Traffic Data Quality Workshop

Work Order Number BAT-02-006

DEFINING AND MEASURING TRAFFIC DATA QUALITY

White Paper

Prepared for

Office of Policy
Federal Highway Administration
Washington, DC

Prepared by

Battelle
. . . Putting Technology To Work

Texas Transportation Institute

Cambridge Systematics, Inc.

December 31, 2002

"Defining and Measuring Traffic Data Quality"

By Shawn Turner

Executive Summary

In developing this white paper, we reviewed current and advanced practices for addressing data quality in three types of user communities: 1) real-time traffic data collection and dissemination; 2) historical traffic data collection and monitoring; and 3) other industries such as data warehousing, management information systems, and geospatial data sharing. The recommendations in this paper follow from this review.

The recommended definition for traffic data quality is as follows:

> *Data quality is the fitness of data for all purposes that require it. Measuring data quality requires an understanding of all intended purposes for that data.*

The following data quality measures are recommended:

- *Accuracy*
- *Completeness*
- *Validity*
- *Timeliness*
- *Coverage*
- *Accessibility*

There are several other data quality measures that could be appropriate for specific traffic data applications. The six measures presented above, though, are fundamental measures that should be universally considered for measuring data quality in traffic data applications.

At this time, we recommend that goals or target values for these traffic data quality measures be established at the jurisdictional or program level based on a better and more clear understanding of all intended uses of traffic data. It is clear that data consumers' needs and expectations, as well as available resources, vary significantly by implementation program, urban area, or state and preclude the recommendation of a universal goal or standard for these traffic data quality measures.

We also recommend that if data quality is measured, a data quality report be included in metadata that is made available with the actual dataset. The practice of requiring a data quality report using standardized reporting is common in the GIS and other data communities. In fact, several metadata standards already exist (FGDC-STD-001-1998 and ISO DIS 19115) for standardized reporting of data quality in datasets. Until a formal traffic data archive metadata standard is approved, the traffic data community should create metadata based upon the core elements (i.e., mandatory metadata items) required in these two other geospatial metadata standards.

Introduction

Although not specifically referring to intelligent transportation systems (ITS), a Wall Street Journal article speaks to the subject of data quality: "Thanks to computers, huge databases brimming with information are at our fingertips, just waiting to be tapped. . . . Just one problem: Those huge databases may be full of junk." (Wand and Wang 1996) As Alan Pisarski noted in his Transportation Research Board (TRB) Distinguished Lecture in 1999, "we are more and more capable of rapidly transferring and effectively manipulating less and less accurate information" (Pisarski 1999).

Recent research and analyses have identified several issues regarding the quality of traffic data available from intelligent transportation systems for transportation operations, planning, or other functions. The Federal Highway Administration (FHWA) is developing an action plan to assist stakeholders in addressing traffic data quality issues. Regional stakeholder workshops and white papers will serve as the basis for this action plan.

As one of those white papers, this document presents recommendations for defining and measuring traffic data quality. This white paper:

- Reviews current data quality measurement practices in traffic data collection and monitoring;
- Introduces data quality approaches and measures from other disciplines; and
- Recommends approaches to define and measure traffic data quality.

Defining Data Quality

Several terms should be defined at the outset. Data and information are sometimes used interchangeably. Data typically refers to information in its earliest stages of collection and processing, and information refers to a product likely to be used by a consumer or stakeholder in making a decision. For example, traffic volume and speed data may be collected from roadway-based sensors every 20 seconds. This traffic data is then processed into information for the end consumer, such as travel time reports provided via the Internet or radio. But the terms are also relative, as one person's data could be another person's information. Throughout this paper the term data quality will be used to refer to both data and information quality. No attempt is made to delineate the point at which data becomes information (or knowledge or wisdom, for that matter).

The literature contains two similar definitions for data quality. Strong, Lee and Wang (1997A) define information quality as "fit for use by an information consumer" and indicate that this is a widely adopted criterion for data quality. English (1999A) further clarifies this widely adopted definition by suggesting that information quality is "fitness for **all** purposes in the enterprise processes that require it." English emphasizes that it is the "phenomenon of fitness for 'my' purpose that is the curse of every enterprise-wide data warehouse project and every data conversion project." In his book, English (1999B) defines information quality as "consistently meeting knowledge worker and end-customer expectations." It is clear from these definitions that data quality is a relative concept that could have different meaning(s) to different consumers. For example, data considered to have acceptable quality by one consumer may be of unacceptable quality to another consumer with more stringent use requirements. Thus it is important to consider and understand **all intended uses** of data before attempting to measure or prescribe data quality levels.

The recommended definition for traffic data quality is as follows:

> *Data quality is the fitness of data for all purposes that require it. Measuring data quality requires an understanding of all intended purposes for that data.*

Current Practices in Measuring Traffic Data Quality

Current practices in measuring traffic data quality are summarized below for three common consumer groups involved in highway transportation:

- Real-time traffic monitoring and control (e.g., traffic management centers);
- Operations/ITS data archives (traveler information systems, data archives, universities, etc.); and
- Historical/planning-level traffic monitoring (traffic monitoring groups in state and local DOTs).

Our review of current practice found that, in general, consistent and widespread reporting of traffic data quality measures was not evident in any of these three consumer groups. Efforts to address data quality were more evident in the latter two groups than with real-time monitoring and control. A few data quality measures have been suggested or are used in each of these groups. These data quality measures are discussed in the following paragraphs:

Real-Time Traffic Monitoring and Control

Data consumers in this group are typically engaged in traffic management and control or the provision of traveler information. Data uses are considered real-time and are generally concerned only with the most recent data available (e.g., typically five to fifteen minutes old). Some agencies are beginning to use historical data to provide additional value to traveler information. In some cases field data collection hardware and software provide rudimentary data quality checks; in other cases, no data quality checks are made from the field to the application database. Field hardware and software failures are common. In some cases, equipment redundancy provides sufficient information to cover gaps in missing data. In other cases, missing data is simply reported "as is" and decisions are made without this data.

Many agencies provide time-stamped traveler information via websites, thus providing an indication of the **data timeliness**. Selected examples can be found at Houston TranStar (http://traffic.tamu.edu), Washington State DOT (http://www.wsdot.wa.gov/PugetSoundTraffic/), and Wisconsin DOT (http://www.dot.wisconsin.gov/travel/milwaukee/index.htm), just to name a few.

Several traffic management centers track failed field equipment through maintenance databases and report such things as the average percent of failed sensors. The Michigan Intelligent Transportation Systems (MITS) Center has defined "lane operability" as the sensor-minutes of failure, which is a product of the number of failed sensors and the duration of the failure in minutes (Turner et al. 1999). These measures can be classified as measures of **coverage** or **completeness**.

Some traffic management centers evaluate the **accuracy** of new types of sensors before widespread deployment. For example, the Arizona DOT traffic operations center in Phoenix used accuracy to measure the data quality from non-intrusive sensors for which they were considering installation (Jonas 2001). In their evaluation, ADOT compared traffic count and speed data from non-intrusive, passive acoustic detectors to calibrated inductance loop detectors under the assumption that the loop detector data represented the most error-free data obtainable. The measures used in the evaluation were absolute and percentage differences between traffic counts and speeds measured with the two types of sensors.

ITS America and the U.S. DOT convened numerous stakeholders in 1999 and developed guidelines for quality advanced traveler information system (ATIS) data (ITS America 2000). The guidelines were developed in an effort to support the expansion of traveler information products and services. One of the explicit purposes of the guidelines was to increase the quality of traffic data being collected. The ITS America guidelines recommended seven data attributes, six of which can be considered data quality measures:

- **Accuracy** – how closely does the data collected match actual conditions?
- **Confidence** – Is the data trustworthy?
- **Delay** – How quickly is the data collected available for use in ATIS applications?
- **Availability** – How much of the data designed to be collected is made available?
- **Breadth of Coverage** – Over what roadways or portions of roadways are data being collected?
- **Depth of Coverage (Density):** How close together/far apart are the traffic sensors?

The ITS America guidelines further defined quality levels of "good", "better", and "best" and provided specific quality level criteria for each attribute. For example, five to ten percent error in travel times and speeds was classified as a "better" quality level under the *Accuracy* attribute.

In another white paper about data quality requirements for the INFOstructure (i.e., a national network of traffic information and other sensors), Tarnoff (2002) suggests the following data quality measures and possible requirements (Table 1):

Table 1. Possible INFOstructure Performance Requirements

Measure	Application	Requirement	
		Local Implementation	National Implementation
Speed Accuracy	Traffic Management	5-10%	5-10%
	Traveler Information	20%	20%
Volume Accuracy	Traffic Management	10%	N/a
	Traveler Information	N/a	N/a
Timeliness	All	Delay < 1 minute	Delay < 5 minutes
Availability	All	99.9% (approx. 10 hours per year)	99% (approx. 100 hours per year)

Source: Tarnoff 2002

Tarnoff presented these data quality requirements as a "starting point for the discussion of these issues" and suggested that there is a tendency in the ITS community to specify performance without a complete understanding of the actual application requirements or cost implications. Thus Tarnoff suggests that any decisions about data quality requirements be grounded in actual application requirements and cost implications.

Operations/ITS Data Archives

Data consumers in this group are typically engaged in off-line analytical processing of data generated by traffic operations. Archived data uses vary widely, from academic research (e.g., traffic flow theory) to traveler information (e.g., "normal" traffic conditions), operations evaluation (e.g., ramp meter algorithms), - performance monitoring, and basic planning-level statistics. Although the operations data in archives are generated in real-time, most of the applications to-date has been historical in nature and outside of the traffic operations area. Data archive applications are still in relative infancy and thus quality assurance procedures are still being established in most areas. Several data archive managers have voiced concerns about the quality of the data generated by operations groups, presumably because the data archive managers have more stringent data quality requirements for their applications than the operations applications. In fact, this concern about archived data quality is part of the genesis for this FHWA-sponsored project. Most current archived data users recognize these data quality issues but maintain an optimistic attitude of "this is the best data I can get for free" and attempt to use the data for various applications. However, interviews conducted in this project revealed several potential data archive consumers that were reluctant to use the data because of real or perceived data quality issues.

As noted previously, data archive applications are still in relative infancy and thus data quality measures are not extensively or consistently used. **Data completeness,** expressed as the number of data samples or the percent of available samples in a summary statistic, is the measure most often used in data archives. The data completeness measure is used frequently because operations data is often aggregated or summarized when loaded into a data archive. For example, the ARTIMIS center in Cincinnati, Ohio/Kentucky reports the number of 30-second data samples (shown in bold in Table 2) that have been used to compute each 15-minute summary statistic.

Table 2. ARTIMIS Reporting of Data Completeness

```
Data for segment SEGK715001 for
07/15/2001
Number of Lanes: 4

#  Time    Samp   Speed   Vol   Occ
00:01:51    30     47     575     6
00:16:51    30     48     503     5
00:31:51    30     48     503     5
00:46:51    30     49     421     4
01:01:52    30     48     274     5
01:16:52    30     42     275    14
...
```

Source: ARTIMIS Data Archives

The Washington State DOT reports **data completeness** as well as **data validity** measures for the Seattle data archives that are distributed on CD-ROM (Ishimaru 1998). In their data archive, they report the number of 20-second data samples in a 5-minute summary statistic (e.g., maximum of 15 data samples possible). A data validity flag (with values of *good*, *bad*, *suspect*, and *disabled loop*) is also included in data reports to indicate the validity of 5-minute statistics (Table 3). Peak hour, peak period, and daily statistics generated by WsDOT's CDR data extraction program also report data validity and completeness summary measures (Table 4). For example, the column headings shown to the right side of Table 4 indicate the number of good ("G"), suspect ("S"), bad ("B"), and disabled ("D") data records. The CDR software also has a data quality mapping utility that allows data users to create location-based summaries of data completeness and validity (Ishimaru and Hallenbeck 1999). This utility is designed for data consumers who would like to analyze the underlying data quality for various purposes.

Table 3. WsDOT Reporting of Data Validity and Completeness

```
***********************************
Filename: 5TO15.DAT
Creation Date: 02/2/98 (Wed)
Creation Time: 03:16:59
File Type: SPREADSHEET
***********************************
ES-145D:_MS___1 I-5 Lake City Way 170.80
09/01/97 (Mon)
---Raw Loop Data Listing---
Time Vol Occ Flg nPds
0:00 49 3.80% 1 15
0:05 37 2.90% 1 15
0:10 38 3.50% 1 15
0:15 34 2.60% 1 15
0:20 48 4.40% 1 15
0:25 44 3.60% 1 15
0:30 35 2.80% 1 15
0:35 33 3.30% 1 15
0:40 28 2.50% 1 15
0:45 30 2.30% 1 15
```

Source: Ishimaru and Hallenbeck 1999

Table 4. WsDOT Reporting of Data Validity and Completeness in Summary Statistics

```
***********************************
Filename: AADT.MDS
Creation Date: 02/2/98 (Thu)
Creation Time: 10:54:09
File Type: SPREADSHEET
***********************************
ES-145D:_MS___1 I-5 Lake City Way 170.80
Monthly Avg for 1996 Jan (Sun)
---Multi-Day Loop Summary Report---
Summary     Valid   Vol    Occ      G   S  B  D  Val Inv Mis
Daily        VAL   19392   7.50% 1133 18  1  0   4   0   0
AM Peak      VAL    1493   3.50%  142  2  0  0   4   0   0
PM Peak      VAL    5069  15.60%  190  2  0  0   4   0   0
AM Pk Hour   VAL    1381  10.00%   47  1  0  0   4   0   0 10:45 11:45
PM Pk Hour   VAL    1576  11.90%   48  0  0  0   4   0   0 13:45 14:45
```

Source: Ishimaru and Hallenbeck 1999

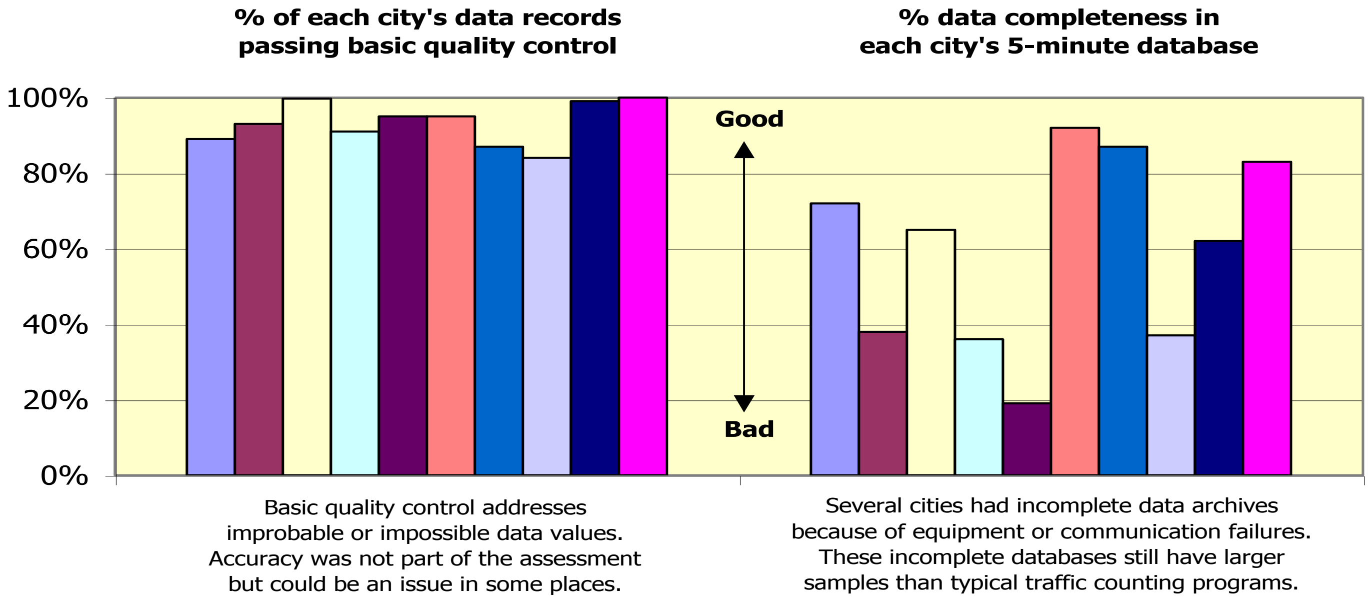
In the FHWA-sponsored Mobility Monitoring Program (http://mobility.tamu.edu/mmp), the Texas Transportation Institute and Cambridge Systematics, Inc. gather archived operations data from numerous traffic management centers nationwide and analyze the archived data to report mobility and reliability trends in the urban areas (Lomax, Turner and Margiotta 2001). As such, the program is an archived data consumer with the primary application of performance monitoring.

The program team performs various data quality checks in the course of processing and analyzing the archived data. In addition to summary statistics on mobility and reliability, performance reports also include information on the following data quality measures:

- **data validity** – percent of records passing quality control checks;
- **data completeness** – percent of records with valid and present values; and
- **data coverage** – percent of freeway centerline-miles with sensor coverage and average sensor spacing.

For example, Figure 1 shows summary information for data validity and data completeness. Significant detail for these data quality measures is also stored in databases. For example, one could do time-based and location-based analyses of data quality using the full database.

Figure 1. Data Quality Statistics for 10 Cities in 2000 Mobility Monitoring Program



Source: Lomax, Turner and Margiotta 2001

Historical/Planning-Level Traffic Monitoring

Data consumers in this group are typically engaged in mid- to long-range (5 to 20-plus years) traffic planning and analysis. Data uses are mostly of an historical nature, so in some cases annual average statistics may not be available (or needed) until six or more months after the past year ends. Thus, the consumer groups' frame of reference for data timeliness differs from the other two groups by an order of magnitude. Whereas operations data consumers may consider data older than 5 minutes unacceptable, planning data consumers may consider waiting up to 9 months for annual statistics to be acceptable. The use of data quality checks or "business rules" for determining the validity of traffic data appears to be fairly common among this group. In many cases, these planning groups serve as the "official source" of traffic data for a particular jurisdiction.

Numerous state departments of transportation (DOTs) use data validation checks or "business rules" when they load traffic data into their information systems. These data quality checks are typically based upon traffic capacity principles, typical traffic trends or patterns, or simply local traffic experience and insight. Thus **data validity** is a common data quality measure using in many historical traffic monitoring groups. For example, the Texas DOT (TxDOT) plans to use 23 business rules for continuous vehicle counts in their Statewide Traffic Analysis and Reporting System (STARS) (TxDOT 2001). Once a data record has failed a business rule, that record is flagged as "suspect" and must be reviewed by a traffic data analyst prior to the beginning of the traffic monitoring program's year-end process. Additionally, STARS uses **data integrity** as a data quality measure as they also run checks on the data file and station integrity.

The traffic monitoring group in the Virginia DOT (VDOT) also uses established business rules to perform traffic **data validity** checks prior to loading them into their information system. As with TxDOT's process, data that fails the business rules are flagged as suspect and must be reviewed by a traffic data analyst. If the traffic data is deemed erroneous, it will not be loaded into the traffic information system. VDOT has a unique contracting arrangement in that they lease the traffic data collection equipment from sub-contractors; thus, they pay the sub-contractors lease payments based upon the quality and completeness of the data collected by the sub-contractors' equipment. For example, a full monthly payment is made for locations "where 25 or more days of useable (for factor creation) classification and volume traffic information are available during a calendar month". A partial lease payment of 50 percent is made "where 15 or more days of useable (for factor creation) volume traffic information, but less than 15 days (useable for factor creation) classification data are available". Thus VDOT's payment for traffic data collection is based on the quality measures of **data validity** and **data completeness**.

VDOT also designates quality levels for their traffic data they distribute. The quality level codes and descriptions are as follows:

- Code 0 - Not Reviewed
- Code 1 - Acceptable for Nothing
- Code 2 - Acceptable for Qualified Raw Data Distribution
- Code 3 - Acceptable for Raw Data Distribution
- Code 4 - Acceptable for use in AADT Calculation
- Code 5 - Acceptable for all TMS uses

These quality codes are designed to indicate to data consumers what the data producers believe to be the fitness of the data for various purposes.

Similar software-based data validity checks are used in several other states. The Pennsylvania and Ohio DOTs both use data validity checks in their traffic information system. These validity checks are performed on a daily basis for all traffic data. The Michigan DOT uses Traffic Data Quality (TDQ), a software tool developed as a result of a pooled-fund study (Flinner and Horsey, no date).

The international experience with traffic data validity checks is comparable to the U.S. experience. A European scanning tour found that several countries perform an automated validation of traffic data (FHWA 1997). All ITS systems observed in the tour countries (the Netherlands, Switzerland, Germany, France, and the United Kingdom) perform some type of automated data validation, usually by comparing current data from a particular site with historical data from that same site during a similar time interval. If an operator identifies questionable data, they use graphic displays to review the data and determine acceptability.

Current Practices in Measuring Data Quality in Other Disciplines

Data quality literature is readily available in several other disciplines, especially the business management and data warehousing industries. The research team conducted a literature review and identified at least two dozen resources that related directly to data quality measures. Selected resources are summarized below with an emphasis on their relevancy to traffic data quality measures.

The geographic information systems (GIS) community has developed standards for documenting data quality in their Spatial Data Transfer Standard (SDTS) (O'Looney 2000; ANSI 1998). The SDTS data quality categories are shown in Table 5. The purpose of the data quality standard within SDTS is not to require acceptable levels of data quality, but to require a data quality report in all GIS data transfers. Following are the SDTS standardized definitions and measures that are to be used in describing and documenting GIS data quality.

Table 5. Five Categories for Data Quality in the Spatial Data Transfer Standard

Category	Definition	Example
Positional Accuracy	The degree of horizontal and vertical control in the coordinate system.	The available precision or detail of longitude and latitude coordinates.
Attribute Accuracy	The degree of error associated with the way thematic data is categorized.	The degree to which a soil description is likely to vary from a soil measurement taken from the corresponding location.
Completeness	The degree to which data is missing and the method of handling missing data.	The ability to estimate crime rates in specific areas may be compromised if data is not available for specific areas.
Logical Consistency	The degree to which there may be contradictory relations in the underlying database.	Location data on some crimes may be based on the place where the crime occurred, while for other crimes the location might be the place where a crime report is taken.
Lineage	The degree to which there is a chronological set of similar data developed using the modeling and processing	Population estimates may not be available for all years; may be estimated on different days of the year; or may be estimated using

	methods.	different estimation techniques and data sources.

Source: O'Looney 2000 and ANSI 1998

Strong, Lee and Wang (1997A, 1997B) suggest four major categories in data quality with 15 dimensions underlying these four categories (Table 6). The authors suggest that traditional quality control techniques (e.g., validity checks, integrity checks, etc.) mostly improve intrinsic data quality dimensions such as accuracy. However, the authors caution that attention to accuracy alone does not correspond to the data consumers' broader data quality concerns. For example, they argue that conventional approaches treat accessibility as a technical systems issues and not a data quality issue. Some data custodians may insist that data is accessible if the physical and software connections are present. The authors suggest, though, that accessibility goes beyond simple technical accessibility; it includes the ease with which the data consumers can manipulate the data to meet their needs.

Table 6. Data Quality Categories and Dimensions

Data Quality Category	Data Quality Dimensions
Intrinsic	• Accuracy • Objectivity • Believability • Reputation
Accessibility	• Accessibility • Security
Contextual	• Relevancy • Value-Added • Timeliness • Completeness • Amount of Information
Representational	• Interpretability • Ease of Understanding • Concise Representation • Consistent Representation

Source: Strong, Lee and Wang (1997A, 1997B)

A relevant analogy to this accessibility issue exists in current practice. Several traffic management centers log detailed traffic data to "file-based archives" where file sizes reach 50-plus MB or the files for a day number in the thousands. These file-based archives are then made available on CD or through the Internet. Some may argue that this data is accessible because it is publicly available. However, the size or nature of the data prevents many data consumers from easily manipulating the data to meet their needs. Thus the authors would argue that these large file-based data archives are not easily accessible to many data consumers.

Wand and Wang (1996) suggest numerous data quality dimensions that distinguish between internal and external views of an information system. External views are concerned with the use and effect of the information system, whereas internal views address the procedures necessary to attain the required functionality that is reflected in an external view. Table 7 contains the various data quality dimensions for both internal and external views.

Table 7. Data Quality Dimensions as Related to Internal or External Views

	Data Quality Dimensions
Internal View (design, operation)	**Data-related** accuracy, reliability, timeliness, completeness, currency, consistency, precision **System-related** reliability
External View (use, value)	**Data-related** timeliness, relevance, content, importance, sufficiency, usability, usefulness, clarity, conciseness, freedom from bias, informative, level of detail, quantitative level, scope, interpretability, understandability **System-related** timeliness, flexibility, format, efficiency

Source: Wand and Wang (1996)

The Department of Defense (DoD) offers a more pragmatic core set of data quality measures for all automated information systems within DoD (Table 8). The DoD also provides guidelines on a total data quality management process and how it can be implemented within the various service units. The guidelines include several real-world examples of data quality management and use of the data quality measures.

The Department of Energy (DOE) has established a Data Quality Objectives (DQO) process and maintains a website on the DQO process at http://dqo.pnl.gov/index.htm. The DQO process is a planning tool for environmental data collection activities that provides a basis for balancing decision uncertainty with available resources. The DQO process is required for all significant data collection projects within DOE's Office of Environmental Management. The DQO process defines 7 steps related to identifying problems, decisions, and inputs, but does not suggest or recommend any specific data quality measures.

Table 8. DoD Core Set of Data Quality Requirements

Data Quality	Characteristics Description	Example Metric
Accuracy	A quality of that which is free of error. A qualitative assessment of freedom from error, with a high assessment corresponding to a small error. (FIPS Pub 11-3)	Percent of values that are correct when compared to the actual value. For example, M=Male when the subject is Male.
Completeness	Completeness is the degree to which values are present in the attributes that require them. (Data Quality Foundation)	Percent of data fields having values entered into them.
Consistency	Consistency is a measure of the degree to which a set of data satisfies a set of constraints. (Data Quality Management and Technology)	Percent of matching values across tables/files/records.
Timeliness	As a synonym for currency, timeliness represents the degree to which specified data values are up to date. (Data Quality Management and Technology)	Percent of data available within a specified threshold time frame (e.g., days, hours, minutes).
Uniqueness	The state of being the only one of its kind. Being without an equal or equivalent.	Percent of records having a unique primary key.
Validity	The quality of data that is founded on an adequate system of classification and is rigorous enough to compel acceptance. (DoD 8320.1-M)	Percent of data having values that fall within their respective domain of allowable values.

Source: DOD Guidelines on Data Quality Management, no date.

The Ken Orr Institute, a systems/software research organization, provides a set of data quality measures very similar to the DoD's data quality measures (Ken Orr Institute, no date). They define these data quality measures as:

- **Accuracy** – The measure or degree of agreement between a data value or set of values and a source assumed to be correct. Also, accuracy is a qualitative assessment of freedom from error.
- **Completeness** – The extent to which values are present in the attributes requiring them.
- **Consistency** – The degree to which data are free from variation or contradiction. Consistency is also a measure of the extent to which a set of data satisfies a set of constraints.
- **Timeliness** – The extent to which a data item or multiple items are provided at the time required or specified. Also, the degree to which specified values are up to date.
- **Uniqueness** – The ability to establish the uniqueness of a data record.
- **Validity** – Data values pass all edits for acceptability, producing the desired results. Also, a measure of the quality of the maintained data, i.e., is it accurate enough to satisfy the acceptance requirements of the classification criteria.

The institute also suggests that quality measures and standards be communicated in several ways:

- Publish organizational data rules for each area by means of metadata;
- Warn of potential missing data sources;
- Clearly establish update schedules; and
- Publish accuracy and deviation results from controlled tests.

Recommended Approaches to Defining and Measuring Traffic Data Quality

Based upon the reviews conducted for this white paper, we recommend the following definition for traffic data quality:

> *Data quality is the fitness of data for all purposes that require it. Measuring data quality requires an understanding of all intended purposes for that data.*

The following data quality measures are recommended:

- **Accuracy** – The measure or degree of agreement between a data value or set of values and a source assumed to be correct. Also, a qualitative assessment of freedom from error, with a high assessment corresponding to a small error.
- **Completeness** (also referred to as availability) – The degree to which data values are present in the attributes (e.g., volume and speed are attributes of traffic) that require them.
- **Validity** – The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values.
- **Timeliness** – The degree to which data values or a set of values are provided at the time required or specified.
- **Coverage** – The degree to which data values in a sample accurately represent the whole of that which is to be measured.
- **Accessibility** (also referred to as usability) – The relative ease with which data can be retrieved and manipulated by data consumers to meet their needs.

There are several other data quality measures that could be appropriate for specific traffic data applications. The six measures presented above, though, are fundamental measures that should be universally considered for measuring data quality in traffic data applications.

At this time, we recommend that goals or target values for these traffic data quality measures be established at the jurisdictional or program level based on a better and more clear understanding of all intended uses of traffic data. It is clear that data consumers' needs and expectations, as well as available resources, vary significantly by implementation program, urban area, or state and preclude the recommendation of a universal goal or standard for these traffic data quality measures.

We also recommend that if data quality is measured, a data quality report be included in metadata that is made available with the actual dataset. The practice of requiring a data quality report using standardized reporting is common in the GIS and other data communities. The American Society of Testing and Materials (ASTM) is currently developing a data archive metadata standard that could be used to document and describe these data quality measures in sufficient detail for data consumers. The ASTM metadata standard under development is being adapted from existing geospatial metadata standards (FGDC-STD-001-1998 and ISO DIS 19115) with their data quality reporting sections intact. Until a formal traffic data archive metadata standard is approved, the traffic data community should create metadata based upon the core elements (i.e., mandatory metadata items) required in these two other geospatial metadata standards.

BIBLIOGRAPHY

American National Standards Institute (ANSI). Spatial Data Transfer Standard (SDTS) –Part 1, Logical Specifications. ANSI NCITS 320-1998, available at http://mcmcweb.er.usgs.gov/sdts/.

English, L.P. *7 Deadly Misconceptions about Information Quality*. INFORMATION IMPACT International, Inc., Brentwood, Tennessee, 1999.

English, L.P. *Improving Data Warehouse and Business Information Quality*. John Wiley & Sons, Inc., New York, New York, 1999.

Federal Highway Administration. *FHWA Study Tour for European Traffic Monitoring Programs and Technologies.* FHWA's Scanning Program, U.S. Department of Transportation, Federal Highway Administration, Washington D.C., August 1997.

Flinner, M. and H. Horsey. *Traffic Data Editing Procedures: Traffic Data Quality (TDQ) Final Report*. Available at http://www.nic-idt.com/tdq/tdq.html, no date.

Ishimaru, J.M. and M.E. Hallenbeck. *FLOW Evaluation Design Technical Report*. Washington State Transportation Center, Seattle, Washington, May 1999.

Ishimaru, J.M. *CDR User's Guide*. Washington State Transportation Center, Seattle, Washington, Version 2.52, March 1998.

ITS America and U.S. Department of Transportation. *Closing the Data Gap: Guidelines for Quality Advanced Traveler Information System (ATIS) Data*. Version 1.0, September 2000.

Jonas, G. "Evaluation of SmarTek SAS-1 Passive Acoustic Detectors on Interstate 17." Arizona Department of Transportation, Phoenix, Arizona, 2001.

Lomax, T., S. Turner and R. Margiotta. *Monitoring Urban Roadways: Using Archived Operations Data for Reliability and Mobility Measurement*. Texas Transportation Institute and Cambridge Systematics, Inc., December 2001.

O'Looney, J. *Beyond Maps: GIS and Decision Making in Local Government*. Environmental Systems Research Institute, Inc., Redlands, California, 2000.

Strong, D.M., Y.W. Lee and R.Y. Wang. 10 Potholes in the Road to Information Quality. *Computer*. Institute of Electrical and Electronic Engineers, August 1997(A), pp. 38-46.

Strong, D.M., Y.W. Lee and R.Y. Wang. Data Quality in Context. *Communications of the ACM*. Association for Computing Machinery, Vol. 40, No. 5, May 1997(B), pp. 103-110.

Tarnoff, P.J. *Getting to the INFOstructure*. White Paper prepared for the TRB Roadway INFOstructure Conference, August 2002.

Turner, S.M., W.L. Eisele, B.J. Gajewski, L.P. Albert, and R.J. Benz. *ITS Data Archiving: Case Study Analyses of San Antonio TransGuide® Data*. Report No. FHWA-PL-99-024. Federal Highway Administration, Texas Transportation Institute, August 1999.

Wand, Y. and R.Y. Wang. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*. Association for Computing Machinery, Vol. 39, No. 11, November 1996, pp. 86-95.