to reinforce the statistical guideline, and data quality review, message throughout DOT. The BTS *Guide to Good Statistical Practice* (BTS 2002a) has been revised to stress the need for transparency and sound statistical methods. The data quality assessment and statistical guidelines projects are being closely coordinated to reinforce each other.

OMB has directed that agencies adopt information quality as a performance goal. Since performance should be measurable, measures of quality will be needed to track performance. With additional development and testing to ensure validity and reliability, the rating questions could form the basis for data quality metrics along the quality dimensions, and for the derivation of a data quality index. More ambitiously, it might be possible to summarize the measures across multiple DOT data systems to form a measure of DOT data quality.

As one of the DOT performance measures, data quality would remain prominent in upper management's attention. If this attention can be translated into additional funding for data quality improvements, the data quality framework can have a lasting impact.

References

Brackstone, G. (1999), "Managing Data Quality in a Statistical Agency," *Survey Methodology*, 25, 139-149.

Bureau of Transportation Statistics (2000), *Strategic Plan 2000-2005*, available at http://www.bts.gov/StrategicPlan.pdf.

_____ (2002a), *BTS Guide to Good Statistical Practice*, available at http://www.bts.gov/statpol/guide/index.html.

_____ (2002b), *Data Gaps: A Vision for Transportation Data*, forthcoming at http://www.bts.gov/datagaps.

Carson, C. S. (2000), "What is Data Quality? A Distillation of Experience," Statistics Department, International Monetary Fund.

Elvik, R. (2002), "Measuring the Quality of Road Safety Evaluation Studies: Mission Impossible?," presented at the 81st Annual Meeting of the Transportation Research Board, Session 539.

Huang, K.-T., Lee, Y. W., and Wang, R. Y. (1999), *Quality Information and Knowledge*, Upper Saddle River, NJ: Prentice Hall.

Office of Management and Budget (2002), "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies" *Federal Register*, February 22, 67 FR 8452.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002), "Data Quality Assessment," Communications of the ACM, 45, 211-218.

Ward, Y., and Wang, R. Y. (1996), "Anchoring Data Quality Dimensions in Ontological Foundations," Communications of the ACM, 39, 86-95.