IntelliDrive℠ Data Capture and Management Program:

Transforming the Federal Role

prepared for the
Intelligent Transportation Systems Joint Program Office
Research and Innovative Technology Administration (RITA)
Federal Highway Administration (FHWA)
Federal Transit Administration (FTA)
Federal Motor Carrier Safety Administration (FMCSA)

May 2010

U.S. Department of Transportation

Introduction

The Data Capture and Management Program seeks transformational change in the capture and management of data from surface transportation systems data. In particular, the program seeks to exploit potential opportunities in flexible and dynamic data communication between vehicles, infrastructure-based technologies, and mobile devices (i.e., transition from *passive* to *active* data capture). In addition the program wishes to explore transformational change enabled by providing travelers and decision-makers with a rich set of integrated data obtained in concert from the full complement of fixed and mobile entities in the transportation system (i.e., transition from *single-source* to *multi-source* data capture and management). In order to achieve these objectives, however, the program must also transform the mechanisms that allow researchers, public sector agencies, the private sector, and other interested parties controlled access to utilize and share data resources. The program envisions a strong federal role in influencing and facilitating enhanced data capture and management practices. This includes identifying and developing mechanisms to achieve program goals, in addition to fostering deployments and implementations consistent with the program's vision.

The Data Capture and Management Program, following the tenet "collect once, preserve, use many times," supports the development of the broadest possible collection of applications and fully leverages federal research investment. This entails a move beyond traditional "collect, use once and neglect" data capture and management. For the Data Capture and Management Program, this implies reconsidering all aspects of how public sector agencies (including the federal government) procures, acquires, captures, stores, manages, and shares data.

Many of the key issues are non-technical, and to a great extent, involve a re-casting of traditional relationships between the federal government and stakeholders in the public and private sector that collect and utilize data. Other issues are more technical in nature, and here the federal government can take a leadership role by prototyping innovative data capture and management systems that successfully deal with these technical challenges.

Background

The United States Department of Transportation (USDOT) Intelligent Transportation Systems Joint Program Office (ITS JPO) is engaged in assessing the potential for systematic and dynamic data capture from vehicles, travelers and the transportation system infrastructure to enhance current operational practices and transform future surface transportation systems management. One foundational element in this USDOT engagement is the Data Capture and Management Program. Program objectives include[1]:

- Enable systematic data capture from multiple sources, e.g., vehicles, mobile devices, and infrastructure;
- Develop data environments that enable the integration of data from multiple sources for use in transportation management and performance measurement;
- Reduce costs of data management and proactively address technical and institutional barriers that are associated with the capture, management, and sharing of data.

[1] Data Capture and Management Program Vision: Objectives, Core Concepts and Projected Outcomes, Version 1.5, April 2010

The Data Capture and Management Program plays a key role supporting other ITS JPO initiatives identified in strategic plans in the areas of Safety, Mobility, and Environment. Many of these initiatives will require systematic capture and management of data over time to realize their objectives. The cross-cutting Data Capture and Management Program is chartered to coordinate across these initiatives to identify comprehensive data needs. Further, the Data Capture and Management Program is jointly responsible for designing laboratory experiments and field tests that meet these identified data needs in the most cost-effective way. Because of a confluence of web-based technologies and asynchronous collaboration methods, it is increasingly possible to structure and share base research data (and its details) with multiple researchers, not just the tools/analyses (usually codified in academic papers or technical reports) resulting from the application of these data. Data collected in these experiments and tests will be systematically structured and documented in one or more *data environments*, a core concept in the Data Capture and Management Program vision. A data environment is defined as[2]:

- a well-organized collection of data of specific type and quality,
- captured and stored at regular intervals from one or more sources,
- systematically shared in support of one or more applications.

The resulting well-documented, distributed data resources will allow data captured from diverse sources to be integrated, shared, and leveraged by a broad range of researchers, private sector partners, and system operators. Program objectives, core concepts and a high-level program activity plan are detailed in the Data Capture and Management Program Vision document [1].

Without a cross-cutting data capture program, the result will be *ad hoc* data capture and management by each application area. This traditional approach results in redundancies and increased expenses in data collection activities, supporting only limited applications by analysts working with individual data sets targeting a single application.

Purpose

The purpose of this document is to establish organizing principles that will guide the federal role in the Data Capture and Management Program. The nature of this federal role will define the rules of engagement for data environments developed in the Data Capture and Management Program. The way data environments are organized and shared will influence mobility applications or other applications developed using these data, and have important implications for IntelliDrive[SM] policy development. A set of rules for data collected under the Data Capture Program must also clearly explain what participation implies – to all stakeholders, including public sector agencies, private sector partners, and academic institutions.

The document begins with a summarized assessment of the current state-of-the-art in data capture and management. This assessment is based on existing transportation data warehousing efforts from various transportation modes. Three desired transformations from current practice are identified, including a transition from single-source, single-mode to multi-source, multi-modal data environments, from passive to active data capture, and from archival to real-time data provision. Next, the document addresses key issues related to achieving these program objectives and discusses a proposed federal role and activity to successfully deal with these key issues. This document concludes with a set of guiding principles and a set of next steps for the program.

[2] Data Capture and Management Program Vision: Objectives, Core Concepts and Projected Outcomes, Version 1.5, April 2010

Characterizing the State of the Practice

Current technologies and archiving platforms already provide researchers and system operators with organized and sizeable data environments. Data warehousing efforts in surface transportation come from various sources. Some of these data warehousing efforts are federally funded programs. Some efforts are state or regionally-supported archives, while others are self-sustaining research communities.

Assessment of Current Practice

Current data environments can be characterized as being highly dependent on passive infrastructure-based sensors. Furthermore, many of these data resources are single-source. For example, there may be separate databases for freeway loop data, arterial occupancy data, and periodic "floating car" travel time runs. In addition, they are often limited to a single mode. For example, personal vehicle, transit and freight data are often gathered by different data captured efforts that are often not integrated with other modes. Traveler data are limited to behavioral data collected from small samples of infrequently conducted surveys. Speed data from a small sample of probe vehicles are obtained from one of several competing technologies. Speed and traffic count data are available from infrastructure-based sensors deployed on selected high-volume freeway segments and some key arterial locations. The capability of a data environment merging current sources to support mobility and other applications remains a relevant research question. Transit data in particular, are widely collected but not well integrated with operational data.

Most current data environments are archival in nature. Data captured and managed in the current data environments tend to be collected over time, assessed for quality and potentially aggregated at some intermediate point, and then at a later date (days, months or even years later) made available to researchers or other interested users.

Desired transformations in data capture and management practices facilitated in the federal Data Capture and Management Program include:

Transformation #1: Move from single source to multi-source and multi-modal data environments

Currently, most data environments are *single-source*, that is, they capture one type of transportation data, such as vehicle crashes, or roadway inventory, or travel times. However these data sets are seldom integrated with other data sets that are pertinent to transportation management applications development and evaluation. Findings from such single-source analyses may benefit from considering findings from other studies in order to reflect a more comprehensive view of field conditions. For example, a review of roadway facilities' delays might benefit from a concurrent consideration of traffic counts, crash history, and weather conditions.

In addition, each data type is usually collected by a different agency and datasets are not necessarily compatible. As an example, the Highway Performance Measure Monitoring System (HPMS) data[3] includes data on the extent, condition, performance, use, and operating characteristics of the Nation's highways. The data are collected by the 50 states and other

[3] http://www.fhwa.dot.gov/policy/ohpi/hpms/index.cfm

jurisdictions and are reported to FHWA. However, the same type of data is made available at each of the state's websites in very different formats, making it difficult to merge datasets together.

One of the Data Capture and Management goals is to transition to *multi-source, multi-modal* environments, where data from diverse sources can be considered concurrently. For example, traveler information, private vehicle, transit vehicle, heavy vehicle, infrastructure, weather and parking data in such an environment might be combined into one data warehousing effort, where users can obtain a coherent and concurrent view of the transportation system. The Data Capture and Management Program will place particular emphasis on the integration of freight and transit data. In the case of transit data, although transit operators collect robust real-time and archived transit data, these data are not generally available to other agencies and researchers. Freight data capture efforts may also capture information on personal vehicle data, such as travel times and speeds but the information is gathered for the purpose of informing the freight industry stakeholders. An example of such an effort is the Cross-Town Improvement Project (C-TIP). However, these data capture efforts gather data that could be valuable for other uses.

Transit agencies currently collect and archive robust data on transit infrastructure, vehicles and ridership. However, there is a lack of integrated and shared transit data. Some transit agencies such as TRIMET (Tri-County Metropolitan Transportation District of Oregon) and WMATA (Washington Metropolitan Area Transit Agency) do provide some publicly available transit data on their public websites. These data include infrastructure information, such as transit lines and stops, and real-time scheduling. However archived data is not available for users outside those agencies. Coordination and integration of transit data into a multi-modal data environment is a key objective of the program.

Integrating parking data is another goal of the program. Parking-related data have started to be collected, archived and disseminated to help with traveler decision making. For example, the University of California Los Angeles (UCLA) Medical Center parking garage is implementing an occupancy monitoring system where users can get information on occupancy and space availability through a data repository web interface that can be sent to users' cell phones[4,5]. These data environments efforts can be integrated with other sources as part of the Data Capture and Management Program.

Having multi-source and multi-modal data environments could enhance performance measures and better inform decision makers, travelers and system managers. For example, this program could provide the ability to obtain a true performance measurement of a broad system in a way that could not be done before.

Currently, most data environments that provide multi-source data types are collected over time and available at a later date. These types of data environments tend to be large survey efforts such as the Census Transportation Planning Package (CTPP)[6] that provide various types of information based on the U.S. decennial census that is collected over a significant period of time.

---

[4] Robert Repas, "Wireless Sensor Network Aids Travelers in Parking Their Cars and Trucks" *Machine Design* September 8, 2009 (http://machinedesign.com/article/wireless-sensor-network-aids-travelers-in-parking-their-cars-and-trucks-0908)

[5] Mary Catherine O'Connor, "UCLA Hospital Hopes Smart Garage Expedites Parking" *RFID Journal*, Aug. 25, 2009 (http://www.rfidjournal.com/article/view/5154)

[6] The CTPP is a set of special tabulations derived from decennial census demographic surveys that includes journey to work flow data (http://www.fhwa.dot.gov/ctpp/)

However, the latency of the data is usually high and the frequency of data collection is usually low. For the CTPP, the data used for the measures provided are only available every 10 years and there is significant lag time before the data are available to the users.

Otherwise, multi-source data type environments are large data repositories that only provide data from secondary sources. One example is Data.gov[7] that serves as a data repository for federal datasets and tools but is not a primary collector of data. The downside to this system is that if the original data sets are modified, the changes will not automatically translate to the datasets provided in the Data.gov repository. Additional coordination would be required to ensure the accuracy of the data. Currently Data.gov does not perform any data integration of data obtained from different sources.

New efforts are under development to integrate multi-source data type environments but are not yet available. Specifically, the Statewide Transportation Engineering Warehouse for Archived Regional Data (STEWARD)[8] is being developed to collect and store statewide information from various sources while integrated with the Florida DOT's planning, safety and maintenance databases.

Transformation #2: Transition from purely archival data to include real-time data provision

The notion of what constitutes *"real-time"* data provision is based on two considerations. The first consideration is the data capture interval or how frequently data is collected. The other consideration is data latency or the time lag between when the data are collected and the time data can be shared with users. Whether or not data provision can be considered real-time depends on how these data are put to use. The Data Capture and Management Program seeks to support development and testing of new or enhanced applications enabled by collecting new or existing types of data at more frequent intervals and enabling these data to be shared more rapidly. How frequent data must be collected and how quickly data must reach users will depend on the requirements for these target applications. Another consideration is that real-time data provision must be coupled with real-time data usability for the promise of these applications to be realized. For example, data may be collected at the appropriate interval and provided to a user with an acceptable latency, but if the data is riddled with erroneous elements, the supported application may still fail.

Data latency among current data management efforts varies widely, as does data collection frequency, depending on the type of data captured and managed. Major data collection efforts that require surveys, such as travel behavior surveys or infrastructure inventories are collected over a period of time and then made available at a later date. For some large data collection efforts, the latency could be months or years. As an example, the National Household Travel Survey (NHTS)[9] interviews over thousands of households nationwide regarding their travel behavior patterns. This large effort takes several months to conduct. After interviews are complete, the data team must then compile, and clean-up the data before presenting to the public. The latency of the data can be quite significant such that the 2008 NHTS data was not made available to the public until 2010.

Focused data collection activities that capture weather or travel time information from roadway sensors can have near real-time data available to users. Although there are fewer data capture

[7] The Data.gov is a data repository for Federal datasets and tools available at www.data.gov
[8] http://trc.ce.ufl.edu/research/CDW/CDW2.htm
[9] http://nhts.ornl.gov/

efforts in this category relative to the number of archival efforts, the data is provided within minutes of collection, appropriate for many current traffic management applications. For example, the MADIS[10] and CLARUS[11] initiatives provide the data to the user within seconds of data collection.

Transformation #3: Move from passive acquisition to active, interrogative methods

Current data environments that gather traveler and vehicle information are highly dependent on passive data capture. For example, traveler survey efforts use pre-determined travel behavior questionnaires to ensure consistency and comparability across responses. Passive infrastructure based sensors also use passive capture techniques obtaining the same information for every vehicle that passes through the sensors.

An example of passive infrastructure-based sensors is the Portland, Oregon's data user service Portland Oregon Regional Transportation Archive Listing (PORTAL)[12] which gathers traffic information including volume, speed, occupancy, and status readings every 20 seconds from the 500 loop detectors on Portland's freeways. Although this is an extensive database, the data collected by the detectors is consistent for every data point.

In contrast, an example of active data capture is an element of the SAE J2735 probe message standard. Vehicles operating under the standard accumulate data on speed and status according to a set of default rules and then communicate these data when the vehicle comes within range of DSRC-based roadside sensors. The rate and type of data generated and stored on the vehicle may be altered on-the-fly, however. System managers can request vehicles collect data at more frequent rates in response to emerging traffic conditions, e.g., in or around incident locations[13]. Additionally, specific applications may request unique vehicle or traveler data to meet their needs.

Several concurrent technological trends have the potential to reshape the traditional infrastructure-based, passive acquisition data paradigm, such as vehicles and hand-held devices that are capable of systematically collecting and communicating a broad range of probe data and modern wireless communication technology that permit an active exchange of data with and between vehicles, travelers, roadside devices, and system operators. An active paradigm allows for a systematic yet dynamic and selective exchange of vehicle status and traveler behavior data. These transformative forms of management may have the potential to increase system productivity and traveler mobility significantly while concurrently reducing environmental and safety impacts.

Outcomes and Federal Role

Identifying desired transformations in data capture and management practice is an important activity. However, listing a set of desired outcomes is not sufficient to achieving program goals. First, the set of key, inter-related issues must be considered that are critical to meeting identified program goals. The role and influence of federal agencies with regard to each issue must be examined, and then a broad plan of action developed. In the next two sections, this document

10 http://madis.noaa.gov/
11 http://www.clarusinitiative.org/
12 http://portal.its.pdx.edu/
13 J2735 Surface Vehicle Standard - Dedicated Short Range Communications (DSRC) Message Set Dictionary, October 2009.

begins the consideration of these key issues, and then outlines a set of guiding principles for federal action to both address these issues and meet program objectives.

Key Issues

This section describes a set of key issues related to achieving the transformational goals in data capture and management identified in the previous section. A brief description of the current state of the practice with respect to each key issue is presented. This description is followed by a discussion on the federal influence that can be brought to bear on the issue to promote the transformational goals of the program.

Intellectual property rights

One of the challenges of this program is to encourage data environment creation that not only broadly shares data but also protects the intellectual property (IP) rights of agencies and private sector entities that collect and assemble transportation data. In many previous federally-funded data capture efforts, there has been little *a priori* consideration about data ownership and what uses the data may be put to prior to data collection. Having no clearly defined IP rights agreements can cause ambiguity to data ownership and rights of use. This is likely sufficient in the case of programs where data collected turn out to have little intrinsic value (e.g., because of technical or other problems), but where data are observed to have value to a broad set of stakeholders this issue can become a serious and contentious issue. When such deadlocks occur, collected data are often frozen and not released to any third party, delaying whatever valuable research or application that might have been developed. In other cases, highly restrictive data sharing agreements are made that prevent broad utilization of data, even in federally-funded research programs.

As an example, the FHWA's Transportation Technology Innovation and Demonstration (TTID) Program[14]funded travel time data collection in several metropolitan areas by NAVTEQ (formerly Traffic.com). However, per an initial agreement associated with a legislative earmark, NAVTEQ and not FHWA has ownership of data obtained as a part of this effort. The agreement has a clear restrictive impact on re-use of captured data for research and other applications. This nature and impact of the agreement between FHWA and NAVTEQ was the subject of a recent report on the TTID Program by the Transportation Department's Office of the Inspector General. Among the report's conclusions was the lost opportunity for broader utilization of obtained data as a result of the initial agreement on IP rights[15].

In contrast, the more recent I-95 Corridor Coalition's Vehicle Probe project is a collaborative effort among the I-95 Corridor Coalition, University of Maryland and INRIX providing real-time travel information using probe technology. In this case, the Coalition ensured data ownership prior to data collection. One section of their original Request for Proposal (RFP)[16] clearly stated rules for data ownership and licensing, guaranteeing that the Coalition has full right of data

[14] http://ops.fhwa.dot.gov/travelinfo/ttidprogram/ttidprogram.htm

[15] Federal Highway Administration -Transportation Technology Innovation and Demonstration Program, Report No. MH-2010-030 , December 8, 2009, available at: http://www.oig.dot.gov/StreamFile?file=/data/pdfdocs/TTID_12_8_2009.pdf.

[16] I-95 Vehicle Probe Project Original RFP, September 2007, available at: http://www.i95coalition.org/i95/Portals/0/Public_Files/uploaded/Vehicle-Probe/RFP_82085N_Final_Final.doc

ownership and use. In addition, the Coalition had a Memorandum of Understanding (MOU)[17] with the University of Maryland stating that the Coalition has the right to use all traffic data procured by the University for the project for its own purposes as well as public dissemination. At the same time, the language in this agreement also protects INRIX from unfair duplication of its data by competing private sector service providers.

Data acquired under the Data Capture and Management Program is intended to be broadly shared but at the same time fairly respect the IP rights of data providers. Drawing on successful agreements like the I-95 Corridor Coalition Agreement, data collected in this program will develop clear rules of engagement. If these rules are clearly identified at the time that data procurements are released into the market, data providers who find the agreements too permissive may choose not to participate or to negotiate modifications to the agreements prior to the start of work. Conversely, if the rules of engagements are not clearly identified, this may result in numerous dispute resolutions and arbitrations. The federal role with respect to this issue is to act as an honest broker balancing IP rights and the value of broad data sharing, communicating with all stakeholders to gain a complete understanding of potential issues as early as possible. The Data Capture and Management Program will seek a set of balanced agreements, assess their impact and utility, and freely provide both the agreements (potentially as templates for other agreements) as well as lessons learned from their implementation.

Note that all data elements within a multi-source environment need not conform to a single uniform policy with respect to each issue (e.g., IP rights). For example, access to individual vehicle drive train status data may be restricted while vehicle speed and location data may have fewer restrictions. Some data may be more easily shared in aggregate form rather than in raw form. The intent of the program is to find the best possible agreement meeting the needs of both data providers and data users.

Privacy

There have been no clearly-defined privacy policies in the majority of the data warehousing activities. In some cases privacy is not a concern. These cases include roadway inventory data sets such as the National Transportation Atlas Databases[18] which is a set of nationwide geographic databases of transportation facilities, transportation networks, and associated infrastructure or the National Transit Database (NTD)[19] which includes annual reports on operational and service characteristics and financial statistics. However, when capturing personal or travel behavior data, privacy can be a major issue. For activities that collect personal information during data collection and provide unrestricted on-line access, the general policy is to remove any identifying personal information from the data sets. For example, the NHTS collects home location and origin and destination coordinates for each trip in a household but does not include this information in the publicly available datasets.

Privacy protection concerns may restrict some data collection programs from releasing disaggregate data from their studies. For example, the Virginia Tech Transportation Institute

[17] I-95 Vehicle Probe Project Sample Memorandum of Understanding, February 2008, available at: http://www.i95coalition.org/i95/Portals/0/Public_Files/uploaded/Vehicle-Probe/I-95%20CC%20VP%20MOU%2012%20Feb%2008.doc

18 http://www.bts.gov/publications/national_transportation_atlas_database/
19 http://204.68.195.57/ntdprogram/

(VTTI)'s 100-cars Naturalistic Driving Study collected pre-crash naturalistic driving data including video of driver's reactions[20].
A challenge of the Data Capture and Management Program will be to gather data in a manner that respects the privacy of individuals, while sharing the data among a variety of stakeholders in the public and private sectors. The privacy challenge and the federal role in managing the issue is similar to the handling of IP rights. The federal role is to act in the role of the honest broker, balancing privacy and data sharing in a way that fairly meets the needs of stakeholders. As in the case of IP rights, the Data Capture and Management Program will actively share agreements and documentation assessing the implementation of these agreements.

Governance

In general, current data warehousing efforts have little or no formal governance. Exceptions include the Next Generation Simulation (NGSIM) program that provides high-resolution vehicle trajectory data sets and core simulation algorithms; and the TRANSIMS open source website that provides access to software, data sets and documentation, and supports community interaction. Both the NGSIM and TRANSIMS programs provide an open source community that is independent and self-governing. There are well defined rules of governance for the NGSIM[21] and TRANSIMS[22] communities that operate according to rules established in the community charter that is described online and enforced by members of the community.

Established governance among the majority of current data warehousing efforts is rarely documented and made publicly available. However, some programs request a statement of how the data will be used before granting access to the data.

Even fewer programs provide a forum for exchange of information among users. One program with this feature is the HPMS community exchange site. This forum allows anyone interested in HPMS to interact with other professional on HPMS issues.

For a successful program, the data should be captured and managed consistently according to governance established for each data environment. In addition, data accessed from the data environments should be used appropriately according to established rules of engagement. As such, a community of stakeholders could be set up to self-police issues around the data environment as they relate to data access, quality, integrity and utilization. A key federal role in this regard is to make governance documents widely available. Sharing these documents will inform potential partners and participants, guide data environment management responsibilities over time, and can be used as templates by other agencies developing data environments.

[20] The 100 Car Naturalistic Driving Study Phase 1-Experimental Design Report, December 2002, available at:
http://www.nhtsa.gov/staticfiles/DOT/NHTSA/NRD/Multimedia/PDFs/Crash%20Avoidance/2002/100CarPhase1Report.pdf
[21] NGSIM Community Charter available at http://ngsim-community.org/index.php?option=com_docman&task=cat_view&gid=36&Itemid=34
[22] TRANSIMS Community Charter available at http://www.transims-opensource.net/index.php?option=com_docman&task=doc_download&gid=110&Itemid=22

Standards and Regulation

Currently, federal data capture programs reflect various levels of standards and regulation. It is especially difficult for programs operating without the benefit of clear set of guidelines or standards when integrating data from various sources.

Detailed standards for data are key for interoperability goals. The SAE J2735 standard for dedicated short range communication (DSRC) for the IntelliDrive probe vehicle data supports interoperability among DSRC applications through the use of standardized message sets, data frames and data elements. This standard identifies a set of specific data, including vehicle location, speed, and status as well as rules on how often these data are to be generated by vehicles and shared with roadside devices. For other mobile entities, such as smartphones and travelers, no such standards have been developed in such detail.

Standards for data formatting are also important. For the HPMS program, the FHWA requires that each state or jurisdiction collect and report geometric data and traffic volumes. However, each jurisdiction provides the data in different formats and on their own websites that may pose difficulties in merging the data either manually or via virtual on-the-fly integration.

The Data Capture and Management Program plans cross-cutting activities that can serve as a key role in motivating and defining emerging standards or data-related rule-making. In many cases this will involve identifying relevant existing standards and any potential changes or enhancements these standards might require. In other cases, where no applicable standard can be identified, this potential need will be noted and any potential federal action to address this gap will be coordinated with the ITS JPO Standards program.

The program must do two things: identify what standards apply to specific technical goals of the program, and then see if there are gaps with no standards or examples where standards can be enhanced or expanded to meet program goals.

A similar approach is expected with respect to federal regulation. Existing regulation may be refined or enhanced, or new regulation identified, if needed.

An example of existing regulation that may play a part in Data Capture and Management Program activities is the Section 1201 regulation regarding nationwide traveler information provision. At the current time, the language within the regulation allows significant leeway in specific data collected, processed and provided to travelers. It is possible that a more refined or specific form of the regulation may result from successful research and testing within the Data Capture and Management Program.

The approach of the program with respect to assessing standards and regulation may be similar but the federal role is different in both cases. Issuing new or revised regulation is the most direct action the federal government can take in pursuit of program goals. Standards may be identified as a part of this regulation, if needed. However, the federal role can be less direct but still influential in the area of standards, influencing and facilitating the enhancement, development and adoption of standards.

Meta-Data

Metadata can be defined as “data about data, or more precisely, definitional and descriptive data that provides information about or documentation of other data managed within an application or

environment."[23] Some efforts provide no meta-data, while other large survey efforts provide very detailed information on how data was captured and organized. Large travel survey type efforts usually provide detailed information on the survey methodology, data captured and organization. The NHTS provides an extensive meta-data document online for the users[24].

The NGSIM effort also provides extensive meta-data documentation of their datasets and algorithms. This includes detailed information on freeway and arterial corridors where NGSIM data were collected. A user of the NGSIM data can view schematics of the roadway geometry and examine images of all posted signage present at the time data were collected. These meta-data are not specifically tied to the well-documented vehicle trajectory data that is the core of the NGSIM effort but provide critical contextual information that enables deeper understanding of the data provided and supports broad re-use of the data for purposes beyond the intent of the original data collection.

Clear guidelines for meta-data consisting of a high-level description of the data environment, what data types it contains, and the general conditions under which data were captured will be explored for the Data Capture and Management Program. Meta-data standards may also be utilized as necessary.

Quality Assurance

Quality of data varies across current data warehousing efforts, both in terms of what constitutes quality and how quality is assessed. In some cases, only highly processed and quality control data are released. The nature of quality control is based on the nature of the data and how quickly the data need to be made available to other users. Efforts with longer lag times have generally more rigorous quality processes, those with short lag times generally do not.

Even with longer lag times, some data environments that aggregate decentralized data may have inconsistent quality control processes depending on the entity that performs the data collection. For example, the HPMS data are collected by states and other jurisdictions and are reported to FHWA. It is the responsibility of each individual state or jurisdiction to perform data quality checks and no uniform data quality controls are in place.

In addition, near real-time data capture efforts have a difficult task balancing a tradeoff between reduced data latency and high data quality. As an example, in order to minimize the delay before data are available, Clarus posts data for users before quality checks are done and then reposts the data once quality checks are performed. In contrast, MADIS performs a series of automated quality checks before posting the data.

Data capture programs have varied ways to perform data quality checks. As an example, FHWA's TTID Program uses an external reviewer for data quality checks. Noblis performs trend analysis to validate quality checks and the results aid in recalibration of the roadway sensors. In contrast, the Minnesota DOT and University of Minnesota as part of the Twin Cities Traffic Data Archive Effort gather loop detector data and utilize probe vehicles to check travel time accuracy.

[23] ASTM E2468-05 - Standard Practice for Metadata to Support Archived Data Management Systems
[24] NHTS User's Guide, available at: http://nhts.ornl.gov/publications.shtml#usersGuide

Other programs such as the I-95 Corridor Coalition's Vehicle Probe project have formal data validation protocols which are made available on the web for users[25].

With respect to data quality, the federal role is to document and provide data that meet clearly defined quality criteria. This can be demonstrated in a set of prototype data environments under the aegis of the Data Capture and Management Program. Another opportunity is in defining data quality standards. The federal role can include the issuing of guidelines or regulations regarding these quality standards. An example of such regulation is Section 1201.

Storage

Various data warehousing efforts have different storage and backup methods. Some such as the NGSIM Community mirror the data and store them at different backup locations. Others such as the HPMS provide a centralized source where all the data can be accessed but the data captured by each jurisdiction is individually stored by them.

Furthermore, centralized data repository efforts such as Data.gov provide access to data sets that primarily duplicates data that are available elsewhere. The issue with this system is that changes or corrections to the data must be made in several locations and there is potential for differences to arise between the same data set in different locations.

The current data warehousing efforts are unlikely to have elements of virtual data warehousing with dynamic data integration tools, where the various data are housed in different locations but can be combined on-the-fly.

The data environment concept for the Data Capture and Management Program does not necessarily imply a single centralized federal repository. If the rules for participation within a data environment are clear, it is possible that disparate data sources can be housed in a distributed form and integrated on-the-fly. Data may stream into the virtual data environment from private or public sources, be integrated by private or public sector entities, and then applied by the same or different array of private and public sector entities.

Access and Security

Access and security protocols vary widely among data warehousing activities depending on the nature of the data. Quite a few efforts have unrestricted on-line access to their data, while others provide some data online but provide additional archival data by request from the user .

Some programs such as the Mobility Monitoring Program only provide summary reports on-line but access is restricted for disaggregate data that must be requested by the user.

In general, the programs providing near-real time data restrict their data access sites to users with valid login accounts. Users must contact the program managers to obtain user account. Access to data sets by international participants may need to be restricted subject to relevant federal policy. However some features may be further restricted. For example, the Freeway Performance Measurement System (PeMS) restricts some account features to only Caltrans affiliates.

25

http://www.i95coalition.org/i95/Projects/ProjectDatabase/tabid/120/agentType/View/PropertyID/107/Default.aspx

It is clear that any federal program with a web presence must consider cyber-security as an integral part of system design and operational practice. The federal role in the area of access and security will be to establish and document practices that best tradeoff a desire to broadly share resources with a desire to maintain security and data integrity.

Operations and Maintenance

Most data warehousing environments have some type of support contact information so that users can report problems or obtain help with their data needs. Some such as the NGSIM program provide several support staff contacts for each type of support needs: technical help, data or program information, website information. Some web-based development environments, like TRANSIMS, also provide frequently asked questions (FAQs) pages to help the user.

It is clear that the notion of how data environments are maintained over time must be a key federal consideration in the stewardship of these data. The federal role in this case is not to be the permanent caretaker of all transportation data collected everywhere forever. Rather, a more nuanced role here is to seek opportunities from innovative data environment concepts, test them and show their value. If data environments are shown to have long-standing research or other application-related value, then transitioning governance and/or maintenance responsibilities to academic partners, non-governmental organizations, and other public/private consortiums may be considered. Another key consideration is the level of maintenance required for different data environments. Some environments may require more maintenance support than others. This issue needs to be taken into account when considering long term stewardship.

Guiding Principles

This section provides a set of guiding principles for the Data Capture and Management Program based on the set of identified key issues. These guiding principles are intended to shape the federal role in all program activities to better the chances of achieving program objectives.

1. ***Data obtained in the Data Capture and Management Program are intended to be broadly shared***. This guiding principle sets broad sharing of data as a foundational element of the program. That does not imply that all data need be available without any restriction. However, to avoid situations where IP, privacy or other issues make broad sharing of data overly difficult or impossible, procurements and agreements associated with the program will clearly identify this as an overarching goal. Respondents to these requests may choose not to participate or to offer potentially restricted data in forms that mitigate or eliminate these restrictions. Organizations willing or capable of broadly sharing data will be sought as partners for collaboration within program activities.

2. ***The program will provide clear guidance and document transparent processes***. This principle is related not only to sharing, but to all aspects of the program. By developing and documenting data environments, the program can influence the broader field to find ways to resolve the key issues described previously in this document. Providing clear guidance and documenting processes and lessons learned carefully, the program will improve its chances of reaching its objectives.

3. ***Establish a leadership role in identifying standard sample agreements that strike a fair and useful balance between competing issues.*** While broad sharing is the over-arching goal of the Data Capture and Management Program, the program cannot afford to be

rigid and inflexible. The program must seek and share standard sample agreements that best balance the value of data obtained and the limitations associated with some data elements. If the inclusion of some data in any form prevents sharing with applications development or other federal research, these data elements will not be provided within the data environments. In addition, data acquisition projects that are funded in whole or in part with program finds shall:

a. Name the US DOT as a co-owner of the data in their contract agreements to procure such data. The FHWA reserves the right to include such data in its Data Capture and Management Program.
b. Utilize standard clauses contained in the Standard Sample Agreements that would ensure readily uploads of the data into the Data Capture and Management Program.

4. ***The program will integrate data from multiple sources, particularly multi-modal sources.*** The Data Capture and Management Program will transition from current single source environments to multi-source and multi-modal data environments. In particular, the program will make special effort to collect, integrate and share transit data and freight data.

5. ***The program will strike a proactive rather than reactive role with respect to key issues.*** Particularly with respect to standards, governance, security, meta-data and quality issues, the program will actively seek to identify opportunities to utilize existing guidance and standards to resolve issues and promote data re-use and interoperability. This includes issuing targeted white papers or other program policy documents that can influence processes outside of direct federal control to achieve program objectives.

6. ***The program will prioritize its resources on the support of innovative applications of federal interest***. The Data Capture and Management Program cannot meet all the data needs for every candidate application nor for every potential research project. The program will focus its limited resources on the development of targeted data environments that leverage new forms of data and support one or more innovative applications of federal interest. These data, the agreements developed to obtain the data, and the developed data environments may have broader application. These ancillary applications or research activities will be supported only within the capability of limited program resources.

7. ***Demonstrate benefits of innovative data capture and management techniques in prototypes and proof-of-concept testing***. Core concepts and practical methods for dealing with key issues can only be effectively demonstrated through development and implementation. Specifically the notions of transitioning from archival data to the inclusion of real-time data provision and moving from passive to active acquisition techniques will be tested by the program. The benefits of adopting these practices will be clearly defined and measured as a part of the overall Data Capture and Management Program.

8. ***Access to data developed in the program will be equitable***. This notion of equity implies that data should, to the extent possible, be shared not only with organizations working on program-funded activities, but with other organizations as well. In addition, the data resources and products developed by the program should be broadly available for use by individuals.

9. ***Plans made within the program will reinforce long-term stewardship of obtained data***. The program will identify sustainable and evolutionary paths for obtained data. Initially data obtained in the program will be managed under federal aegis. However, a flexible plan to sunset this support by transitioning the data to other federal programs (e.g. Data.gov) or other organizations (e.g. academic or non-profit) will be included in any developed data resource. The program will also consider how data can be archived and sustained over time. Migration to private sector data management programs for long-term storage of data sets and fees associated with storing the data, are among the options that this program will evaluate.

Next Steps

The Data Capture and Management Program has planned a range of activities to begin in 2010. Although a complete program plan has not yet been developed, a number of activities planned in the first phase of the program are intended to specifically address the federal role identified in this paper.

For example, the program has planned a detailed state-of-the-practice assessment with respect to institutional and policy issues. Findings from this assessment will shape the types of agreements used to obtain data within the program.

Test data sets and prototype data environments are planned to be developed and made operational in the first phase of the program. These efforts will require practical demonstration of how competing issues can be successfully resolved to support broad data sharing.

In addition, once a set of candidate applications is identified in the Dynamic Mobility Applications Program, a detailed state-of-the-practice assessment of relevant standards will be initiated. Findings from this assessment will influence the role these standards will play not only in the prototype data environment developed in the first phase of the Data Capture and Management Program, but also in data environments developed in later phases of the program.