GUIDELINES

FOR

Developing ITS Data Archiving Systems

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Texas Department of Transportation (TxDOT) or the Federal Highway Administration (FHWA). This report does not constitute a standard, specification, or regulation. The engineer in charge of the project is Shawn Turner, P.E. #82781.

The United States Government and the State of Texas do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of this report.

The TransGuide trademark is owned by the Texas Department of Transportation.

ACKNOWLEDGMENTS

The author wishes to acknowledge the support and guidance of Ms. Regina Flores, current project director from TxDOT's Fort Worth District, and Mr. Abed Abukar, former project director currently with Dallas Area Rapid Transit. Additionally, the author acknowledges the comments and input from other members of the project review panel:

Ms. Natalie Bettger, North Central Texas Council of Governments

Ms. Sholeh Karimi, City of Arlington

Mr. Dan Rocha, North Central Texas Council of Governments

The author also acknowledges the following persons from TTI, who assisted in the research project at various times and with various tasks:

Mr. Robert Benz

Mr. Scott Cooner

Mr. Jason Crawford

Mr. William Frawley

Mr. Daryl Puckett

Guidelines for Developing ITS Data Archiving Systems

by

Shawn M. Turner, P.E.

Assistant Research Engineer

Texas Transportation Institute

Report 2127-3

Project Number 0-2127

Research Project Title: Developing Guidance for Sharing Archived/Warehoused ITS Data

Sponsored by the

Texas Department of Transportation

In Cooperation with the

U.S. Department of Transportation

Federal Highway Administration

September 2001

TEXAS TRANSPORTATION INSTITUTE

The Texas A&M University System

College Station, Texas 77843-3135

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

WHAT IS DATA ARCHIVING?

Intelligent transportation system (ITS) data archiving is defined as the systematic retention and re-use of transportation data that is typically collected to fulfill real-time transportation operation and management needs. Data archiving is also referred to as data warehousing or operations data archiving. Transportation operations and their respective sensors and detectors are a potentially rich and detailed source of data about transportation system performance and characteristics. Examples of the most common data elements potentially available from operations include:

- traffic monitoring and detection systems – vehicle volume, speed, travel time, classification, weight, and position trajectories;
- traveler information systems – current traffic conditions (e.g., travel time, speed, or level of congestion), traffic incidents, work zone and/or lane closures;
- traffic control systems – time and location of traffic control actions (e.g., ramp metering, traffic signal control, lane control signals, message board content);
- incident and emergency management systems – location, cause, extent, and time history of roadway incident/emergency detection and clearance; and
- advanced public transit systems – transit vehicle passenger boardings by time and location, vehicle trajectories, passenger origins and destinations, and priority control information.

Later sections of the report (see Table 3 on page 27) provide a comprehensive inventory of the data items that can potentially be collected by ITS applications and operations groups.

WHY ARCHIVE OPERATIONS DATA?

The primary reasons for archiving operations data are:

- provide more and better information in managing and operating the transportation system – The first step in proactive management is knowing where problems are likely to occur before they actually do, then preventing or mitigating the impacts of those problems. Archived operations data can be used to predict when and where problems may occur again, as well as helping to evaluate alternative strategies for preventing or mitigating the problem.

Intelligent transportation system (ITS) data archiving is defined as the systematic retention and re-use of transportation data that is typically collected to fulfill real-time transportation operation and management needs.

Data archiving permits transportation agencies to maximize their investments in data collection infrastructure.

- maximize cost-effectiveness of data collection infrastructure – Data archiving permits transportation agencies to maximize their investments in data collection infrastructure by re-using the same data for numerous transportation planning, design, operations and research needs.
- much less expensive than manual data collection – Data archiving is significantly less expensive than having a planning or design workgroup re-collect even a small percentage of the data using manual methods or special studies.
- established business practice in other industries – The retention and analysis of operational data is an established practice in most competitive industries that use data to manage their business activities. For example, the retail sales industry uses data warehouses full of customer transactions and inventories to better understand the basics of supply and demand in numerous markets around the world.

IF OUR TRANSPORTATION SYSTEM WAS A FACTORY . . .

Consider an analogy that our transportation system is a factory, and that the department of transportation is the factory owner and manager that produces and sells widgets. Now assume that we have implemented technology in the factory that permits us to track the number of widgets that come off the production line every minute. The operations manager reviews this widget tracking data in real-time on the factory floor to make sure that none of the widget machines are malfunctioning.

Now consider the planners for the factory, who are in charge of making sure that sufficient floor space and equipment are available to make widgets. Since the planners are located in the factory's administrative headquarters across town, they send their staff over to the factory one day per year to manually count the number of widget machines in production mode and the number of widgets being produced.

Is the widget data collection one day per year by the factory planners necessary, given that it duplicates the detailed widget tracking data already collected by the operations manager? No, the extra effort by the factory planners costs the factory extra money, and because the manual loading dock counts are not that accurate or detailed, the factory has to keep a large inventory of raw materials. The lack of operational data sharing is ultimately affecting the factory's widget production and profit margin.

WHY DOESN'T EVERYONE WITH ITS ARCHIVE THEIR DATA?

A similar situation is occurring in many areas where technology (i.e., ITS) is being used to operate and manage transportation systems. Some "factory operations managers" are deploying technology and not saving or analyzing "widget data" for other "factory units" to use in making more informed decisions that could increase "factory output and effectiveness."

Implementation and analysis of operations data archives in the U.S. have been somewhat limited as a typical practice (see Table 1) despite the fact that several early "pioneers" have been archiving and analyzing operations data from traffic control sensors and detectors for at least 20 years. As early as the

1970s, the Illinois Department of Transportation (DOT) was saving aggregated loop detector data in Chicago to report "minute-miles" of congestion (1). Similarly in 1968, the Texas Highway Department and the Texas Transportation Institute (TTI) were using an IBM 1800 computer to save and analyze loop detector data along the Gulf Freeway in Houston (2). The archived Houston data were used to support level of service and merging research studies, as well as to demonstrate and quantify the effects of incidents on the freeway corridor. The Washington State DOT (WSDOT) and the University of Washington have been archiving loop detector data from Seattle's freeway traffic management system since 1981, with researchers and planning agencies being primary users (3). Loop detector data from Highway 401 in Toronto, Ontario (Canada), also have been used extensively since the 1980s for traffic flow theory and capacity research.

There are several reasons why operations data archiving and analysis are not more widely implemented:

- some operating workgroups/agencies are focused on crisis management and do not see the utility of anything other than "real-time" data;
- operating workgroups/agencies see data archiving as the responsibility of planning workgroups/agencies, who they feel are the primary beneficiary of archived data;

Data archiving and warehousing is an established business practice in many other competitive industries.

Table 1

1999 Deployment Levels of Operations Data Archiving

Type of System	Type of Data	Agencies Reporting Data Archiving (%)
Freeway Management	Vehicle traffic volumes	87% (59 of 68)
	Vehicle classification	76% (37 of 49)
	Traffic incidents (time sequence of events, location, cause, number of lanes blocked, etc)	67% (35 of 52)
	Vehicle speeds	66% (31 of 47)
	Current and scheduled work zones (location, number of lanes closed, scheduled duration, etc)	53% (34 of 64)
Arterial Street Management	Vehicle traffic volumes	83% (134 of 162)
	Turning movements	83% (94 of 113)
	Traffic incidents	83% (34 of 41)
	Phasing and cycle lengths	80% (91 of 114)
	Vehicle speeds	79% (80 of 101)
	Traffic signal preemption info	75% (46 of 61)
	Current work zones	72% (52 of 72)
	Scheduled work zones	67% (43 of 64)

Source: U.S. DOT Metropolitan ITS Infrastructure Deployment Tracking Database, FY 1999.

- planning workgroups/agencies are typically not involved in the operational data collection, thus they are not aware of or are not comfortable with the quality of the data to be archived;
- data archiving was not considered an essential component of traffic control/management software during system development; and
- there may be data ownership, maintenance, or control issues that cannot be resolved between workgroups/agencies that collect data and archive data.

In some cases, operations data are archived but have not been widely distributed or analyzed for several reasons:

- proprietary data formats and data storage devices (e.g., magnetic tape cartridges) hinder archived data distribution;
- distributing archived data to users sometimes places an unreasonable burden on operations personnel (if the distribution is not automated); and
- before the Internet and CD technology arrived in the early 1990s, it was difficult to distribute the large quantities of data that were typically stored in proprietary data formats and data storage devices (e.g., magnetic tape cartridges).

Interest in operations data archiving has grown in the late 1990s, due in part to the formation of the archived data user service (ADUS) in the National ITS Architecture in 1999 (4). The increased visibility of data archiving in the National ITS Architecture has presumably established its legitimacy and importance, as there are more agencies that are now planning data archiving systems than five years ago. Table 1, which shows the 1999 level of deployment of data archiving, serves as a useful benchmark as data archiving becomes more integrated into the operations processes.

OVERVIEW OF GUIDE

This report contains three basic chapters that summarize guidance on data archiving systems. The three chapters are as follows:

Chapter 1. Introduction – provides an introduction to data archiving and the relevant issues;

Chapter 2. Basic Principles of Data Archiving – provides a summary of data archiving principles that can be applied regardless of archive size or design. This chapter also includes case studies that illustrate these basic principles.

Chapter 3. Summary of Technical Issues – provides a summary of technical issues related to a) what data to save and how much? b) performing quality control on archived data; and c) using the National ITS Architecture and related data standards to develop data archiving systems.

FURTHER READING

The following are suggested as further reading for those interested in data archiving:

- *ITS as a Data Resource: Preliminary Requirements for a User Service*, Report No. FHWA-PL-98-031, April 1998 (5).
- "Archived Data User Service (ADUS): An Addendum to the ITS Program Plan," Version 3, September 1998 (4).
- ITS America's Archived Data User Service Resource Page, http://www.itsa.org, search for "Archived Data User Service Resource Page."
- Archived Data User Service in the National ITS Architecture, http://www.iteris.com/itsarch.
- *ITS Data Archiving Resources: Resources for Implementing ADUS*, CD developed for Federal Highway Administration by Texas Transportation Institute, 2000.

CHAPTER 2

BASIC PRINCIPLES OF DATA ARCHIVING

This chapter provides a summary of data archiving principles that can be applied regardless of archive size or design. This chapter also includes case studies that illustrate some of these basic principles. The basic principles are as follows:

- Determine the workgroup(s) or agency(ies) that should have primary responsibility for operating and maintaining the data archive;
- Start small but think long-term, and begin with modest prototypes focused on a single source of data (e.g., freeway or arterial street detector systems);
- Develop the data archiving system in a way that permits ordinary users with typical desktop computers to access and analyze the data;
- Provide access to and distribution of archived data through the Internet or portable storage devices such as CDs or DVDs;
- Save original data as collected from the field for some specified period of time, but make summaries of this data available for most users;
- Use quality control methods to flag or remove suspect or erroneous data from the data archive; and
- Provide adequate documentation on the data archive and the corresponding data collection system.

BASIC PRINCIPLES

In surveying and talking with people operating data archiving systems outside of Texas, we found several areas that had developed effective data archiving systems that enable ordinary computer users to access large databases of archived ITS data. We conducted in-depth studies of those areas that have already developed effective data archiving systems and found that these systems and programs have several common characteristics. These common characteristics developed the basic principles that are discussed in the next few pages.

Determine the workgroup(s) or agency(ies) that should have primary responsibility for operating and maintaining the data archive. This may seem like a simple matter; in many cases, though, data archiving systems have not been further developed because no one has taken responsibility for their operation and maintenance. The lead workgroup or agency in operating a data archive will vary from location to location depending upon the primary data users as well as the institutional relationships and resources of the stakeholders in a region.

The lead workgroup or agency in operating a data archive will vary from location to location depending upon the primary data users as well as the institutional relationships and resources of the stakeholders in a region.

Focus on detector data as the first element in data archiving prototypes.

Discussion and dialogue in early stages among all stakeholders should assess the demand for archived data as well as the strengths and weaknesses of which agency or workgroup in a region maintains data archives. Numerous agencies in the state of Maryland used this approach to determine who should have primary responsibility for operating and maintaining data archives for each region and the state (6).

In some cases, there may be several agencies that each operate their own data archive, but which are connected and integrated through a "virtual data warehouse" (see the data archiving architecture section in Chapter 3). In other cases, it may be logical for a regional planning agency with strong information management capabilities to warehouse data that can be shared among other agencies in the region. In other situations, the operating workgroup or agency may wish to operate and manage a data archive because they have the most resources and will be primary users of the data.

Start small but think long-term, and begin with modest prototypes focused on a single source of data (e.g., freeway or arterial street detector systems). Our research indicates that several of the most effective data archiving systems started as small prototypes that took existing detector data files (which are large, multi-million record text files) and made this data easily accessible to typical computer users. This "start small but think long-term" approach comes from other industries, where large, complex data warehousing efforts have failed or struggled for years to get started by trying to "be all things to all people."

Several user requirement studies consistently show that detector data (i.e., volumes and speeds) are the most desired archived data (7,8,9). Thus, it makes sense to focus on detector data as the first element in data archiving prototypes. Most data archiving systems provide the capability to summarize the detector data to various levels in space (e.g., lane-by-lane, across all lanes for a roadway link, or a collection of links along a facility) or in time (5-, 15-, and 60-minute summaries).

Develop the data archiving system in a way that permits ordinary users with typical desktop computers to access and analyze the data. Our review shows that effective data archiving systems make large operations data archives available to ordinary computer users without requiring them to have specialized database or programming skills. These systems use a "point-and-click" interface, either through a Windows-based application or a web browser, to provide access to the data archives. These data archive or data warehouse interfaces are often available as commercial, off-the-shelf products, or they may come pre-packaged with some relational databases.

The most widespread practice for data archiving, though, is logging original field-collected data to a single large or numerous text files. The size and format of these files make them difficult to use for most engineers and planners. Researchers and other "power users" are the most common user groups that have been able to access and analyze these large and cumbersome text file-based archives. It is clear that data archives will have to be oriented more toward typical computer users if they are to be most effectively utilized by the agencies that need the data.

Provide access to and distribution of archived data through the Internet or portable storage devices such as CDs or DVDs. Our research indicated that Internet-based access and distribution of data were some of the most common

and effective means to share archived data. In some cases, Internet-based archives consisted of a "point-and-click" query interface that allows users to summarize or analyze the data using tools or applications on the web site. In other cases, the Internet or FTP site simply allowed users to download data, requiring users to have the analysis tools or applications on their own desktop computer.

CDs or DVDs are used as an alternative to Internet-based data archives. They permit the data archiving agency to maintain greater control and security over the data. Intranets or "private networks" have also been proposed as a way to exercise greater control and security over data archives than what is available through the Internet. As with Internet-based archives, some CD-based archives provide data and reporting tools, whereas other archives simply contain raw or summary data.

Save original data as collected from the field for some specified period of time, but make summaries of this data available for most users. Most detector data is collected from the field at a very detailed level (between 20 seconds and 1 minute); however, most users do not need this level of detail. Many data archiving systems aggregate data to a consistent time interval (5 minutes is most common) for loading into a data archive. Because there will always be some users interested in the original data, a mechanism should be developed to store this for a short period of time or to store it permanently off-line. Plummeting data storage costs make off-line compressed storage (such as a CD or DVD) an inexpensive option that preserves the original data.

Use quality control methods to flag or remove suspect or erroneous data from the data archive. Although most quality control methods being used in data archiving systems are relatively simple, they typically identify the majority of serious data errors or problems. Some areas are experimenting with more sophisticated quality control methods. The rigor of the quality control ultimately depends upon how and for what purpose the data will be used. An entire section in Chapter 3 contains more details on quality control methods.

Two different philosophies exist for what to do with data that has failed quality control: 1) simply identify or flag the data records that have failed quality control; or 2) remove the data records that have failed quality control and replace with better estimates. Again, these business rules (for how to deal with data failing quality control) will depend upon who will be using the data and for what purpose. There is no single correct answer for quality control.

Provide adequate documentation on the data archive and the corresponding data collection system. With data archiving systems, many data users will be from outside the operations workgroup or agency that collected the data. Thus, they may have little knowledge about the operations data that is collected, how it is collected, and how it is processed by operations before it is archived. Documentation is often the last task in computer system development and integration but, in this case, it may be the most important.

Adequate documentation for data archives primarily includes (but is not limited to) these things:

- overall documentation on the data archive, such as the data elements (and definitions) that are available, the various databases and tables that are maintained, and other items that might typically fall under a "data schema";

Effectice data archiving systems make large operations data archive available to ordinary computer users without requiring them to have specicalized database or programming skills.

- an "audit trail" of how the data have been processed since they were collected in the field. This audit trail includes information about the results of quality control, any summarization or aggregation steps, and any estimates or changes that have been made to original, field-collected data; and
- documentation on the data collection system, such as the type, location, and other identification for detectors, the detectors that were considered "online" for a particular hour or day, and information about equipment calibration and maintenance.

CASE STUDIES

This section provides case study summaries of several data archiving systems as a way to illustrate some of the basic principles discussed above. The case studies are from:

- Austin, Texas;
- PeMS in California;
- Seattle, Washington; and
- FHWA's Mobility Monitoring Program.

Austin, Texas

The Operations Section of TxDOT's Austin District currently archives freeway detector data to a computer server and then provides the data on CD upon request. The data are currently saved in a comma-separated values (csv) ASCII-text format with a separate file created for each hour of the day for each freeway corridor being monitored (e.g., approximately 8,760 files per corridor per year). Analysis of these original data files has proven difficult and time-consuming for typical engineers or analysts. To date, advanced computer skills and sophisticated database software have been necessary to use the data in any meaningful way.

The Austin case study is provided as an example of a modest ITS deployment that currently archives detector data to a large number of text files. TTI has proposed a simple approach (*10*) that does the following to improve the usability of TxDOT's original detector data archive:

- maintains the original 1-minute data as collected from the field on CD;
- summarizes the original 1-minute data to 5-, 15-, and 60-minute summary statistics so as to fit into most spreadsheets;
- performs quality control and identifies number of failed records in summary statistics;
- re-formats the data into a similar csv-text file format that can be imported into most spreadsheets;
- organizes the data by date and by location; and
- distributes the compressed and summarized text files on a single CD or through an Internet site.

PeMS in California

The Operations Division in Caltrans' Headquarters office has worked with researchers at the University of California at Berkeley in creating PeMS, a freeway Performance Measurement System (*11*). PeMS gathers raw freeway detector data in real-time from several of Caltrans' districts, including Los Angeles, Orange County, and Sacramento. The detector data for these participating districts are summarized to a common 5-minute time interval, then loaded into the PeMS data warehouse. The data archives are then made available through the Internet (http://transacct.eecs.berkeley.edu) for anyone that has access privileges (i.e., the site is password-protected). PeMS has several built-in data summary and reporting tools on the web site. As its name states, the primary use of PeMS is for monitoring freeway performance using speeds, estimated travel times, and vehicle volumes.

The impetus for this data archive was state legislation that required Caltrans to monitor the performance of their transportation system. Because Caltrans has extensive detector coverage on freeways in several districts, they chose to archive existing data rather than manually re-collect system performance data. Caltrans' PeMS data warehouse is unique because it is one of the few statewide operations data archives in existence. Time and experience will reveal how useful a centralized statewide data archive is to local agencies and workgroups at the district level.

Seattle, Washington

The Washington State DOT and the Washington State Transportation Center (at University of Washington) have developed a CD-based data archive for Seattle freeways, which they have used to distribute archived operations data for at least the past five years. The freeway detector data are collected every 20-seconds from field controllers, but the data are summarized to the 5-minute level in the data archive. Quality control is also performed before the detector data is loaded into the archive, and the archive documents the number of data records that have failed quality control. Each data archive CD contains data extraction and summary tools (*12*).

WSDOT has been archiving freeway detector data since 1981 in some shape or form, although early efforts were difficult because of the expense of data storage and the difficulty of data transfer (pre-Internet). The agencies have made numerous improvements to their data archive over the years and, for the most part, the data archives have been institutionalized within WSDOT. In their data archiving system, a CD is used to hold three months of 5-minute summary data and the CDs are available upon request. Seattle is an example of starting small but making incremental improvements as demand for the data increases. The data are used for a wide variety of purposes, including testing and evaluating of operational improvements such as ramp metering or HOV lanes, freeway performance monitoring, pavement design, and freight performance analysis.

FHWA's Mobility Monitoring Program

The FHWA, with support from TTI and Cambridge Systematics, has gathered archived freeway detector data from ten cities for the year 2000 to develop a performance monitoring program (*13*). The archived data are gathered in a variety of formats but are then summarized to a standard 5-minute, lane-by-lane format for further processing and analysis. The primary purpose of this multi-

Documentation is often the last task in computer system development and integration but, in this case, it may be the most important.

city data archive is to support performance monitoring of mobility and reliability at the city and national level; however, the data archives can and will be used for a variety of other analyses and applications.

The ten cities participating in this past year's (year 2000 data) program were as follows: Atlanta, Cincinnati, Detroit, Hampton Roads, Houston, Los Angeles, Minneapolis-St. Paul, Phoenix, Seattle, and San Antonio. Another ten to fifteen cities that have freeway detector data archives will be added to next year's (year 2001 data) program. Additionally, TTI and Cambridge Systematics are planning to experiment with using arterial street detector data archives for performance monitoring in the near future.

The archived data processing and analysis in this program follow many of the basic principles, such as:

- Initial efforts were focused on a single source of data – freeway detector data.
- Original data were saved off-line on CDs and summary data are kept on-line for most analyses.
- Basic quality control methods were used to identify and remove suspect or erroneous data.
- Documentation was provided at the data record level as well as the database level for the data processing that had been performed, as well as the data collection system for each city.

CHAPTER 3

SUMMARY OF TECHNICAL ISSUES

This chapter provides a summary of technical issues related to:

- what data to save and how much;
- performing quality control on archived data; and
- using the National ITS Architecture and related data standards to develop data archiving systems.

WHAT TO SAVE AND HOW MUCH?

The potential uses and applications of archived ITS data are as diverse as the data user groups that wish to obtain the data (Table 2). The data needs for some of these applications are currently fulfilled through the manual collection of traffic data, which often suffers from inadequate breadth or depth. Other data needs are met through estimation and computer simulation techniques, while some data needs simply continue to go unmet.

Given the wide variety of potential archived data users, it is likely that all data elements generated by ITS sources could be useful to other archived data users at some time. In developing data archiving and warehousing systems, though, it may be necessary to prioritize certain data elements by their inherent value to archived data stakeholders. If prioritization of data to be archived is necessary, it should be done by archived data managers in consultation with the archived data users. Past experience has indicated that several types of archived data are valuable to more than one data stakeholder group:

- traffic condition data: traffic volumes, vehicle speeds and travel times, vehicle classification, and closed-circuit television (CCTV) images;
- construction and work zone data: location, time, date, and extent of blockage/closure;
- traffic incident logs: time sequence of events (detection, notification, arrival, and clearance), location, extent/severity, and cause; and
- traffic control responses: dynamic message sign (DMS) messages, ramp meter timing, etc.

Table 3 contains a comprehensive inventory of data elements that potentially could be archived. The table was created from the data flows in the National ITS Architecture at the time the archived data user service (ADUS) was created in 1999. This table can be used as a starting point in determining what types

In developing data archiving and warehousing systems, though, it may be necessary to prioritize certain data elements by their inherent value to archived data stakeholders.

Table 2

Stakeholders and Example Applications for Archived ITS Data

Stakeholder Group	Primary Transportation—Related Functions	Example Applications
MPO and state transportation planners	Identifying multimodal passenger transportation improvements (long- and short-range); congestion management; air quality planning; develop and maintain forecasting and simulation models	▮ congestion monitoring ▮ link speeds for TDF and air quality models ▮ AADT, K- and D-factor estimation ▮ temporal traffic distributions ▮ macroscopic traffic simulation ▮ HOV, paratransit, and multimodal demand estimation ▮ congestion pricing policy
Traffic management operators	Day-to-day operations of deployed ITS (e.g., Traffic Management Centers, Incident Management Programs)	▮ pre-planned control strategies (ramp metering and signal timing) ▮ highway capacity analysis ▮ microscopic traffic simulation ▮ dynamic traffic assignment ▮ incident management ▮ congestion pricing operations ▮ evaluation and performance monitoring
Transit operators	Day-to-day transit operations: scheduling, route delineation, fare pricing, vehicle maintenance; transit management systems; evaluation and planning	▮ capital planning and budgeting ▮ corridor analysis planning ▮ maintenance planning ▮ market research ▮ operations/service planning ▮ performance analysis planning
Air quality analysts	Regional air quality monitoring; transportation plan conformity with air quality standards and goals	▮ emission rate modeling ▮ urban airshed modeling
MPO/state freight and intermodal planners	Planning for intermodal freight transfer and port facilities	▮ truck O-D flow patterns ▮ HazMat and other commodity flow patterns
Safety planners and administrators	Identifying countermeasures for general safety problems or hotspots	▮ safety reviews of proposed projects ▮ high crash location analysis ▮ generalized safety relationships for vehicle and highway design ▮ countermeasure effectiveness (specific geometric and vehicle strategies) ▮ safety policy effectiveness
Maintenance personnel	Planning for the rehabilitation and replacement of pavements, bridges, and roadside appurtenances; scheduling of maintenance activities	▮ pavement design (loadings based on ESALs) ▮ bridge design (loadings from the "bridge formula") ▮ pavement and bridge performance models ▮ construction and maintenance scheduling

Table 2–Continued

Stakeholders and Example Applications for Archived ITS Data

Stakeholder Group	Primary Transportation–Related Functions	Example Applications
Commercial vehicle enforcement personnel	Accident investigations; enforcement of commercial vehicle regulations	▮ HazMat response and enforcement ▮ intermodal access ▮ truck route designation and maintenance ▮ truck safety mitigation
Emergency management services (local police, fire, and emergency medical)	Response to transportation incidents; accident investigations	▮ labor and patrol planning ▮ route planning for emergency response ▮ emergency response time planning ▮ crash data collection
Transportation researchers	Development of forecasting and simulation models and other analytic methods; improvements in data collection practices	▮ car-following and traffic flow theory development ▮ urban travel activity analysis
Private sector users	Provision of traffic condition data and route guidance (Information Service Providers); commercial trip planning to avoid congestion (carriers)	

Source: adapted from Margiotta 1998 (5), pp. 4-5.

of data are being collected in a particular region and which data are of most interest for archiving.

Because of the detailed nature of ITS detector data (typically collected every 20 to 30 seconds), data aggregation is often a consideration when archiving ITS data. Aggregation refers to the time interval at which data are summarized. For example, several data archiving systems aggregate 20-second speeds and volumes to 5-minute average speeds and volume subtotals. Aggregation is done primarily to save computer storage space and to reduce data processing time when analyzing or further summarizing archived data. Additionally, aggregation is mostly considered only for traffic condition data (i.e., speed, travel time, volume, occupancy) from detectors or sensors and not for event-based data such as incident response information.

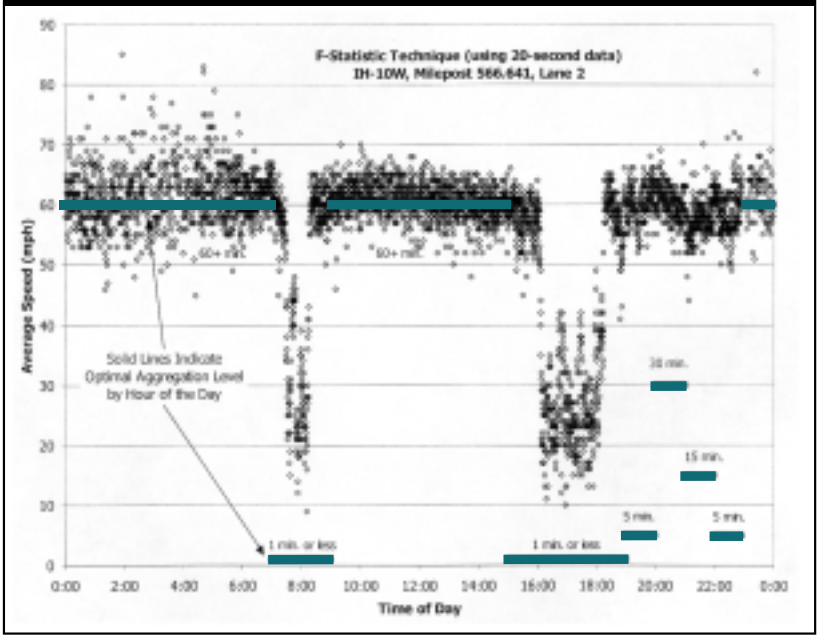
Selecting the "best" aggregation level is a local decision that is best informed by data user needs/requirements as well as available data management resources. Aggregation levels used in ITS data archiving systems around the country vary considerably, ranging from saving raw data (20 to 30 seconds) to summarizing data (15 minutes). Transportation planners typically only require 15-minute summaries at most; whereas, researchers may require the most detailed data possible for sophisticated analyses. Because of the wide-ranging nature of uses for archived data, in some cases the recommendation for aggregation level is "save as much as you can afford."

In some areas, the data user needs/requirements may not be clearly defined, or someone may suggest that the user needs will change or evolve as more detailed ITS data becomes available. In situations like this, the "best" data

Because of the wide-ranging nature of uses for archived data, in some cases the recommendation for aggregation level is "save as much as you can afford."

Figure 1

Typical Speed Profile and Optimal Statistical Aggregation Level



aggregation level can be based upon the statistical variability of the data itself. This approach ensures that, regardless of the use of the data, the aggregation level will capture all of the variation within the data. For example, consider Figure 1, which shows average speeds at a location throughout the day. One can clearly see that the variation of speeds increases during the peak hours. Statistical techniques have been developed to calculate the appropriate aggregation level (as shown in Figure 1) given this variation throughout the day (*14*).

Previous research (*14*) has indicated that there is a range of possible aggregation solutions that range from simple to complex (Figure 2). The particular solution for an area will depend upon local capabilities, needs, and resources. To date, the focus in many traffic management centers has been mostly on simple solutions, such as selecting 5- or 15-minute aggregation levels based on existing data needs and data management capabilities.

QUALITY CONTROL FOR ARCHIVED DATA

Quality control has been defined as a system of techniques for economically producing goods and services that meet the customer's requirements (*15*). As it pertains to archived operations data, quality control means using methods to produce databases and information of a sufficient quality to meet data users' needs. Quality control techniques for archived data should encompass at least these three data attributes:

- suspect or erroneous data – identifying and "treating" illogical or improbable data values that do not fall within expected ranges or meet established principles or rules;
- missing data – identifying and "treating" expected data values that are missing because of hardware/software malfunction or quality control edits; and

Table 3

ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	Spatial coverage	Temporal coverage		
FREEWAY AND TOLL COLLECTION						
Freeway traffic flow surveillance data	▮ volume ▮ speed ▮ occupancy	▮ loop detectors ▮ video imaging ▮ acoustic ▮ radar ▮ microwave	usually spaced at <=1 mile; by lane	sensors report at 20- to 60-second intervals	▮ ramp meter timing ▮ incident detection ▮ congestion/queue identification	▮ congestion monitoring ▮ link speeds for planning and air quality models ▮ AADT, K- and D-factors ▮ saturation flow rates ▮ pre-planned TMC operations
	▮ vehicle classification ▮ vehicle weight	▮ loop detectors ▮ weigh-in-motion ▮ video imaging ▮ acoustic	usually 50-100 per state; by lane	usually hourly	pre-screening for weight enforcement	▮ truck percents by time-of-day for demand forecasting and air quality models ▮ truck flow patterns ▮ pavement loadings
Ramp meter and traffic signal preemptions	▮ time of preemption ▮ location	field controllers	at traffic control devices only	usually full-time	Priority to transit, HOV, and EMS vehicles	▮ network details for microscopic traffic simulation models
Ramp meter and traffic signal cycle lengths	▮ begin time ▮ end time ▮ location ▮ cycle length	field controllers	at traffic control devices only	usually full-time	Adapt traffic control response to actual traffic conditions	▮ network details for microscopic traffic simulation models (e.g. traf, transims) ▮ preplanned tms operations
Visual and video surveillance data	▮ time ▮ location ▮ queue length ▮ vehicle trajectories ▮ vehicle classification ▮ vehicle occupancy	▮ cctv ▮ aerial videos ▮ image processing technology	selected locations	usually full-time	▮ coordinate traffic control response ▮ congestion/queue identification ▮ incident verification	▮ congestion monitoring ▮ car-following and traffic flow theory

Table 3—Continued

ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	Spatial coverage	Temporal coverage		
FREEWAY AND TOLL COLLECTION, continued						
Vehicle counts from electronic toll collection	▮ time ▮ location ▮ vehicle counts	electronic toll collections equipment	at instrumented toll lanes	usually full-time	automatic toll collection	▮ traffic counts by time of day
TMC-generated	▮ link congestion indices	TMC software	selected roadway	usually full-time	▮ incident detection	▮ congestion monitoring
Traffic flow metrics	▮ stops/delay estimates		segments		▮ traveler information ▮ control strategies	▮ effectiveness of prediction methods
ARTERIAL AND PARKING MANAGEMENT						
Arterial traffic flow surveillance data	▮ volume ▮ speed ▮ occupancy	▮ loop detectors ▮ video imaging ▮ acoustic ▮ radar ▮ microwave	usually midblock at selected locations only ("system detectors")	Sensors report at 20- to 60-second intervals	▮ progression setting ▮ congestion/queue identification	▮ congestion monitoring ▮ link speeds for travel forecasting models (free flow only) ▮ AADT, K- and D-factors
Traffic signal phasing and offsets	▮ begin time ▮ end time ▮ location ▮ up/downstream offsets	field controllers	at traffic control devices only	usually full-time	adapt traffic control response to actual traffic conditions	network details for microscopic traffic simulation models
Parking management	▮ time ▮ lot location ▮ available spaces	field controllers	selected parking facilities	usually day time or special events	real-time information to travelers on parking availability	parking utilization and needs studies

Table 3—Continued

ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	Spatial coverage	Temporal coverage		
TRANSIT AND RIDESHARING						
Transit usage	▮ vehicle boardings (by time and location) ▮ station origin and destination (O/D) ▮ paratransit O/D	electronic fare payment systems	transit routes	usually full-time	used for electronic payment of transit fares	▮ route planning/run-cutting ▮ ridership reporting (e.g., Federal Transit Administration Section 15)
Transit route deviations and advisories	▮ route number ▮ time of advisory ▮ route segments taken	TMC software	transit routes	usually full-time	transit route revisions	transit route and schedule planning
Rideshare requests	▮ time of day ▮ O/D	computer-aided dispatch (CAD)	usually areawide	daytime, usually peak periods	dynamic rideshare matching	▮ travel demand estimation ▮ transit route and service planning
Transit priority control	▮ time, location and duration of priority vehicle preemption	field controllers and on-vehicle equipment	intersections under priority control	usually full-time (during periods of operation)	priority vehicle preemption at signalized intersections	▮ transit route and service planning ▮ signal re-timing/adjustments

Table 3—Continued

ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	Spatial coverage	Temporal coverage		
INCIDENT MANAGEMENT AND SAFETY						
Incident logs	▮ location ▮ begin, notification, dispatch, arrive, clear, depart times ▮ type ▮ extent (blockage) ▮ HazMat ▮ police accident report reference ▮ cause	▮ CAD ▮ computer-driven logs	extent of incident management program	extent of incident management program	incident response and clearance	▮ incident response evaluations (program effectiveness) ▮ congestion monitoring (e.g., percent recurring vs. nonrecurring) ▮ safety reviews (change in incident rates)
Train arrivals at highway-rail intersections	▮ location ▮ begin time ▮ end time	field controllers	at instrument highway-rail intersections	usually full-time	▮ coordination with nearby traffic signals ▮ notification to travelers	grade crossing safety and operational studies
Emergency vehicle dispatch records	▮ time ▮ O/D ▮ route ▮ notification, arrive, scene, leave times	CAD	usually areawide	usually full-time	coordination of emergency management response	▮ emergency management labor and patrol studies ▮ emergency management route planning
Emergency vehicle locations	▮ vehicle type ▮ time ▮ location ▮ response type	automatic vehicle identification (AVI) or GPS equipment	usually areawide	usually full-time	▮ tracking vehicle progress ▮ green wave and signal preemption initiation	▮ emergency management route planning ▮ emergency management response time studies

Table 3—Continued

ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source			Real-time uses	Possible multiple uses of ITS-generated data
		Typical collection equipment	Spatial coverage	Temporal coverage		
INCIDENT MANAGEMENT AND SAFETY, continued						
Construction and work zone identification	▮ location ▮ date ▮ time ▮ lanes/shoulders blocked	TMC software			traveler information	congestion monitoring
COMMERCIAL VEHICLE OPERATIONS						
HazMat cargo identifiers	▮ type ▮ container/package ▮ route ▮ time	commercial vehicle operations (CVO) systems	at reader and sensor locations	usually full-time	▮ identifying HazMat in specific incidents ▮ routes for specific shipments	▮ HazMat flows ▮ HazMat incident studies
Fleet activity reports	▮ carrier ▮ citations ▮ accidents ▮ inspection results	CVO inspections	N/A	usually summarized annually	May overlap with SAFETYNET functions	
Cargo identification	▮ cargo type ▮ O/D	CVO inspections	at reader and sensor locations	usually full-time	clearance activities	freight movement patterns
Border crossings	▮ counts by vehicle type ▮ cargo type ▮ O/D	CVO inspections	at reader and sensor locations	usually full-time	enforcement	freight movement patterns

Table 3—Continued

ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source: Typical collection equipment	Features of the Data Source: Spatial coverage	Features of the Data Source: Temporal coverage	Real-time uses	Possible multiple uses of ITS-generated data
INCIDENT MANAGEMENT AND SAFETY, continued						
On-board safety data	▮ vehicle type ▮ cumulative mileage ▮ driver log (hours of service) ▮ subsystem status (e.g., brakes)	CVO inspections	at reader and sensor locations	usually full-time	enforcement and inspection	special safety studies (e.g., driver fatigue, vehicle components)
ENVIRONMENTAL AND WEATHER						
Emissions management system	▮ time ▮ location ▮ pollutant concentrations ▮ wind conditions	specialized sensors	at sensor locations	usually full-time	identification of hotspots and subsequent control strategies	▮ trends in emissions ▮ special air quality studies
Weather data	▮ location ▮ time ▮ precipitation ▮ temperature ▮ wind conditions	environmental sensors	at sensor locations	usually full-time	traveler information	▮ congestion monitoring (capacity reductions) ▮ freeze/thaw cycles for pavement models
VEHICLE AND PASSENGER INFORMATION						
Location referencing data	Special case; pertains to all location references in ITS and planning					need conversion from latitude/longitude to highway distance and location (e.g., milepost references for queue lengths)

Table 3—Continued

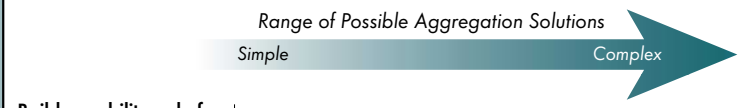
ITS Data Relevant for Archiving

ITS data source	Primary data elements	Features of the Data Source: Typical collection equipment	Features of the Data Source: Spatial coverage	Features of the Data Source: Temporal coverage	Real-time uses	Possible multiple uses of ITS-generated data
VEHICLE AND PASSENGER INFORMATION, continued						
Vehicle probe data	▮ vehicle ID ▮ segment location ▮ travel time	▮ probe readers and vehicle tags ▮ GPS on vehicles	GPS is areawide; readers restricted to highway locations	usually full-time	▮ coordinate traffic control response ▮ congestion/queue identification ▮ incident detection ▮ real-time transit vehicle schedule ▮ adherence ▮ electronic toll collection	▮ congestion monitoring ▮ link speeds for travel forecasting models ▮ historic transit schedule adherence ▮ traveler response to incidents or traveler information ▮ O/D patterns
VMS messages	▮ VMS location ▮ time of message ▮ message content	TMC sofware	VMS locations	hours of TMC operation	traveler information	effects of VMS message content on traveler response
Vehicle trajectories	▮ location (route) ▮ time ▮ speed ▮ acceleration ▮ headway	▮ AVI or GPS equipment ▮ advanced video image processing	AVI restricted to reader locations; GPS is areawide	1- to 10-second intervals	collected as part of surveillance function	▮ traffic simulation model calibration for local conditions (driver type distributions) ▮ modal emission model calibration ▮ traffic flow research
TMC and Information Service Provider generated route guidance	▮ time/date ▮ O/D ▮ route segments ▮ estimated travel time	TMC/information service provider software	usually areawide	hours of TMC operation	traveler information	▮ O/Ds for traffic simulation model inputs ▮ interzonal travel times for traffic simulation model calibration
Parking and roadway pricing changes	▮ time/date ▮ route segment/lot ID ▮ new price	TMC software	facilities subject to variable pricing	hours of TMC operation	demand management	▮ special studies of traveler response to pricing ▮ establishment of pricing policies

Source: *Margiotta 1998 (5), pp.10-13*

Figure 2

Application of Statistical Methods to Data Aggregation



Build capability only for existing data needs	Build flexibility into archiving system for unanticipated needs		
	Single Fixed Aggregation Level	Multiple Fixed Aggregation Levels	Dynamic Aggregation Levels
▪ Statistical methods not necessary ▪ Aggregation level based on existing data needs and analysis methods ▪ Example: Save only 15-minute data because that is the minimum level that is currently needed	▪ Use these research results without further analysis OR Use statistical methods with local data to determine single or multiple optimal aggregation levels ▪ May also consider tradeoffs between aggregation and available resources		▪ Archiving system automates statistical methods, applied at regular intervals ▪ Aggregation level may vary during a day due to differences in traffic variability ▪ Example: Aggregation is automatically performed at end of day using statistical algorithms; differences in aggregation are transparent to users
	▪ Aggregation level same throughout the day based on minimum optimal aggregation for all time periods ▪ Example: Save raw data throughout the day because that is the minimum for the day	▪ Aggregation level may vary for fixed periods during a day (peak vs. off-peak) ▪ Example: Save raw data during peak period, 15-minute data during off-peak period	

- inaccurate data – identifying and "treating" data values that are systematically inaccurate (but within the range of plausible values) because of equipment measurement error (e.g., equipment improperly calibrated).

This section contains an overview of quality control processes for archived data, and provides some specific examples for each of the three quality control attributes listed above.

Overview

Data quality has been noted as one of the primary concerns of archived operations data users (*14*). Some of the concerns may be attributed to the fact that these large data sets are new to many data users; thus, there is some unfamiliarity with the inherent quality of the data. In some cases, the operations center may not need data as accurate as archived data users; thus, they are less concerned with detailed accuracy. The concern with data quality also may be relevant because, in some cases, only minimal error detection is performed as the data are being collected in real-time.

As with any data collection or analysis effort, data quality should be an important consideration in designing data archiving or analysis systems. Quality control procedures are especially critical with operations data for several reasons: 1) the potentially large volume of operations data makes it difficult to detect errors using traditional manual techniques; 2) the continuous monitoring nature of operations data implies that equipment errors and malfunctions are more likely than during periodic data collection efforts; and 3) archived data users may have different (potentially more stringent) quality requirements than real-time users of that same data.

Quality control can be performed at one or several places as operations data are being collected, transmitted, and processed. For example, roadside detector controllers may do simple error checking in real-time; then more extensive

quality control can be performed as the data are loaded into a permanent data archive or data warehouse. Regardless of where quality control is performed, it is important to mark or "flag" data values that have failed quality control or have been modified by quality control processes. These quality control flags help database managers and analysts to more accurately interpret and manage suspect or erroneous data. These quality control flags could also help maintenance personnel easily identify problem locations where maintenance is needed.

Identifying Suspect or Erroneous Data Values

Error detection capabilities are a critical component of data archiving systems. Even though many ITS deployments have traffic management software or field controllers with basic error detection (16,17,18), additional advanced error detection capabilities may be desirable for data archiving systems. Most data screening techniques used to detect such errors at traffic management centers (TMCs) are based on comparing reported volume, occupancy, and speed values to minimum or maximum threshold values. These minimum or maximum thresholds are typically defined as the lower or upper limit of plausible values. These data screening techniques in place at many TMCs have been criticized as providing only a ". . . minimal examination of credibility" (18).

Several of the data quality procedures developed specifically for planning applications provide guidance for replacing erroneous or suspect data (also known as imputation); however, this is not the recommended practice of the *AASHTO Guidelines for Traffic Data Programs* (19). Data archives may contain suggested replacement values, but these values should be flagged as estimates and not direct measurements. A good example of user-specified data error thresholds is provided in the CD-R data archival software developed at the Washington State Transportation Center (12).

For example, the Texas Transportation Institute and Cambridge Systematics, Inc. analysis of archived operations data in ten different cities used these basic tests for quality control (13):

- maximum volume threshold (e.g., greater than 250 vehicles per lane for 5 minutes);
- maximum occupancy threshold (e.g., greater than 90 percent for 5 minutes);
- maximum speed threshold (e.g., greater than 80 mph for 5 minutes);
- minimum speed threshold (e.g., less than 3 mph);
- inconsistency of traffic data values (volume, occupancy, and speed) within the same data record or with traffic flow theory (e.g., occupancy is less than 3 percent but speed is less than 45 mph; speed equals zero but volume is non-zero; and, occupancy is greater than zero but volume and speed are zero); and
- sequential volume test (e.g., if the same volume is reported for 4 or more consecutive time periods, assume that the detector is malfunctioning).

These very basic tests will identify blatant data errors; however, more advanced tests may be required if a more rigorous quality control process is sought. Advanced quality control can include these tests:

- sequential data checks – identifies rapid fluctuations in data values for consecutive time periods (e.g., speeds typically do not go from 60 mph to 20 mph and back to 60 mph in consecutive 5-minute periods);

Regardless of where quality control is performed, it is important to mark or "flag" data values that have failed quality control or have been modified by quality control processes.

Missing data are inevitable because of the continuous nature of ITS data collection.

- spatial/corridor data checks – identifies inconsistencies between detectors in adjacent lanes or between upstream/downstream detectors (e.g., volume into a link should approximately equal volume out); and
- historical data checks – examines the changes from one year to the next for reasonableness (e.g., high increases in volume or drastic changes in speeds without a corresponding change in traffic volume).

Data quality checks are only the first step in the quality control process. Once suspicious or erroneous data are detected, an action must be taken. Possible actions include simply flagging or marking the data, or entirely replacing the data. Methods for replacing data that fails quality control, as well as for imputing missing data, offer the chance to improve data completeness. Such methods would be based on "good" data from surrounding locations for the same time period as well as using historical data at that same location.

Identifying Missing Values

Several reports note missing data as a common attribute of ITS traffic monitoring data because of the continuous operation of the traffic monitoring equipment (5,12,20). The typical causes of missing data, as well as how the causes affect missing data, are shown in Table 4. The characteristics of missing data may vary considerably depending upon the type of traffic monitoring equipment, field controllers, and central traffic management systems. It is important to not only identify and fix missing data in data archives, but also to evaluate the cause(s) of missing data. This requires analyzing patterns in missing data and working closely with software developers, maintenance personnel, or others that may be able to fix the missing data problem.

The nature and extent of missing ITS traffic monitoring data should be identified and reflected in the design of an ITS data archiving and/or analysis system. Missing data are nearly inevitable; therefore, knowing the characteristics of the

Table 4

Typical Causes and Characteristics of Missing ITS Traffic Monitoring Data

Cause of Missing Data	Characteristics of Missing Data	
	Spatial Attributes	Temporal Attributes
Construction activity that disrupts the traffic monitoring installation	data missing at a single location or several consecutive locations along a corridor	data typically missing for extended period of time (i.e., several months, but depends upon type of construction activity)
Failure of traffic monitoring equipment (could include the inductance loop hardware or the field controller software)	data missing at a single or several isolated locations	data typically missing for short or long periods of time (i.e., several minutes to several weeks)
Disruption of communications between field controllers and central traffic management system	data missing at a single or several isolated locations	data typically missing for short periods of time (i.e., less than several minutes)
Failure of central traffic management system or data archiving system (hardware or software) failure	data missing at all locations (or all locations on a given data server)	data typically missing for short periods of time (i.e., several hours to less than one day)

missing data will help in identifying how best to handle the missing data in aggregation, summarization, or analysis algorithms. Regardless of the methods or algorithms used, it is important that data users be informed of missing data when summary or analysis results are presented. For example, when an average hourly speed is calculated and presented, an additional missing data statistic (e.g., percent complete value, Equation 1) should also be calculated. The percent complete value assists users in determining the reliability of average or summary statistics.

Equation 1

$$\text{Percent Complete Value} = \frac{\text{actual number of records/observations}}{\text{total expected number of records/observations}}$$

As an example of data completeness, consider the following example. If we are using 5 minutes as our analysis interval, we expect to have 288 data values or records per day (e.g., 1,440 minutes per day divided by 5-minute periods equals 288 records) per detector lane or location. The total number of records we expect to see within an area is 288 records per day per detector, or (288 × number of detectors) per day. A detector inventory that lists installed and functional detectors can be used to calculate percent complete values and data completeness. Table 5 illustrates another example of calculating data completeness for an entire system (in this case, we use San Antonio's TransGuide® system).

Because the requirements of various data analyses and applications can vary significantly, a common practice is to simply flag missing data with no edited or replacement values. Data users with specific application needs can then edit or replace missing data values as appropriate to their individual analysis.

In summary, the following are important findings related to handling missing data values in archived operations databases:

It is important to not only identify and fix missing data in data archives, but also to evaluate the cause(s) of missing data.

Table 5

Identification of Missing Data for San Antonio's TransGuide®

Steps in Identifying Missing Data	Calculation for San Antonio TransGuide®
Determine the frequency of observations (e.g., polling cycle) at each location and lane. Use missing data score at this step to identify location-specific missing data problems.	Loop detectors polled every 20 seconds, producing **4,320 possible records** at each lane and location every day.
Determine the number of unique traffic monitoring locations for each computer server.	Two computer servers: Poll Server A and B Server A: 297 unique lane detectors Server B: 230 unique lane detectors
Determine the total possible number of records per computer server. Use missing data score to identify server-specific missing data problems.	Server A = 4,320 records × 297 detectors Server A = 1,283,040 records per day Server B = 4,320 records × 230 detectors Server B = 993,600 records per day
Determine the total possible number of records per day for the entire system. Use missing data score to identify overall missing data problems.	Entire System (Phase One) = Server A + B Entire System = 2,276,640 records per day Missing Data Records = Total Possible Records - Observed Records

- Missing data are inevitable – Because of the continuous nature of the data collection systems, intermittent failures should be expected.
- Identify the nature and extent of missing data for system design – Archived data management system designers should perform basic analyses to determine the nature and extent of missing data. By performing these simple analyses, system designers will be able to minimize the adverse effects of missing data on databases and analysis tools.
- Account for missing data in summary statistics – Summary or aggregated statistics should reflect the amount of missing data by reporting a missing data statistic (percent complete values). Cumulative statistics, such as vehicle-miles of travel (VMT), are most affected by missing data and will likely require data users and system designers to collaborate on how to account for the missing data.

Identifying Inaccurate Data Values

Accuracy is another attribute of data quality that is often a concern for archived data users. In this sense, accuracy refers to the sensor's ability to truly reflect actual traffic conditions (e.g., reported vehicle counts closely approximate actual number of vehicles). Accuracy typically is a concern with archived operations data because the primary data collectors (e.g., TMCs) may have different accuracy requirements than the majority of archived data users. For example, TMCs may only require vehicle speeds to the nearest 5 or 10 mph for congestion or incident detection, whereas simulation model validation or performance monitoring may require archived speed data to the nearest 2 or 3 mph for accurate results.

Additionally, the primary data collectors' accuracy requirements may be for shorter periods of time (i.e., less than 15 minutes for real-time operations and management), whereas archived data users typically have accuracy requirements for much longer periods of time (i.e., one hour to a full year). A small bias or calibration error at the 5-minute level can accumulate significant error in aggregated statistics, such as in average weekday traffic (AWDT) or average annual daily traffic (AADT).

Inaccurate data may not be immediately obvious to database managers or data analysts. For example, it may be difficult to tell whether vehicle counts are consistently 20 percent higher or lower than actual values unless one has prior knowledge about actual vehicle counts. Most studies of sensor accuracy compare measured data to an independent benchmark or "ground truth" value. For example, ground truth in vehicle counts is frequently determined by manually counting vehicles several times (typically from video), until all manual counts fall within a certain error range (typically 2-3 percent error).

This ground truth method has been used in several accuracy assessments of archived operations data (14,21,22). Accuracy assessments have shown varying results. In San Antonio, one detector location had vehicle counts within ±3 percent of ground truth, whereas another detector location had vehicle counts that range from +20 percent to -38 percent of ground truth. Similar findings have been made in Atlanta, Orlando, and New York.

Figure 3 shows typical charts that are used to compare sensor measurements to benchmark or ground truth. In this figure, the benchmark is a permanent vehicle count station maintained by a planning group.

DATA ARCHIVING ARCHITECTURE AND STANDARDS

Defining a system architecture and identifying relevant data standards are important steps in designing a data archive. These are such important steps that FHWA issued a policy in January 2001 that requires all ITS projects to conform to the National ITS Architecture (and regional architecture) and relevant standards.

The National ITS Architecture was revised in 1999 to include an archived data user service (ADUS), which defines a general framework for data archiving. Archived data standards, however, are in the early development stages, so little definitive information is available. This section provides an overview of ADUS and its possible functions and includes some information on archived data standards.

Inaccurate data may not be immediately obvious to database managers or data analysts.

Figure 3

Typical Traffic Volume Comparison Results—ATR to TransGuide® Detector

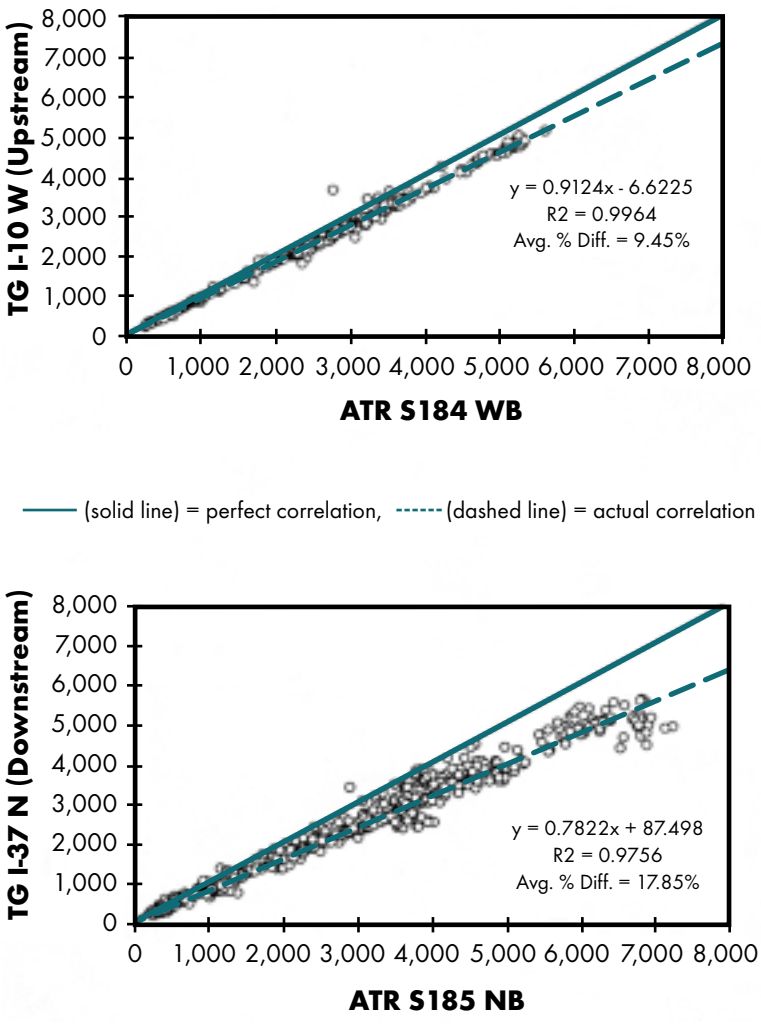


Figure 4

National ITS Architecture, Version 3.0

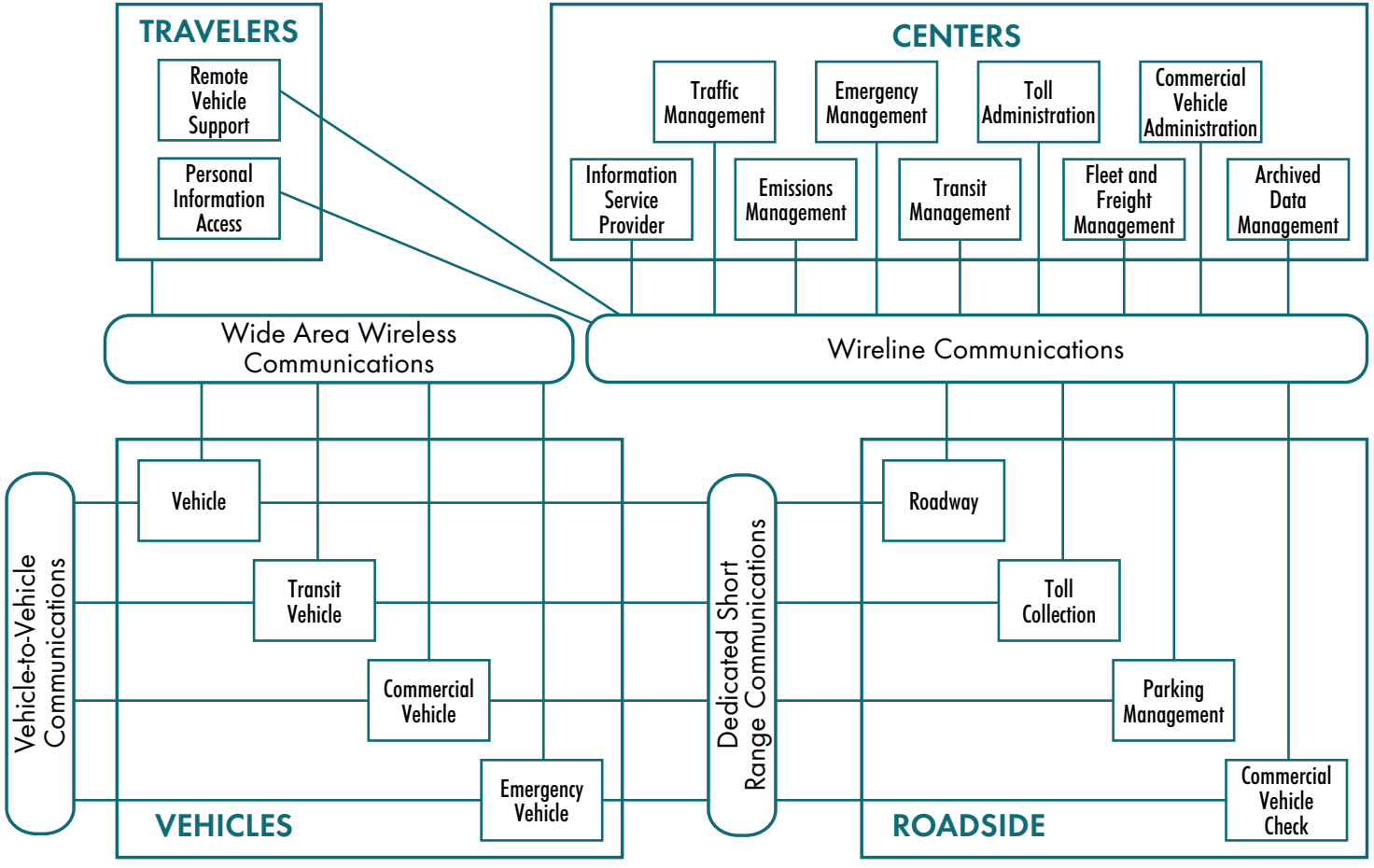
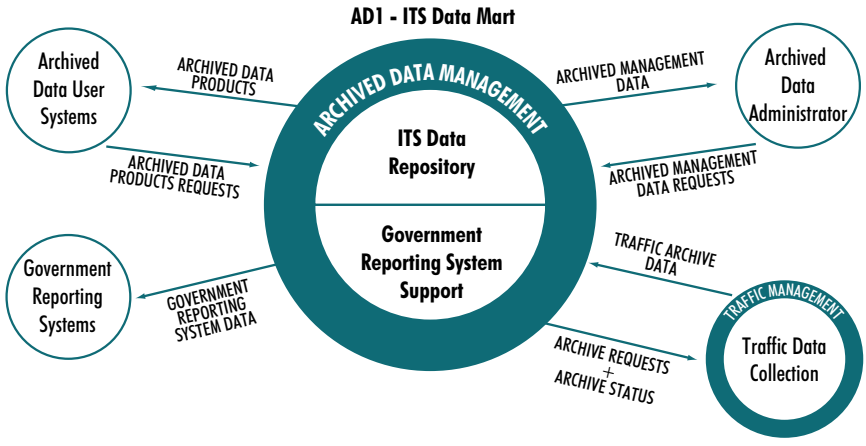


Figure 5

ITS Data Mart Market Package
National ITS Architecture Version 3.0



Any of the following ITS data sources can be the source for an ITS Data Mart. The Traffic Management Subsystem is shown as an example.

Data Sources:

- Commercial Vehicle Administration
- Emergency Management
- Emissions Management
- Information Service Provider
- Parking Management
- Roadway Subsystem
- Toll Administration
- Traffic Management
- Transit Management
- Construction and Maintenance
- Intermodal Freight Depot
- Map Update Provider
- Multimodal Transportation Service Provider
- Other Data Sources
- Weather Service

Archived Data User Service (ADUS)

ADUS was officially incorporated into Version 3.0 of the National ITS Architecture, which was released in December 1999. The important functions or services that ADUS can provide are as follows:

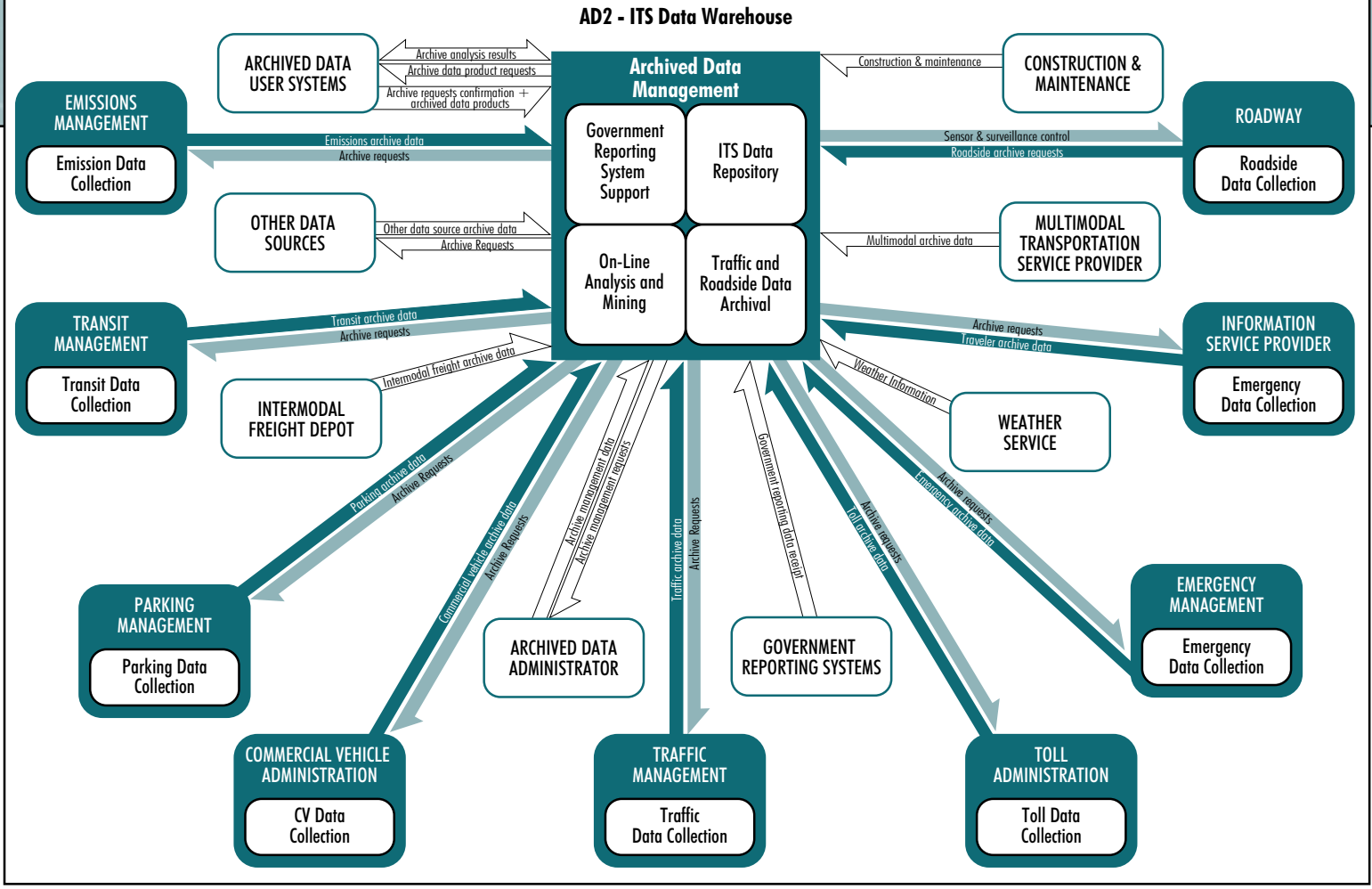
- collects, archives, manages, and distributes data from ITS sources;
- provides proper formatting, quality control, and assigns necessary metadata (i.e., information about the conditions under which data were collected);
- performs data fusion (i.e., the association and joining of data elements from numerous disparate sources); and
- prepares "data products" for input to federal, state, and local data reporting.

User requirements for ITS data archiving were defined by stakeholders in the early stages of developing ADUS. These user requirements are documented in the Architecture and can serve as a useful starting point for defining local user requirements. In Architecture terminology, ADUS describes the **function** that is provided, whereas the physical entity that provides the function is the archived data management subsystem (ADMS). As such, the ADMS is one of 19 currently in the National ITS Architecture (Figure 4).

In the Architecture, market packages are groups of technologies or services that fit real-world transportation problems and needs. Implementing ITS improvements will frequently be a matter of combining individual elements; the market packages defined below are illustrations of ADMS products that could be installed:

Figure 6

ITS Data Warehouse Market Package
National ITS Architecture Version 3.0



- ITS Data Mart (Figure 5) – contains data from a single agency or jurisdiction and for a single mode of transportation. For example, an ITS Data Mart could consist only of loop detector data from a local TxDOT district.

- ITS Data Warehouse (Figure 6) – contains data from multiple agencies, jurisdictions, and transportation modes. For example, an ITS Data Warehouse could consist of loop detector data from TxDOT, environmental data from the weather service, and transit data from the local transit agency.

- ITS Virtual Data Warehouse (Figure 7) – serves as a regional clearinghouse for physically distributed ITS data archives that are locally managed. For example, an information service provider (ISP) may develop an interface that permits data users to access numerous databases in regions that are maintained by different agencies at different locations.

The market packages described above are a collection of one or more equipment packages, which group similar processes together and are the most detailed elements of the physical architecture. The equipment packages for ADMS are:

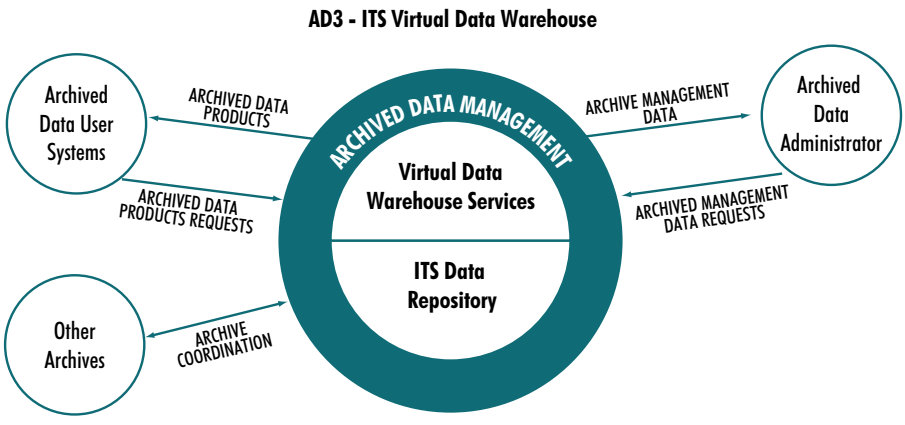
- Government Reporting Systems Support – selects and formats data residing in an ITS archive to facilitate local, state, and federal government data reporting requirements. For example, this package could be used to facilitate reports to FHWA's Highway Performance Monitoring System (HPMS) database.

- ITS Data Repository – collects data and data catalogs from one or more data sources and stores the data in a focused repository that is suited to a particular set of ITS data users. This package is the basic data storage and management function that most relational databases provide.

- On-Line Data Analysis and Mining – provides advanced data analysis, summarization, and mining features that facilitate discovery of information, patterns, and correlations in large data sets. This package provides additional analysis functions that enable typical users to analyze large relational databases.

Figure 7

ITS Virtual Data Warehouse
National ITS Architecture Version 3.0



- Traffic and Roadside Data Archival – collects and archives traffic, roadway, and environmental information for use in off-line planning, research, and analysis. This package enables data from devices outside the ITS domain to be imported into an ITS data archive. For example, traffic volumes from sensors maintained by transportation planners could be imported into an ITS data archive.

- Virtual Data Warehouse Services – provides capabilities to access "in-place" data from geographically dispersed archives and coordinates information exchange with a local data warehouse. This package enables the sharing of data between agencies that maintain their own separate databases. For example, this package could allow you to access and analyze the effects of incidents on traffic speeds, where the incident data are maintained in a state highway patrol database and the traffic speed data are maintained in a DOT data archive.

The U.S. DOT has developed numerous tools (e.g., Turbo Architecture) and training courses that are helpful in developing regional ITS architectures using the National ITS Architecture as a framework. Additional information can be found at http://www.its.dot.gov/arch/arch.htm.

Archived Data Standards

A plethora of ITS data standards activities are ongoing at this time with numerous activities related to data archiving. The standards most relevant to data archiving are those currently being developed by the American Society for Testing and Materials (ASTM) through the E17.54 subcommittee. These ASTM standards are focused on better documenting the source, content, and quality of archived data. Other data standards relate to how data archiving systems communicate with other systems, such as freeway management or traveler information systems. Additional and the most up-to-date information on ITS data standards can be found at http://www.its-standards.net/.

Additional information on ADUS can be found in the following resources:

- Archived Data User Service in the National ITS Architecture (http://www.iteris.com/itsarch);
- *ITS as a Data Resource: Preliminary Requirements for a User Service*, Report No. FHWA-PL-98-031, April 1998 (http://www.fhwa.dot.gov/ohim/its/itspage.htm); and
- "Archived Data User Service (ADUS): An Addendum to the ITS Program Plan," Version 3, September 1998 (http://www.itsdocs.fhwa.dot.gov/jpodocs/repts_pr/41401!.htm).

The ITS Data Registry (http://standards.ieee.org/regauth/its/) is another tool that could be useful in identifying relevant data elements and standards that have already been created and tested. According to the Institute of Electrical and Electronic Engineers (IEEE) web site, the ITS Data Registry web site is a "centralized data dictionary or repository for all ITS data elements and other data concepts that have been formally specified and established for use with the U.S. national ITS domain. Its primary objective is to support the clear-cut interchange and reuse of data and data concepts among the various functional areas of intelligent transportation systems." The ITS Data Registry, along with the ITS standards web site mentioned above, are currently the most comprehensive sources for information about existing data elements and standards.

REFERENCES

1. McDermott, J.M. "Update: Chicago Area Freeway Operations." Illinois Department of Transportation, June 4, 1991.

2. Goolsby, M.E. "Digital Computer Serves Effectively in Traffic System Study." *Texas Transportation Researcher*, Vol. 4, No. 2, College Station, Texas, April 1968.

3. Nihan, N., L.N. Jacobson, J.D. Bender, and G. Davis. Detector Data Validity. Report WA-RD 208.1. Washington State DOT, Washington State Transportation Center, Seattle, Washington, March 1990.

4. "Archived Data User Service (ADUS): An Addendum to the ITS Program Plan." Version 3, September 1998, http://www.itsdocs.fhwa.dot.gov/jpodocs/repts_pr/41401!.htm.

5. Margiotta, R. *ITS as a Data Resource: Preliminary Requirements for a User Service*. Report FHWA-PL-98-031. Federal Highway Administration, Washington, DC, April 1998, http://www.fhwa.dot.gov/ohim/its/itspage.htm.

6. Winick, R. *An Assessment of Architecture Approaches for Data Integration and Archiving: A Review of the Needs and Requirements of Potential Users of an Archived Data User Service for CHART II*. Tech Memo, December 15, 1999.

7. *TranStar Data Warehouse User Data Needs*. TDW-Version 1.0, Prepared by Southwest Research Institute for Texas Department of Transportation, March 13, 2000.

8. *Scalable AZTech™ Data Server Enhancements for Planning and Operations: User Services Requirements Study*. Prepared by Kimley-Horn and Associates, Inc. for Maricopa Association of Governments, Phoenix, Arizona, November 8, 1999.

9. "Forth Worth TransVISION Data Inventory and Needs Assessment: Documentation and Explanation of Survey Results." Tech Memo submitted to Texas Department of Transportation by Texas Transportation Institute, 2001.

10. Turner, S.M. "Improving and Using TxDOT Freeway Operations Data Archives in Austin, Texas." Tech Memo to Texas Department of Transportation, Austin District, July 2001.

11. Chen, C., K.F. Petty, A. Skabardonis, P. Varaiya, and Z. Jia. *Freeway Performance Measurement System: Mining Loop Detector Data*. TRB Preprint 01-2354. Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington DC, January 2001.

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

ASTM standards are focused on better documenting the source, content, and quality of archived data.

13. Lomax, T., R. Margiotta, and S. Turner. *Monitoring Urban Roadways: Using Archived Operations Data for Reliability and Mobility Measurement*. Texas Transportation Institute and Cambridge Systematics, Inc. for Federal Highway Administration Operations Core Business Unit, Draft, May 31, 2001.

14. 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, College Station, Texas, August 1999.

15. Asaka, T. and K. Ozeki, ed. *Handbook of Quality Tools: The Japanese Approach*. Productivity Press, Portland, Oregon, 1990.

16. Chen, L. and A.D. May. Traffic Detector Errors and Diagnostics. *In Transportation Research Record 1132*. Transportation Research Board, Washington, DC, 1987, pp. 82-93.

17. Jacobson, L.N., N.L. Nihan, and J.D. Bender. Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System. *In Transportation Research Record 1287*. Transportation Research Board, Washington, DC, 1990, pp. 151-166.

18. Cleghorn, D., F.L. Hall, and D. Garbuio. Improved Data Screening Techniques for Freeway Traffic Management Systems. *In Transportation Research Record 1320*. Transportation Research Board, Washington, DC, 1991, pp. 17-23.

19. *AASHTO Guidelines for Traffic Data Programs*. American Association of State Highway and Transportation Officials, Washington, DC, 1992.

20. Wright, T., P.S. Hu, J. Young, and A. Lu. *Variability in Traffic Monitoring Data: Final Summary Report*. Oak Ridge National Laboratory, University of Tennessee-Knoxville, Oak Ridge Tennessee, August 1997.

21. Schmoyer, R., P. Hu, and R. Goeltz. *Statistical Data Filtering and Aggregation to Hour Totals of ITS Thirty-Second and Five-Minute Vehicle Counts*. TRB Preprint 01-0873. Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC, January 2001.

22. Grant, C., B. Gillis, and R. Guensler. Collection of Vehicle Activity Data by Video Detection for Use in Transportation Planning. In *ITS Journal*, Volume 5, Number 4, 1999.