

Implementing the ITS Archived Data User Service in Portland, Oregon

Robert L. Bertini, Andrew M. Byrd, and Thareth Yin

Abstract—This paper describes the current status of the ITS archived data user service for the Portland, Oregon metropolitan area. Included is a discussion of the definitions, benefits and history of ITS data archiving. Case studies of three current ITS data archiving projects are also presented. The requirements and characteristics of a long term data archive for ITS data are summarized and the implementation plan for this service is presented. Design features of preexisting data archives and guiding principles are described. Design choices made in response to these principles are discussed. The final database design is then outlined.

I. INTRODUCTION

DATA ARCHIVING for intelligent transportation systems (ITS) refers to the systematic retention and reuse of operational ITS data. Though the original and primary purpose of generating these data is often for real-time management of transportation systems, archiving these otherwise-discarded data offers a rich source of information to various entities and agencies with a heightening need to evaluate system performance and characteristics [1]. Once retained, the vast amount of ITS-generated data can be used in transportation planning, administration and research by key stakeholders including metropolitan planning organizations (MPOs), state transportation planners, traffic management operators, transit operators and transportation researchers [2].

ITS data may be used to supplement or replace conventional data sources that are both time consuming and costly to implement, and the high level of spatial and temporal detail found in ITS-generated data lends itself to novel methods of analysis [2]. Cost effectiveness of the existing data collection infrastructure is maximized and sampling bias is reduced as a result of the continuous collection of data, and a better understanding of variability in system performance and response can be achieved [2].

Revised manuscript received July 18, 2004. This work is supported by the National Science Foundation, the Oregon Department of Transportation and Portland State University. R.L. Bertini is with the Department of Civil & Environmental Engineering at Portland State University, Portland, OR 97207 USA (phone: 503.725.4249 fax: 503.725.5950 email: bertini@pdx.edu). A. Byrd is a student in Computer Science (e-mail: abyrd@cs.pdx.edu) and T. Yin is a student in Civil Engineering at Portland State University (email: thareth@thareth.com).

Use of archived data can also assist in diagnosing defective ITS components and aid in maintenance programming.

To foster the retention and use of ITS-generated data in these areas, the United States Department of Transportation (USDOT) revised the National ITS Architecture in 1999 to include the Archived Data User Service (ADUS), envisioning “the unambiguous interchange and reuse of data and information throughout all functional areas.” ADUS requires that data from ITS systems be collected and archived for historical, secondary and non real-time uses, and that these data be made readily available to users [3]. The addition of ADUS to the ITS architecture in 1997 has substantially increased national interest and guided the implementation of archiving projects, and in 2000, a 5-year ADUS Program Plan was implemented, beginning major Federal activities to support ADUS deployment.

II. CHARACTERISTICS OF SUCCESSFUL ARCHIVES

Though there has been a recent surge of interest in the reuse of archived ITS data, the concept is not new. Projects involving sensor and detector data were begun as early as thirty-five years ago, long before they were associated with the terms ITS or data archiving. Starting in 1968, loop detector data from the Gulf Freeway in Houston were being collected by the Texas Highway Department and the Texas Transportation Institute (TTI) for investigations on the effects of incidents and merging on level of service. In the early 1970s, congestion information was derived from aggregated loop detector data by the Illinois Department of Transportation [4]. Researchers and planners at the Washington State DOT and the University of Washington have been using archived loop detector data from Seattle’s freeway system since 1981, and data from the Gardiner Expressway in Toronto, Ontario and from Interstate 880 in Hayward, California have been used extensively in the study of capacity and traffic flow theory.

This early work with data archiving uncovered numerous potential problems in data archiving systems [4], including:

- Data volume and availability was limited by the storage technology in use. Magnetic tape cartridges and proprietary data formats were specific concerns.
- Operations centers were burdened with the task of archiving large amounts of data which were not

perceived as useful to their organization.

- The environment for processing and analyzing these large quantities of data was not user-friendly.
- Poor data quality due to missing or erroneous data.

In contrast to these problem areas, several characteristics were found in successful data archiving programs, leading to the following guidelines [1]:

- Determine the workgroup(s) and/or agency(ies) that should hold the primary responsibility and accountability for operating and maintaining the data archive.
- Begin with a simple prototype focusing on a single source of data. Experience in other industries has shown that efforts to begin with large, complex data warehouses often struggle for many years and user studies show that volume and speed information are in highest demand. It follows that loop detector data should be the first element in a data archiving system, with increases in complexity occurring after the framework is in place.
- Provide users easy access to data through the Internet or by distribution on optical storage media. Web-based systems were found to be the most effective means of distributing data, and it is especially desirable to utilize a point-and-click query interface which can present summary information graphically.
- Save the original raw data and make aggregate data available to users. Sample times of 20 sec to 1 min are typical for raw loop detector data, and 5-min aggregation may be adequate for most users. Permanent off-line storage should be developed to accommodate those needing detailed data.
- Quality control methods should flag, remove or impute suspect or erroneous data. Addressing this issue can help avoid propagation of error into end-user analyses.
- Perhaps most importantly, provide adequate documentation of the archive and the data collection infrastructure. It is critical that users working outside the implementing agency be able to understand the nature of the archived data and the tools through which they may access it. Documentation should include a presentation of the data schema, including database and table definitions and relationship specifications. Metadata should be included which describe the aggregation algorithms and quality control procedures in use, and any modifications to raw data should be noted. Information on the data collection infrastructure should include the type and location of each detector. Location should be provided in a format that can be universally understood (e.g., latitude/longitude rather than a route/milepost).

III. CASE STUDIES

This section describes three existing data archiving systems which illustrate practical design principles.

A. San Antonio, Texas

The San Antonio district of the Texas Department of Transportation (TxDOT) designed the TransGuide Advanced Traffic Management System (ATMS) to monitor and report freeway performance. TransGuide integrates data from loop detectors, overhead acoustic devices and transponders in volunteer probe vehicles that use an automatic vehicle identification (AVI) system [5]. The AVI system covers 155 km (98 miles) of freeways and limited access highways [4]. The loop detectors are the main source of traffic data. Completed since 1995, Phase I of the TransGuide Project has double-loop detectors covering 42 km (26 miles) of freeway encircling San Antonio [4]. The detectors are spaced at approximately 0.8 km (0.5 miles) and there are detectors on all entrance and exit ramps [6]. The dual loop detectors (speed traps) on the main freeway lanes are spaced about 10 m (30 ft) apart [6].

Every 20 sec, the TransGuide computer servers poll or retrieve aggregated data in a sequential pattern from the local controller units (LCUs) in the field [6]. The following data are collected: average spot speed (mph), volume (number of vehicles) and lane occupancy (percent of time loop is occupied) [6]. The most recent month's raw 20-sec data are posted to a file transfer protocol (FTP) server at (<ftp://www.transguide.dot.state.tx.us/lanedata/>) and are available to anyone with Internet access. The files contain a date stamp, a location code and corresponding speed, volume and occupancy measurements.

Researchers from TransLink at Texas Transportation Institute (TTI) interested in using the data developed a prototype ITS data management system called DataLink to archive and mine the TransGuide data. The concept originated from the desire to ease the access and facilitate analysis of the large amounts of data (approximately 120 megabytes per day) being collected by the TransGuide operation center. DataLink was developed with these features and/or functions in mind: [6]

- Store, access, and analyze large amounts of data.
- Access database from remote locations without burdensome or costly software requirements.
- Intuitive graphical user interface, no programming or query language required.
- Summarize raw data as well as calculating performance measures.
- Output results in various tabular and graphic formats.

The DataLink system warehouses 5-min. aggregated loop detector data in a 40+ GB relational database. System architecture consists of five major components including the Oracle database management system (DBMS), common gateway interface (CGI), gnu plot graphics software, e-mail service and apache web server [6].

DataLink has a point-and-click interface. It has the ability to select query variables such as date, time period,

data aggregation, and roadway facility of interest. Query output is flexible and can be viewed in tables in the browser itself, comma-separated values through e-mail, and 2-D and 3-D graphics viewable using Adobe Acrobat Reader.

B. California PeMS

PeMS is a freeway performance measurement and data archiving system developed by the Operations Division of the California Department of Transportation (Caltrans) and the University of California at Berkeley. The ITS data management tool collects raw freeway data from several Caltrans districts. Archived data are made available through the Internet at <http://transacct.eecs.berkeley.edu> for anyone with access privileges. PeMS is used for monitoring freeway speeds, estimated travel times and volumes.

There are 23,138 loop detectors measuring 30-sec vehicle count and occupancy from California's freeways [7]. Detector data are transmitted via phone lines to a central computer and then sent to PeMS via Caltrans' Wide Area Network (WAN). Raw data are aggregated into 5-min intervals and loaded into PeMS' data warehouse consisting of an on-line Oracle database with two terabytes of data, which grows at two gigabytes per day. PeMS was developed with open architecture with options for adding capabilities as the need arises and technology improves. [7]

Unlike the double loops in San Antonio, California's are single loop, which do not directly measure speed. Speed can be deduced if average vehicle lengths are known. Using algorithms, PeMS estimates both vehicle length and speed. PeMS also uses algorithms based on empirical models and fitted to historical data to generate standard freeway performance measures used to track trends in congestion, locate bottlenecks and monitor service quality. With regard to planning, the performance measures can point to priorities in funding projects and provide information for drivers to make informed decisions about travel behavior and route decisions.

PeMS also diagnoses detector health to deal with the issue of data quality. Reports on detector health can focus maintenance resources on problematic detectors. An algorithm diagnoses each 5-min sample, comparing it with the previous day's data and neighboring measurements. [7]

The data itself is used to detect bad data. The detection and imputation algorithms simplify the design and operation of higher-level algorithms by completing a grid of values using original data. The imputation algorithm models a linear relationship between each pair of neighbors fitting the parameters to historical data. This method was found to be better and more robust than alternative techniques [7].

PeMS has been an ongoing project since 1998 that was first motivated by legislation requiring Caltrans to

monitor the performance of their transportation network [1]. Instead of manually re-collecting data, they use existing data from the network of existing freeway loop detectors.

C. Seattle, Washington

Washington State Department of Transportation (WSDOT) and the Washington State Transportation Center (TRAC) have developed a CD-based data archiving system for the central Puget Sound area. Data are collected from loop detectors that collect volume and occupancy data and are spaced approximately 0.8 km (0.5 mile) apart. An algorithm is used to estimate speeds [4]. The data are collected every 20-secs from field controllers, summarized into 5-min intervals and archived to CD. It takes four CDs to store one year's worth of the 5-min aggregated data. The data are filtered through a quality control process before archiving, and a record is kept of any data that have failed quality control. Each CD contains data extraction and summary tools [4].

Since 1981, WSDOT has been performing freeway detector data archiving. Early efforts were difficult because of the expense of data storage and pre-Internet data transfer. Over time, the system has made incremental improvements as the demand for archived data has increased. The archived data have since been used for: [8]

- Testing and evaluating operational improvements (e.g. ramp metering or HOV lane).
 - Freeway performance monitoring.
 - Pavement design.
 - Freight performance analysis.
- Software tools and algorithms are used to analyze the usage and performance of the freeway system. The CDR (Compact disc Data Retrieval) tool is a catalog of data processing options available for freeway network analysis. The evaluation tool set uses archived CD data to summarize raw data, compute performance measures, present analyzed data in text and graphical formats and reformat raw data for use by CDR Analyst (CDR Auto). [8]

IV. THE PORTLAND REGIONAL ADUS

Although there can be many beneficiaries, data archiving has been historically and continues to be the domain of a single user group (e.g., researchers and planners) [1]. The mostly informal arrangements for data archiving with researchers and planners is due to the inaccessibility and difficulty in using the large data sets. In the Portland, Oregon, metropolitan area (see Fig. 1), regional agencies across two states, including cities, counties, MPOs, transit operators, a port, a university and state DOTs cooperate by means of TransPort, the regional ITS coordinating organization. The region's ITS Architecture has been developed under the auspices of TransPort.

It can be difficult to determine who should bear the

responsibility and cost of maintaining an ITS data archive because the organizations that produce the data are frequently different from those who benefit from its retention. In cooperation with the Oregon Department of

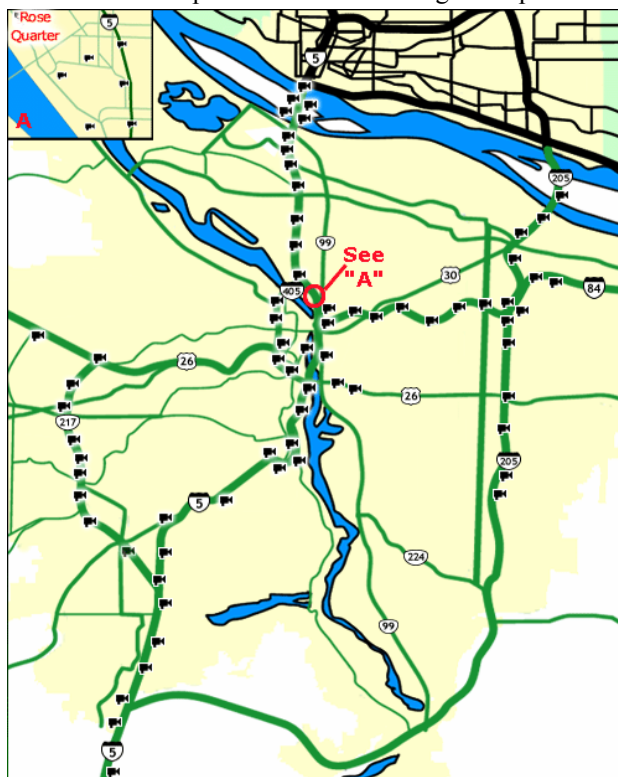


Fig. 1. Portland Metropolitan Area.

Transportation (ODOT), TriMet and the City of Portland, Portland State University (PSU) is a major user of regional ITS data for transportation research. Other recent uses of ITS data by ADUS stakeholders such as Metro, the Portland regional planning organization, have been encouraged and facilitated by PSU researchers already familiar with manipulating this type of data.

As a result, PSU is in an ideal position to host transportation data, and its Intelligent Transportation Systems Laboratory (www.its.pdx.edu) has been designated as the regional archiving site for ITS data from Portland and adjacent areas of southwestern Washington State. The design and implementation of this archive are being carried out in accordance with the functional requirements for ADUS set forth in the National ITS Architecture, while also taking into account suggestions found in the *Guidelines for Developing ITS Data Archiving Systems* [1].

The following sections describe how PSU has fulfilled the functional specifications for an ITS data archive presented in the National ITS Architecture ADUS addendum [10].

V. APPLICABLE TECHNOLOGIES

The Portland regional ITS data archive makes full use

of available information technologies as recommended in section 1.5 of the ADUS addendum. Data are stored and accessed using a PostgreSQL relational database management system (RDBMS) running on a Linux platform, and new generation hardware has been employed to minimize user wait time for dynamic queries. The RDBMS stores data physically on a redundant array of independent disks (RAID) providing both high-speed access and increased reliability through redundancy in the event of hardware failure. The web-based user interface is provided by the Apache web server, with database connectivity and additional processing by the PHP hypertext preprocessor module. Detector locations within the database have been geocoded for interoperability with geographic information systems (GIS).

This archive implements a data warehousing strategy in that it retains large amounts of raw operational data for analysis and decision making processes, and in that these data are stored independently of their operational sources, allowing the execution of time-consuming queries with no impact on critical operations uses.

VI. DATA SOURCES AND ARCHIVE STRUCTURE

This section corresponds to the National ITS Architecture ADUS addendum [10], section 1.4.1.1 and 1.4.1.2. Several ITS data sources are available in the Portland metropolitan region, primarily originating at the ODOT, Washington Department of Transportation, and TriMet, the regional transit agency. The structure of the Portland ADUS is shown in Fig. 2

The ODOT Region 1 transportation management operation center (TMOC) gathers data from 75 video cameras and 436 inductive loop detectors which comprise the Portland area's ATMS. These data are used in the day to day management of the transportation system to identify congested areas and incidents and to dispatch incident response or emergency vehicles to the appropriate locations. Drivers can be alerted to problems and rerouted using variable message signs and television and Internet-based traveler information systems (www.tripcheck.com).

The vehicle count passing each loop detector and the average speed of these vehicles is reported every 20 sec, along with the occupancy, or percentage of the sample period when a vehicle was over the detector. ODOT currently archives 15-min aggregate data which are adequate for operations purposes but unsuitable for research, planning, and simulation. The raw 20-sec data are retained for a short time prior to aggregation, and an arrangement exists which now allows their transmission to PSU through an existing fiber optic connection between the two institutions.

TriMet gathers extensive data regarding the ridership, location, speed, and scheduling of its fleet of over 650 transit vehicles. In order to facilitate research on and planning of the region's transit system, it will be desirable to integrate this rich data source into the regional archive.

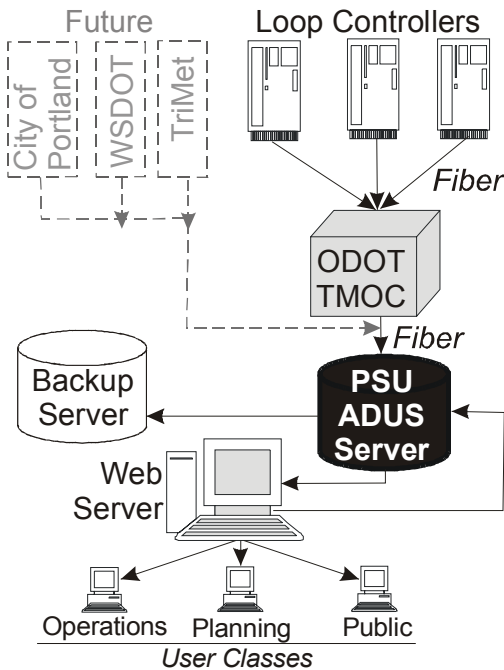


Fig. 2. Portland Metropolitan Area ADUS Data Flows.

However, in keeping with the suggestion that data archives begin operations with a single data source capable of providing highway volume and speed information [1], the inclusion of transit data will begin now that the basic archive system is in place and is operating effectively. Freeway loop detector data from the adjacent region of Vancouver, Washington, and data from arterials in the City of Portland, and weather data will also be included.

The inclusion of multiple data sources in the archive will necessitate a choice between a centralized or decentralized archive structure. At present a centralized approach is anticipated, where data from each individual ITS source is transferred to the central repository at PSU for storage and analysis. This approach allows data users to request, compare, and jointly analyze data from disparate sources through a single portal.

VII. DATA PROCESSING

This section corresponds to the National ITS Architecture ADUS addendum [10], section 1.4.1.3.1, which requires the fulfillment of five distinct processing functions during the receipt of incoming data from other ITS subsystems. First, the archive must store the data in the same format as it is received from its source. Loop detector data are exported from the relational database in use at the TMOC and transmitted directly to PSU in an equivalent tabular form. No loss in time resolution or data accuracy occurs, and the volume, speed, and occupancy data from each detector are inserted into a similar relational database table in the ADUS. Only fields corresponding to operations

center-specific functions and not representing direct loop detector observations are ignored.

Second, the archive must accommodate levels of aggregation appropriate to the type of data being recorded. The PSU regional archive provides a wide variety of aggregate and processed views of the underlying high resolution data, including both simple summaries such as vehicle miles traveled (VMT) or average daily traffic (ADT) and more advanced performance measures requiring additional manipulation. Several static aggregates are prepared automatically for each day's detector data and saved independently of the raw data, while others are computed dynamically, allowing a focus on specific facilities or time frames of interest to the user.

Third, raw data must be sampled and stored online, offline, or both. In our case raw data are retained indefinitely at full resolution. No additional downsampling occurs below the 20-sec intervals used by the detectors themselves. Incoming data are stored online in the database server itself, from which tables are exported daily to a compressed format small enough to be archived offline at yearly intervals on optical (DVD) media.

Fourth, quality control procedures must be applied which identify and mark suspect or erroneous data. Currently, defective loop detectors frequently report values to the TMOC despite equipment failure. Valid loop detector data exhibit daily cycles, limited range, some level of smoothness, and other patterns which make them relatively easy to distinguish from erroneous data. However, the inclusion of these meaningless values in the calculation of aggregates and performance measures would result in the propagation of error into results where it might be difficult or impossible to detect. This problem can be avoided by analyzing the raw loop data and flagging those values that appear to be incorrect. Highly effective techniques for spotting malfunctioning detectors have been documented by researchers working on the California PeMS system [7], and a similar system is used in Portland's archive.

Fifth, distinctions must be maintained between raw, edited, aggregated, and transformed data. Care has been taken in the database design to ensure that data of one type will never be stored in the same database table with any other type, and the level of aggregation is clearly defined and documented for each table in the database.

According to the National ITS Architecture, these five functions may be split apart and performed by different agencies or personnel. However, in keeping with the desire to develop a centralized repository all functions are performed locally by the PSU archive system.

VIII. DATA STORAGE

This section corresponds to the National ITS Architecture ADUS addendum [10], section 1.4.1.3.2, which requires certain considerations in the online and offline storage of raw and processed data by an ADUS.

The ITS Architecture requires that the original, unaltered data be retained in a master archive for some period of time, unaffected by subsequent aggregation and processing. Furthermore, all user-defined transformations and processing must be performed on a working copy of the raw data rather than within the master archive. In the Portland regional archive, two complete copies of the raw data set are maintained; one resides on the archive's main relational database server and serves as the working copy from which data sets fulfilling user requests are retrieved, while the other is stored off-site in a compressed format and serves as the master copy.

Second, metadata must be provided which allow users to understand the data collection conditions, data quality, and any transformations that the data have undergone. Quality control, aggregation, and reduction procedures should be noted to promote careful data usage. Because the archive is initially operating with a single data source polled under uniform conditions, the majority of relevant metadata are contained in the system documentation rather than in the database tables themselves. Notable exceptions include error detection algorithm results and data representing the weather conditions under which the data were collected. Temperature, precipitation, and wind information are retrieved once per hour from the National Oceanic and Atmospheric Administration (NOAA) and stored in a table independent of the raw loop data. Algorithms are run once daily to locate faulty detectors, and suspect locations are noted in another table, allowing warnings to end users of possible inaccuracies. Aggregation is documented and the aggregation level is clearly indicated with each data set returned to the user.

Third, a standardized location referencing system is used throughout the ADUS to allow the linking of spatially related elements. To this end, the position of each detector is stored as a latitude-longitude pair. This allows the comparison and integration of archived loop detector data with the extensive geographic information system (GIS) databases in use by Portland area researchers and planners.

IX. DATA RETRIEVAL

This section corresponds to the National ITS Architecture ADUS addendum [10], section 1.4.1.3.3, which requires the Data Retrieval function in an ADUS—the interface between the data repository and its end users. As suggested in *Guidelines for Developing ITS Data Archiving Systems* [1], user access to the data archive is through a web-based interface providing point-and-click access to both raw data sets and a wide range of common summary data and performance measures. Large tables may be downloaded in common formats for analysis on the user's own machine, while smaller sets of summary data are presented as both inline tables and plots. A map displaying speed or occupancy data for the entire region is available, and contour plots providing an overview of these data for an entire day on an entire facility are produced. Users can

limit queries to a specific time frame or facility of interest. Access to advanced features and processor-intensive queries are limited to registered users with accounts, which are provided to interested parties with legitimate uses for transportation data.

X. DATA SECURITY

Though not directly mentioned in the ADUS requirements document, data security is also regarded as an important component of our regional data archive. As mentioned above, the working copy of the database maintained on the primary server is replicated in a compressed format at a remote site. These daily backups will ensure that if one copy of the database is lost due to equipment failure or fire, the archive service can be rapidly returned to operation with no significant loss of data. Both the primary database server and the backup storage are located in climate controlled machine rooms with uninterruptible power supplies (UPS) and generator backup power, preventing data loss or gaps in data availability due to power outages. The working copy of the database is stored on a RAID device, providing error detection, redundancy, and the ability to rebuild missing data upon device failures. Finally, hardware maintenance and security updates are provided for all computer systems by experienced systems administration staff.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the National Science Foundation, ODOT, TriMet and the City of Portland as well as the TransPort organization for their ongoing support and guidance in the development of the Portland ADUS.

REFERENCES

- [1] S. Turner. *Guidelines for Developing ITS Data Archiving Systems*. Report 2127-3. FHWA, U.S. Department of Transportation, Texas Department of Transportation, and Texas Transportation Institute, 2001.
- [2] ADUS Program. *ITS Data Archiving: Five-Year Program Description*. USDOT, March 2000.
- [3] ADUS Program. *Archived Data User Service (ADUS): An Addendum to the ITS Program Plan*. USDOT, September 1998.
- [4] S. Turner. *ITS Data Archiving: Case Study Analysis of San Antonio TransGuide Data*. Texas Transportation Institute, August 1999.
- [5] M. Carter. *Metropolitan Model Deployment Initiative San Antonio Evaluation Report*. Science Applications International Corporation, May 2000.
- [6] S. Turner, R. Brydia, J. Liu and W. Eisele. *ITS Data Management System: Year One Activities*. TransLink® Research Program: Research Report 1752-2 in Research Study No. 0-1752. Texas Transportation Institute, August 1997.
- [7] C. Chen. *Freeway Performance Measurement System (PeMS)*. California PATH Research Report UCB-ITS-PRR-2003-22. Institute of Transportation Studies, University of California, Berkeley, 2003.
- [8] J. Ishimaru and M. Hallenbeck. *FLOW Evaluation Design Technical Report*. Washington State Transportation Center, May 1999.
- [9] National ITS Architecture ADUS Addendum