Texas A&M
Transportation
Institute

Data Management Life Cycle

Final report

PRC 17-84 F

TRANSPORTATION
Policy Research CENTER

Data Management Life Cycle

Texas A&M Transportation Institute

PRC 17-84 F

March 2018

Authors

Kristi Miller

Matt Miller

Maarit Moran

Boya Dai

Data Management Life Cycle

Transportation inefficiencies cost money, reduce safety, increase pollution-causing emissions, and take time away from people's lives. In transportation, decision-makers use data to assess alternatives, weigh tradeoffs, and to evaluate performance. Stakeholders use data to assess the comprehensive performance of a transportation organization. The public uses data to inform their personal decisions and travel behavior. Transportation data is a key component for policy research and performance management.

This report provides a roadmap of data management to be used for high-level prioritization for future research efforts. Researchers developed the data management life cycle to organize data, characterize its nature and value over time, and identify policy implications of cross-cutting data management issues.

The report discusses the seven phases data moves through in its life cycle:

- Collection.
- Process.
- Store and secure.
- Use.
- Share and communicate.
- Archive.
- Destroy or re-use (concurrent phases).

The following cross-cutting issues in the data management lifecycle, which occur and can change over the life cycle, but effect each of the life cycle phases, are also identified and discussed:

- Purpose and value.
- Privacy.
- Data ownership.
- Liability.
- Public perception.
- Security.
- Standards and Data Quality.

The volume of transportation data expands continually. Technological advances are happening at a rapid pace, generating large amounts of data that appear to be valuable in understanding the issues that form transportation policy. As data continues to expand, it is important for policy makers to know the value of data and the return on investment for collection and analyzing data. Data-driven insight can serve to inform policy decisions at all levels, helping to conserve limited public funds and ensure the most efficient and effective use of transportation systems.

Table of Contents

List of Figures

List of Tables

Introduction

Transportation inefficiencies cost money, reduce safety, increase pollution-causing emissions, and take time away from people's lives. The solution is not always to build more roads, create parking spaces, or add more bus routes. Sometimes, the better solution is to do more with the infrastructure we already have, and for that, you need information on which to base decisions.

Data are raw material representing actions or transactions in the real world that are recorded, classified, processed, stored, and potentially repurposed to create information that supports policy and decision making. The end user interprets the meaning to draw conclusions and identify implications of the information (*1*). In transportation, decision-makers use data to assess alternatives, weigh tradeoffs, and to evaluate performance. Stakeholders use data to assess the comprehensive performance of a transportation organization. The public uses data to inform their personal decisions and travel behavior.

Transportation data are a key component for policy research and performance management. Examples of data that reflect the wide range of data sources used for transportation purposes include the following:

- Crash records that reveal incident location and contributing factors.
- Probe speed and volume data to inform congestion mitigation and management efforts.
- Census data to show demographic and socioeconomic characteristics, population distribution, and change.
- Roadway inventory to estimate the supply and demand of infrastructures.
- Travel behavior data to identify patterns and trends.
- Public opinion data to reflect attitudes and awareness of transportation issues.
- Road weather information data to alert travelers to roadway conditions and traffic operations.

The volume of transportation data expands continually. Technological advances are happening at a rapid pace, generating large amounts of data that appear to be valuable in understanding the issues that form transportation policy. As data continues to expand, it is important for policy makers to know the value of data and the return on investment for collecting and analyzing data.

The importance of data in this era of data-driven decision making, the swift increase in the volume of data due to improved collection methods, new uses such as automated and connected vehicles, and increased interest on the part of the public in factors underlying decision making, suggests that policymakers may have an interest in understanding and addressing the quantity, quality, creation, collection, storage, retention, privacy, security, and availability of transportation data across agencies.

This paper attempts to bring clarity to the topic of data—to simplify and organize it into something that is digestible. By better understanding the data landscape as a whole, policy makers can better understand the role of each piece of data as it relates to transportation, as well as in other areas. This report provides a roadmap of data management to be used for high-level prioritization for future research efforts.

The report is organized as follows:

- Data Management Life Cycle. This section describes the process used to categorize data topics and develop the data management life cycle, as well as introduces the components of the data management life cycle.
- Data Management Life Cycle Phases. This section describes each of the eight phases in the data management life cycle in detail.
- Cross-cutting Issues in Data Management. This section describes eight issues that cut across all phases of data management.
- Summary. This section summarizes the data life cycle and provides suggestions for future research efforts.

Data Management Life Cycle

Accurate, timely data is an important input for making accurate, timely transportation planning and policy decisions. However, the management of data is challenging and must be addressed over the life span of a piece of data. Transportation agencies already manage many of their physical assets: roads, bridges, signs, lights, etc. Data can be treated like other physical assets. Data is a key component in decision-making, so it is important to also carefully manage and maintain data to know what data exists, where it is located, how it can be obtained, and if it is accurate. Furthermore, data are often expensive to procure, so one would want to make sure the right data are available to support key decisions.

Data as a topic is so broad; it can be overwhelming and difficult to grasp all the elements it encompasses. Through a cyclical and iterative process, researchers at TTI identified possible aspects and uses of data in the transportation context and developed a framework of what data exists, and then condensed the topics into cross-cutting issues and main themes in the data management life cycle. This life cycle presents a way to organize data, characterize its nature and value over time, and identify policy implications of cross-cutting data management issues.

Illustrated in Figure 1, the data management life cycle describes key aspects of data from creation to destruction, as well as cross-cutting issues that affect data in each phase of the life cycle. Data moves through seven phases in its life cycle:

- Collect.
- Process.
- Store and secure.
- Use.
- Share and communicate.
- Archive.
- Destroy or re-use (concurrent phases).

Researchers at TTI also identified seven cross-cutting issues in the data management lifecycle, which occur and can change over the life cycle, but affect each of the seven life cycle phases (Figure 1). Some cross-cutting issues are pivotal to each life cycle phase, and all have policy implications. The cross-cutting issues are:

- Purpose and value.
- Privacy.
- Data ownership.
- Liability.

- Public perception.
- Security.
- Standards and Data Quality.

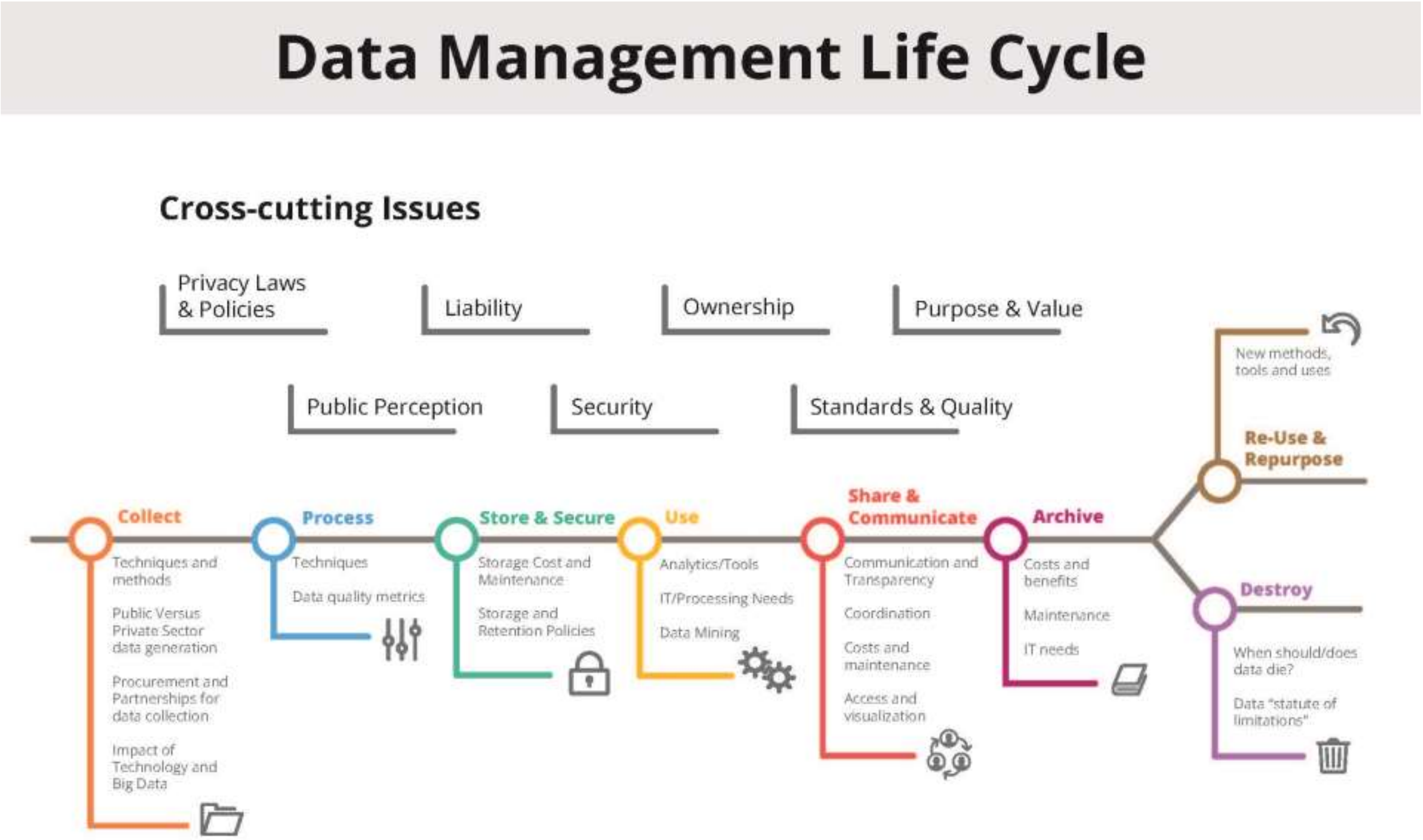


Figure 1. Data Management Life Cycle and Cross-Cutting Issues in Data Management.

Data Management Life Cycle Phases

The stages of the data management life cycle—collect, process, store and secure, use, share and communicate, archive, reuse/repurpose, and destroy—are described in this section.

Collect

The first phase of the data management life cycle is data collection. Data is being collected for a myriad of reasons, such as operations, maintenance, planning, performance measures, or to address a certain policy goal or objective. The key factors in this stage are:

- Techniques and methods for collection.
- Public versus private sector data generation, procurement, and partnerships for data collection.
- Impact of technology and big data.

Techniques and Methods for Data Collection

Transportation data relate to people, vehicles, assets, physical infrastructure, and travel. Users of the information derived from the data are key stakeholders in the data collection and analysis process. Depending on the needs of the user, the data collection type and methodology vary at different geographic and jurisdictional levels. Data collection systems should be designed to meet both internal and external user needs and the agency's legislative mandates. The planning and design of data collection system includes establishing data needs and objectives, identifying data providers, planning and designing methods to meet data needs and objectives, and documenting data collection and designs (*2*).

Data collection methods should be determined based on factors such as funding availability, data quantity, length of collection period, research questions, and target populations. Future research should be focused on examining ways that public agencies can harness big data from private entities.

Partnerships for Data Collection

Data collection can be challenging for transportation agencies with limited time, resources, and technology. The process of identifying and collecting accurate and useful data requires technical expertise and well-developed tools. A public-private partnership in this case could help to facilitate data collection and enhance agencies' ability to be data-driven development practitioners and decision makers. Currently, Texas' public-private partnership mostly focuses on the State's facilities and infrastructure projects. There is a lack of formal guidance on potential collaboration of data collection. Before entering a public-private partnership, it is critical to be aware of existing data ownership policies and clearly describe rights and obligations so data integrity is not compromised.

There are multiple ways vehicle data is collected by public and private sector sources. In the public sector, sensors on roads put in place by local and state DOTs collect vehicle speed and volume data that is not associated with the personal identity of the vehicle owner. In the private sector, individual vehicle telematics data is obtained via cellular backhaul transmissions by telecommunications companies who have agreements in place to route the data to automotive manufacturers who then use it for various purposes.

Impact of Technology and Big Data

Collection and exploitation of large data sets for transportation operations, planning, and safety purposes is not new; in the past data were acquired, processed, and discarded. Now with low-cost and widespread sensing across all modes and types of infrastructure, they are acquired, processed, and stored for some later currently unknown use.

It is important to understand what data have been generated and how to use them to shape the future of transportation in Texas. Millions of devices have been equipped with Internet of things (IoT) technology. The IoT refers to "the network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data" (*3*). Application of the IoT extends to all aspects of transportation systems, (i.e., the vehicle, the infrastructure, and the driver or user). It automates data collection and generates a massive pool of data (Big Data) from diverse locations that is aggregated very quickly. For example, Google has crowd sourced the collection of real time traffic data via mobile phones. If the Google Maps app is installed on a mobile phone with GPS capabilities enabled, Google can collect the location and travel data of the phone user in real time. When Google combines the speed collected from all the phones on road, they are able to evaluate live traffic conditions and send it back to user for navigation.

Federal and state laws and rules place requirements on the collection of certain data related to various aspects of transportation. One example in the field of new transportation-related technologies are recently developed laws involving data collection requirements surrounding automated license plate reader systems (ALPR) mounted on police cars, road signs, and traffic lights that capture geo-located and temporal data aligned with PII data from these systems. Given the PII, data collection requirements have been created for ALPRs across various states. Table 1 describes some of these laws (*4*).

Table 1. 2015 Status and Description of Select ALPR Laws in the United States.

Arkansas (Passed)	Highway police **division can utilize the automatic license plate reader system to collect ALPR data** for the electronic verification of registration, logs, and other compliance data **for commercial vehicles on a state highway and for installation at an entrance ramp at a weigh station** facility for the review of a commercial motor vehicle entering the facility.
California (Passed)	**Imposes specified requirements on an automated license plate recognition operator to ensure that the information the operator collects is protected** with certain safeguards, and **implements specified security procedures and a usage and privacy policy with respect to that information**. **Requires the operator to maintain a specified record of any information access**. Requires public input regarding any public entity program. Includes specified information to be considered personal information for breach purposes.
Illinois (Pending)	Allows law enforcement agency to use ALPR data and historical ALPR data only for legitimate law enforcement purposes. Prevents ALPR data from being traded or shared for any other purpose.
Texas (Failed)	Law enforcement agency may use an automatic license plate reader. All images and data produced from an automatic license plate reader shall be destroyed not later than the 90th day after the date of collection unless the image or data is evidence in a criminal investigation or prosecution.
Minnesota (Pending)	Relates to data practices; classifies data related to automated license plate readers; requires a log of use; requires data to be destroyed in certain circumstances.

Data sets, often referred to as Big Data, of this magnitude and complexity are proliferating in part because data is increasingly being continuously gathered by ubiquitous information-sensing mobile devices, GPS devices, remote sensing technologies, software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks. Examples of Big Data sources in transportation research include probe data, GPS data, Bluetooth sensors, mobile devices, and cameras.

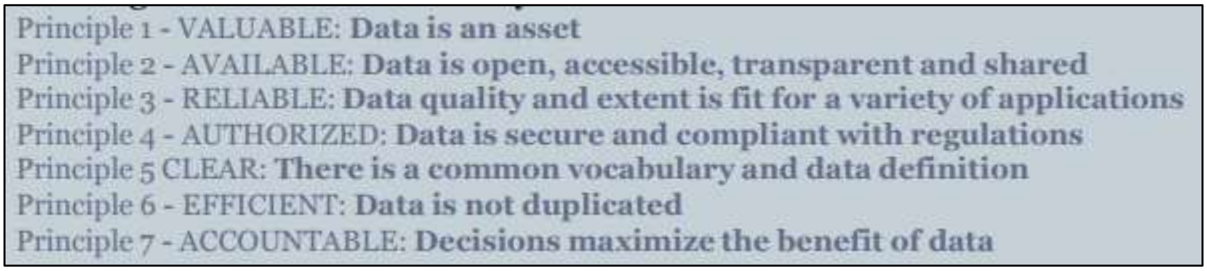
Process

Data processing is the second phase of the data management life cycle that takes a primary role in converting the data collected in the first stage of the life cycle to meaningful information. When data is collected, it may not be in a readily usable form. The process starts with discovering inconsistencies and other anomalies in the data into raw data, as well as data cleansing to improve the data quality. Users could then conduct analyses to produce meaningful information based on the data that may lead to a resolution of a problem or improvement of an existing situation. The key factors in this stage include:

- Data quality metrics.
- Quality assurance and quality control.
- Data processing techniques.

Data Quality Metrics

Data quality metrics identify data errors and erroneous data elements and measure the impact of various data-driven processes. A data quality assessment enables transportation agencies to understand the condition of their safety and traffic data, for example, in relation to expectations. It could assist agencies in understanding how effectively data represents the objects, events, and concepts it is designed to represent. AASHTO has developed seven core data principles to have consistency among states, listed in Figure 2.



Principle 1 - VALUABLE: **Data is an asset**
Principle 2 - AVAILABLE: **Data is open, accessible, transparent and shared**
Principle 3 - RELIABLE: **Data quality and extent is fit for a variety of applications**
Principle 4 - AUTHORIZED: **Data is secure and compliant with regulations**
Principle 5 CLEAR: **There is a common vocabulary and data definition**
Principle 6 - EFFICIENT: **Data is not duplicated**
Principle 7 - ACCOUNTABLE: **Decisions maximize the benefit of data**

Figure 2. AASHTO Core Data Principles.
Source: (*5*)

Data Processing Techniques

Transportation agencies, research entities, and private companies are seeking to tap the information power within big data to create more effective decision making. It poses challenges to the traditional management and analysis, which lacks the capabilities to handle the complex data sources and amount of information. To extract and mine massive transportation data from various databases, it is important to understand and use advanced data processing techniques and tools. The Bureau of Transportation Statistics provides general instructions on data processing in the *Guide to Good Statistical Practice in the Transportation Field* (*6*). This guide includes principles and guidelines on data editing and coding, handling missing data, production of estimates and projections, and data analysis and interpretation.

Stakeholders can save time and increase capacity by using the advanced tools to enable more efficient and accurate real-time transportation data processing. For example, researchers at TTI have studied potential methodologies to realize the benefits from big data resources (*7*). One of the best alternatives is cloud computing. Cloud computing is described as, "a type of Internet-based computing that provides shared computer processing resources and data to computers and other devices on demand" (*8*). Alternatively, MapReduce is "a computation process that can process a large data set simultaneously utilizing multiple nodes (processors) in a cloud platform or in a local cluster environment."

Technological advances allow for the generation of increasingly large amounts of data collected from information sensing devices such as smartphones, GPS devices, software logs, cameras, microphones, and other sensors. As the volume of data increases, transportation professionals need to have the technical skills and computer processing power to effectively use this robust data.

Store and Secure

The third phase of the data management life cycle is data storage and security. When data is secure and appropriately regulated, there is greater trust and confidence in its use. Data must be trustworthy and safeguarded from unauthorized access, whether malicious, fraudulent or erroneous. Transportation agencies at all levels of government (federal, state, and local) hold a wealth of diverse data sets, but it is often stored in different databases that are incompatible with each other or difficult to find.

The key factors in this stage include:

- Storage cost and maintenance.
- Storage and retention policies.

The global volume of electronically stored data is doubling every two years (*9*). The rapid growth in the volume of transportation data due to the innovation in data generation and collection leads to great demand of cost-effective storage technologies. More and more organizations are considering outsourcing storage services or cloud storage options because the availability of cloud computing resources opens up possibilities for users to transition to purchasing access to computing power and storage space as a service instead of maintaining it themselves. This way, providers are responsible for the performance, reliability, and scalability of the computing environment, while users can concentrate on data analysis and production (*10*).

It is important to note the risks related to cloud-based computing: unauthorized access to data by cyber-security attacks against cloud service providers, security risks internal to the cloud service provider, compliance and legal risks associated with liability for data breach, key feature price changes over time, and critical data availability risks for cloud server downtime.

Use

Data use is the fourth stage in the life cycle. Transportation data is used in numerous ways to study, plan, design, construct, operate, and monitor our transportation system. It helps planners understand traveler behavior and helps policymakers identify ways to make the system more efficient and cost-effective. It is also used to understand traveler behavior. These different uses are what make data an asset. The potential for infinite possible uses of data also creates challenges throughout the data life-cycle, from data collection to data destruction. How data can and will be used is dependent on how it is collected, processed, and stored.

A model of how data is used by departments of transportation in the United States to inform their activities, developed by Cambridge Systematics, is shown in the diagram in Figure 3 (*11*).

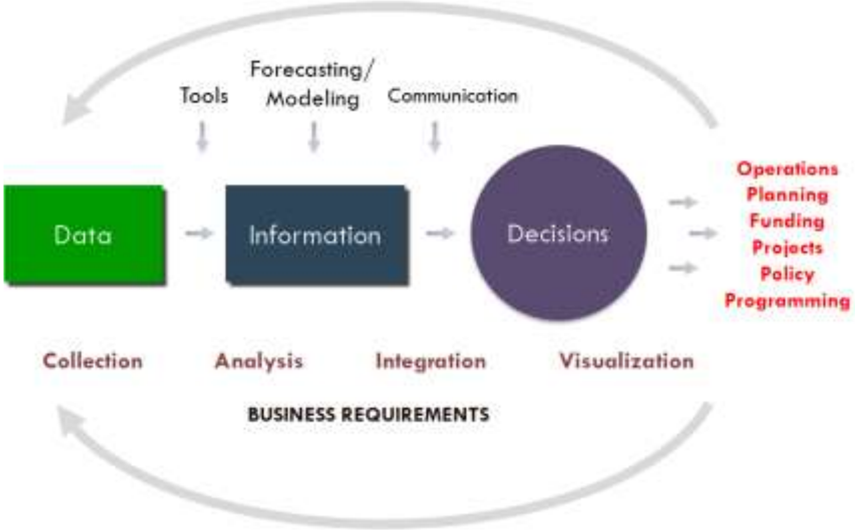


Figure 3. Model of Data Use by DOTs. Source: Cambridge Systematics.

There are several issues to consider when reviewing data use for transportation purposes, including:

- Larger and more detailed data sources can create challenges for analytic capacity among researchers and processing tools, as well as challenges sharing data across an enterprise or with partners.
- As access and availability of data increase, users need to weigh this against their ability to process and interpret the data.
- Balancing valid data uses with security concerns about access to data.
- Privacy and proprietary restrictions on the use of collected data.

To address the transportation problems the state is currently facing, it is important to first determine the questions and the demand of information. For instance, in order to prioritize transportation funding and meet individual travel needs, it is important to understand travel behaviors and patterns. The U.S. DOT has been collecting traveler information across the nation through National Household Travel Survey since 1969. The data are used by Congress, policy makers at all level of government, and transportation planners to understand the performance of the current transportation system and develop strategic plans for the future. It has also contributed to improving safety, reducing congestion, tracking air quality improvements, and planning for future transportation investments (*12*). In Texas, TxDOT started a comprehensive travel survey program in the 1990s.

A *Big Data Scan* of the Texas A&M Transportation Institute in 2015 found that large or complex data sets are used by transportation researchers in topic areas such as mobility, safety and operations, operations and energy, and transportation modeling. However, the research also suggested that there were technical, institutional, and financial limitations on the capacity for researchers to explore new uses of data. Deployment strategies for organizations to capitalize on advancing data analytics include supporting collaboration with commercial data providers and private entities specializing in big data analytics, building internal capacity to leverage existing data sources, and offering data management as a service to clients and partners (*13*).

Share and Communicate

As transportation organizations work with more stakeholders and external partners to incorporate them into decision making, planning, and operations, there is an increased pressure to also share data. Shared data can help improve decisions since agencies/researchers will be able to obtain a more comprehensive picture of the impacts their decisions have based on contributions of new data sets from a wider variety of sources, both internally and externally. At the same time, shared data will also drive a decision maker to require more quality and clarity from data gathered, which will likely result in fewer sources of more accurate and timely managed data for decision-making.

Data sharing is the fifth stage of the data management life cycle. Open sharing of information and the release of information via relevant agreement must be balanced against the need to restrict the availability of classified, proprietary, and sensitive information. There are several issues to consider for sharing and communicating data, including:

- Communication and transparency.
- Coordination within the agency, with external partners, with private sector, with the public.
- Costs and maintenance of shared data.
- Interoperability across systems (tolling, connectivity, telematics).
- Access.

Communication and Transparency

Sharing public datasets is part of government efforts to communicate with the public, maintain transparency, and engage the public in decision-making processes in transportation. For example, various transportation-related data sets TxDOT utilizes in planning and decision-making are found on their web page titled “OneDOT Data Shop.” The site is an example of a state effort to provide information about its public data sets. It provides a set of basic identifying information about each data set, including the title, description, contact person, source, and update frequency.

Coordination

As transportation organizations partner more often with stakeholders and external partners in decision making, planning, and operations, there is an increased interest and need to share data. Furthermore, agencies are increasingly asked to "do more with less." Transportation systems management and operations (TSMO) are a long-standing transportation activity in which transportation agencies collect roadway data to help manage congestion on roadways, improve incident response, and provide traffic information services. As computing technology improved, state and regional entities developed advanced traffic management systems (ATMS) that combine data from multiple public agencies and through contracts with the private sector to coordinate the transportation data networks of an entire region, across modes, jurisdictions, and organizations. Data-intensive TSMO activities often coordinated among these agencies include:

- Traffic flow performance monitoring.
- Incident detection and response.
- Traffic signal timing coordination.
- Integration of road weather information systems with the provision of traveler information.

This type of regional coordination adds complexity to the coordination of information technology system procurement and design, and how data is shared across multiple organizations, both public and private (*14*).

Costs and Maintenance of Shared Data

As data becomes increasingly available, data sharing can be a tool to combat rising costs for data storage, processing, and analysis and to identify cost-effective and efficient transportation solutions.

Access

Sharing data is a key step in reducing the burden on staff time as data becomes more accessible. Users must have access to the data critical to their duties and functions. Wide access to properly processed and packaged data can lead to efficiency and effectiveness in decision-making, and affords timely responses to information requests.

The benefits of sharing information and the release of information with public and private partner agreements must be balanced against the need to restrict the availability of classified, proprietary, and sensitive information. How transportation data is or is not shared has broad policy implications, particularly in cross-cutting areas such as data ownership, security, privacy, and liability. For example, the rapidly expanding presence in new vehicles of *vehicle telematics systems* that collect and transmit vehicle data present potential privacy risks for drivers (*15*).

Telematics systems incorporate numerous on-board communications, positioning, and computing technologies to provide services such as navigation, infotainment, remote diagnostics, and transmission of vehicle performance data for insurance purposes. At the same time, these products can collect and transmit vehicle data that is of value to public transportation agencies and to private entities with a commercial interest in developing and selling data associated with smart phones, GPS, and Bluetooth technologies in vehicles (*16*). Existing consumer protection and insurance policies may have to address the privacy issues raised by vehicle telematics and, in the future, highly automated and connected vehicles.

Archive

The sixth stage of data management is archiving. Data archiving is "the process of identifying and moving inactive data out of current production systems and into specialized long-term archival storage systems." This serves two objectives: 1) moving inactive data out of active systems and databases to optimize current performance, and 2) storing inactive data in specialized archival systems that are more cost-effective and allow for retrieval when needed (*17*). A data archive may also be called a data bank or data center. There are several issues to consider when reviewing data archiving, including:

- Storage costs.
- IT needs.
- Cost/benefit.
- State and federal requirements to backup data.
- Other issues related to data backup.

Archiving is not a new concern for transportation data users, but the complexity and costs for data archiving are growing as data is collected faster and in larger amounts. Data archiving requires a variety of software, database, and electronic data storage technologies. It also requires staff to maintain systems, develop reports, and provide IT and administrative support. In 2011, the cost to operate and maintain one multiple agency data archive, with data fusion and visualization systems, was estimated to cost approximately $400,000 per year. Estimates for other statewide or regional data archive systems range from $300,000 to over $4 million, as costs can vary widely depending on the size and features of a system (*18*).

Transportation planners and policymakers are increasingly focused on data-driven decision-making and benchmarking for performance monitoring. Archived transportation data can enable better benchmarking and tracking of improvements to the transportation system over time. Crash records are an example of transportation data that is often archived and used for numerous purposes. Under the Texas Transportation Code, TxDOT is responsible for maintaining crash data submitted by Texas law enforcement officers. Since 2007, TxDOT has been developing Crash Records Information System (CRIS), the state repository for vehicle crash data, into a comprehensive electronic crash data system (*19*). In 2011, TxDOT launched the Crash Reporting and Analysis for Safer Highways (CRASH) internet application to speed up the transfer of crash data from law enforcement agencies to TxDOT by collecting reports electronically. Use of CRASH allows for faster and more efficient submission of data from the office or a patrol car. Quality of data entry is ensured through CRASH training, which is scheduled as part of the set-up process for each agency (*20*). In 2015, the retention period for Texas CRIS data was increased from 5 years to 10 years (*21*).

Data archiving plays a critical role in ongoing efforts to design and manage intelligent transportation systems (ITS) across the United States. ITS is "an operational system of various technologies that, when combined and managed, improve the operating capabilities of the overall system" (*22*). For ITS purposes, 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" (*23*). ITS programs primarily focus on collecting real-time operational data that can be used for incident management, traffic signal control or travel information systems. In addition to providing more and better information for operations, this data can have other uses and avoid costly efforts to re-collect data for special studies (*24*). For decades, ITS programs have included efforts to support and expand the use of ITS data for transportation planning and other needs. In 1999, the Archived Data user Service (ADUS) was added to the National ITS Architecture, documentation that outlines how users should design and use ITS. ADUS was added to facilitate the use of ITS-generated data for multiple uses. Figure 4 demonstrates the various ways data from ITS sources can be used for many other transportation purposes.

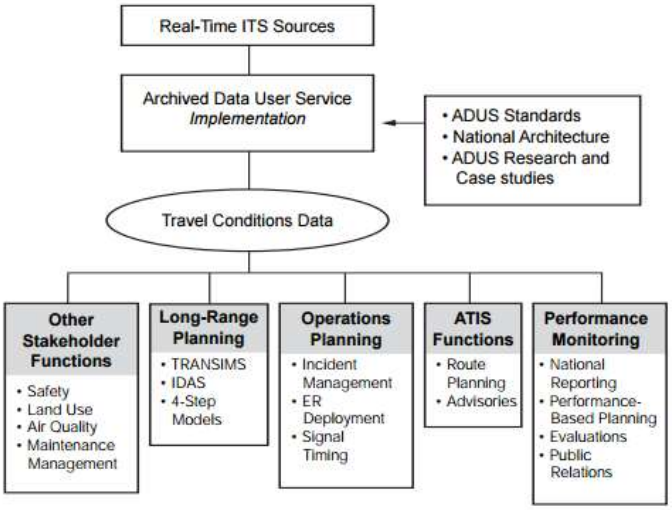


Figure 4. Use of ITS Data for Other Transportation Purposes.

Source: (*24*)

ITS programs are expanding to accommodate emerging connected and automated vehicle and infrastructure technology (*25*). The U.S. DOT Intelligent Transportation Systems Joint Program

Office’s ITS Strategic Plan 2015–2019 notes several focus areas related to data archiving, including the following:

- Enterprise data management focused on capturing, managing, and integrating “big data” from the range of ITS enabled technologies.
- Focus on ensuring interoperability within increasing complex technical systems by evolving standards and architectures to ensure that technological advancements are reflected and the required backward compatibility and interoperability are maintained.

Reuse/Repurpose or Destroy

At the end of data life cycle, data ultimately are either processed for reuse/repurpose, or destroyed when their utility has been exhausted. With data reuse and repurpose, the data management life cycle is no longer linear but has become circular. When data are appropriately handled, it can have a long life with many uses beyond its original one and serve projects yet to be planned. Data reuse refers to using the same data more than once for the same purpose; data repurpose means using the same datasets to serve a new purpose that is different from the original purpose of the datasets.

Data destruction refers to the process of removing information in a way that renders it unreadable (for paper records) or irretrievable (for digital records), so that it is completely unreadable and cannot be accessed or used for unauthorized purposes. Failure to do so can lead to serious breaches of data-protection and privacy policies, compliance problems, and storage issues.

Reuse/Repurpose

The repurposing of data enables the continuous extraction of value from data and leverages the data to solve new problems. This could also help to justify the expense of accumulating and managing huge volumes of data when organizations are monetizing or productizing their information assets. For instance, backup and archive data has represented the most comprehensive data set in many organizations. But this data is rarely used for any purposes other than restoring deleted, corrupted, or lost data. Mining the existing data for potential value creates the opportunity to turn some of the cost of backup into a resource.

There is no end in the data life cycle as far as data being continually reused and repurposed, creating new data products that may be processed, distributed, discovered, analyzed, and archived. The IoT generates a massive volume of data. For example, connected cars are equipped with more than 100 sensors creating a constant stream of data by measuring location, performance, physical parameters, and driving behavior, often several times per second. According to a 2015 Hitachi whitepaper, a single connected car will produce more than 25 GB of data per hour of use (*26*). These data can be analyzed in real-time to keep the vehicle’s performance, efficiency, and safety in check. It also provides vital feedback for cities and states about traffic volume and roadway design.

Experts say the value of vehicles will likely pale in comparison to the riches from our cars' data (*27*). The data could be reused and repurposed for different goals. Car manufacturers can analyze vehicle operating performance, assess automotive telematics, and track performance of electrical components performance in different models. Vehicle owners can be notified of scheduled maintenance and repair requirements. Providing better, more proactive maintenance support using captured data may ultimately be a factor in what makes one type of vehicle more attractive to consumers. Additionally, insurance companies can potentially track speed and driving behavior in order to reward good drivers with lower premiums. Law enforcement may use connected car data to investigate accidents or to prosecute criminals.

The reuse and repurpose of data is encouraged, especially for open data. Government agencies like the US DOT make their data available for public use, and encourage use of data by a variety of means including "hackathons" to promote interest among analytically oriented innovators and entrepreneurs. It also encourages innovative use of its data by commercial ventures including the Federal Aviation Administration data for private pilot iPads and the analysis of truck incident data patterns by insurance companies. Such uses may not always be specified by DOT's enabling legislation (*28*).

Despite the advantages of data re-use and repurpose, there are barriers such as data quality and perceived risk of reusing others' data. While secondary data analysis entails reusing data created from previous projects for new purposes, trustworthiness of data sources could be an issue. Oftentimes, there is lack of documentation of what has been done to the datasets, which becomes a significant disincentive to reusing data. Not knowing how the data was collected and cleaned poses a potential risk of generating invalid results. Standardization of procedures and formats could help to address the problem. For example, if cleaning procedures within an organization, or even across a subject field, were standardized, recipients of data would know exactly how it was cleaned. Also when data follows a standard format, it can be easily integrated for analysis by different users. Right now, extra effort is required from secondary users to preserve data interconnectedness in order to guarantee the data's understandability and informative value.

Destroy

The destruction method is normally selected based on the underlying sensitivity of the data being destroyed, or the potential harm they could cause if they are recovered or inadvertently disclosed. There are several issues to consider if the owner chooses to destroy the data, including:

- Determining when data should die, how to make the choice, and who makes the choice.
- Document retention laws.
- Data "statute of limitations."
- Usefulness of historical data versus the need for new, updated data sources.

There are three main effective data destruction approaches (*29*): overwriting, degaussing, and physical destruction. Each of these techniques has benefits and drawbacks (*30*) discussed in the following list:

- **Overwriting:** Using software or hardware appliances to overwrite data. This is one of the most common ways to address data remanence. The advantage of this approach is that it is relatively easy and low-cost. It can be used selectively on part or all of a storage medium. On the downside, it takes a long time to overwrite an entire high-capacity drive. It also may not be able to sanitize data from inaccessible areas such as host-protected areas. In addition, this process can only be used when the storage media is not damaged and is still writable.

- **Degaussing:** using a device to remove or reduce the magnetic field of a storage disk or drive. The key advantage of this approach is it makes data completely unrecoverable. However, a strong degausser can be expensive and heavy. It may even produce collateral damage to vulnerable equipment nearby due to its strong electromagnetic fields. Also, the damage to the drive is destructive; the drive will be unusable after degaussing.

- **Physical destruction:** physical media can be shredded or shattered using various physical destruction methods to keep the data from being recovered. For very low risk information, this may mean simply deleting electronic files or using a desk shredder for paper documents. However, these types of destruction methods can be undone, making these methods inappropriate for more sensitive data. For more sensitive data, stronger methods of destruction at a more granular level are needed to assure that the data are truly irretrievable. On the other hand, physical destruction can provide the highest assurance of absolute destruction of the data since it is impossible to reconstruct or recover the data from a disk or drive that has been physically destroyed. But this involves high capital expenses and is considered an unsustainable and a costly way to dispose of data.

For the purpose of protecting privacy, data destruction is a critical and often required process. Personally identifiable information (PII) is often collected by businesses and the government and then stored in various formats. In the United States, at least 31 states and Puerto Rico have enacted laws that require entities to destroy, dispose, or otherwise make personal information unreadable or undecipherable. The Federal Trade Commission’s Disposal Rule also requires proper disposal of information in consumer reports and records to protect against “unauthorized access to or use of the information.” The rule applies to consumer reports or information derived from consumer reports (*31*). In Texas, the Business and Commerce Code includes regulations about disposal of certain business records (Tex. Bus. & Com. Code § 72.004). The Texas Administrative Code, Title 13, Chapter 7 establishes the minimum requirements for destruction of local governments’ source documents. However, there is no established law in Texas regulating how the government’s data must be destroyed.

A common question is how long data should be retained before being destroyed. The answer varies depending on the kind of data. For research records, it is recommended they be kept for at least five years and possibly longer, depending on the longest applicable standard (*32*). For Texas state agencies, the answers can be found in the agencies' retention schedules. A records retention schedule is a document that identifies and describes a state agency's records and the lengths of time that each type of record must be retained. Texas state agencies are required to submit their retention schedules to the Texas State Library and Archives Commission (TSLAC) on a timetable established by administrative rule. If a record series does not appear on a certified records retention schedule, it may not be destroyed without obtaining special permission of TSLAC's executive director. Historically, TxDOT retained five years of crash data. However, in 2015, an update to TxDOT's TSLAC retention policy was made, and TxDOT moved to a 10-year retention period for crash data. As a result, TxDOT has crash data from Jan. 1, 2010, to present, and will accrue data for 10 calendar years. Records prior to Jan. 1, 2010, have been purged and are no longer available.

Cross-Cutting Issues in Data Management

Several issues affect all stages of the data management life cycle. The research team identified seven cross cutting issues: purpose and value, privacy, data ownership, liability, public perception, security, and standards and quality, each discussed in the following sections.

Purpose and Value

Data is collected, produced, and reported to serve certain purposes. It is important to identify the purpose and value of each stage of data throughout the data management life cycle, and to whom the data is valuable. Each phase of the data management life cycle centers around the purpose and value of the data. In each phase, questions about why it is being done and why it is important need to be answered.

Understanding the purpose and value of data is important in the decision to re-use or repurpose data, as well as when monetizing data. Research indicates stakeholders are currently taking steps to monetize vehicular data from automated and connected vehicle technologies. Data as an asset has high potential future value. For example, the expected growth of the value pool from car data and shared mobility could add up to more than $1.5 trillion by 2030 (*33*).

The value of data also increases when it is used with other data and in a variety of applications. The use of multiple datasets together can significantly contribute to transportation planning decision-making. For example, the development of the Rural San Antonio Bike Plan for TxDOT San Antonio District involves an assessment of existing bicycle conditions. The data collected and used in the review of current conditions cover three aspects: network supply (shoulder width, speed limit, and traffic volume), bicycle trends (bicycle commuter mode share, bicycle destinations, and bicycle crash), and public input (rural county planning and annual bike meeting). When looking at each factor separately, the information provided is limited. But, when combined, planners have a comprehensive understanding of the current bicycle environment in San Antonio. The plan identifies how attractive each roadway section is for bicycle activity, which will guide the development of the prioritization framework for bicycle accommodations.

Data is a core industry asset that has measureable value and is managed accordingly. The question must no longer be just "who owns the data?" but "who can use that data for what purpose?" If this question is answered creatively, pragmatically, and transparently, data could be used collaboratively to create bigger value.

Privacy

Data privacy is an issue affecting each stage of the data management life cycle. Privacy is especially important when determining standards and degree of conformance, allowable uses, and processes. There has been a growing concern about privacy protection in transportation data collection. Central to this complex subject is location privacy, described by Beresford and Stajano as, "the ability to prevent other parties from learning one's current or past location" (*34*).

Transportation and location data can reveal personal travel habits as it provides both spatial and real-time data on the traveler's activities. A recent MIT research study that found that four pieces of location data combined with finance data could re-identify 90 percent of individuals, despite the data set lacking any names or other traditional identifying characteristics (*35*). One result of the study's findings is that it is no longer safe to assume that anonymized data does not require the same high level of controls as information with personally identifying characteristics. If location and time are, by themselves, intrinsically identifying, then many sources of transportation data may need stricter protections than previously thought.

Many Americans highly value the privacy of their vehicular data, but currently available and emerging vehicle technologies may make it difficult to secure this information. For example:

- A 2015 survey conducted by U.S. Senator Ed Markey's office determined that "nearly 100 percent of cars on the market include wireless technologies that could pose vulnerabilities to hacking or privacy intrusions" (*36*).
- A 2013 survey of 2000 adults by the Auto Alliance found that privacy is an issue for consumers, with 75 percent of survey respondents indicating that they were very/ somewhat concerned that companies would collect data from the software operating self-driving cars (*37*).
- A 2013 telephone survey conducted by the American Automobile Association (AAA) found that 86 percent of the 1,007 U.S. adults surveyed thought there should be laws and policies to protect their vehicle data (*38*).

An important process used to evaluate the collection of personal data in information systems is Privacy Impact Assessment (PIA). The E-Government Act of 2002, Section 208, establishes the requirement for agencies to conduct PIAs. The assessment is "a practical method for evaluating privacy in information systems and collections, and documented assurance that privacy issues have been identified and adequately addressed" (*39*). In practice, the privacy of vehicular data is addressed by different stakeholders (OEMs, aggregators, public agencies) and the different approaches align with this notion: where the data is recorded matters (*40*). For example:[1]

- Transportation agencies collect toll tag data from RFID tags, Bluetooth data sets from Bluetooth sensors, and weigh-in-motion data from freight vehicles, and view this data as public information. They are governed by law in their protection of PII. As agencies begin to use more data from the car, they will need to address protection.
- Transportation Data aggregators (TomTom, INRIX, HERE, Air Sage) create a new product from data generated from inside the car, and view themselves as owner of the product. The companies use agreements with telecommunications providers and their cellular networks to transmit location and other forms of data (speed, heading,

[1] The examples provided are not a comprehensive list of data gathered and put to use by stakeholders.

acceleration, etc.) associated with sensors in roadway infrastructure, vehicles, and smart phones to their database to generate products related to business competition analysis (number of visits by customers), real-time speeds and volume on a transportation roadway network. They have the most stringent and developed procedures for PII since data is their primary business.

- OEMs see the driver as the owner of the data generated in the vehicle, but through user agreements view themselves as "stewards" of the data. Car makers have been traditionally focused on the profit from vehicle sales, but have realized the car data has value, sometimes greater than profit from the sale of the car itself. In 2014, various industry groups adopted guidance for privacy protection, but it is not legally binding. OEMs use backhaul frequencies tied to cellular networks in order to collect data containing unique vehicle identifiers and telematics information on the current operating condition of a vehicle.
- Financial transaction records tied to location are often shared with big data providers through the same cellular networks to more effectively market consumers by their location.

This variety of stakeholders each have gathered PII and location data that may be anonymized alone, but when combined could lead to the re-identification of consumers (*41*). Major telecommunications providers, data aggregators, auto manufacturers, transportation agencies, and financial institutions work closely with one another to ensure that PII is effectively anonymized. However, there exists a large liability when it comes to the possibility of hackers combining these location-based data sets with finance data sets in a way that could re-identify individuals.

Presently, Texas laws do address PII, the risk of potential combinations of anonymized data to re-identify individuals, and the requirements/responsibilities of stakeholders who have experienced a data breach. The protection of PII is a current topic of interest for Texas, as well as federal policy makers. Table 2 describes PII-related legislation the Texas legislature has proposed and enacted.

Table 2. Proposed and Enacted Privacy Legislation in Texas.

Proposed 85th Legislature (2017)	Enacted 80th Legislature (2007)	Enacted 76th Legislature (1999)
Removal of PII from car crash information maintained by the state.	Section 521.002 of the Texas Business and Commerce Code defines personally identifiable information as information that "alone or in conjunction with other information identifies an individual" alongside what is considered identify theft.	Section 32.51 of Title 7 Texas Penal Code criminalizes fraudulent use or possession of PII, which is referred to as "identifying information."

The capture and use of data about an individual's location has not been specifically listed, but the following types of data can be included in current Texas law (521.002) as a type of personal identifying information:

- Name.
- Social security number.
- Date of birth.
- Government-issued identification number.
- Unique biometric data.
- Unique electronic identification numbers.
- Addresses.
- Routing codes.
- Telecommunication Access Device (card, plate, code, account number, personal identification number, electronic serial number, mobile identification number, etc.).
- Vehicle location data associated with Bluetooth and GPS devices that rely on cellular towers, and GPS satellites.

Personally identifiable information (PII) is defined by the General Services Administration (GSA) as: "information about a person that contains some unique identifier, including but not limited to name or Social Security Number, from which the identity of the person can be determined" (*42*). Texas Business and Commerce Code 521.002 describes PII as information that "alone or in conjunction with other information identifies an individual," which recognizes that re-identification is a risk as a result of the combination of different available data sets.

Texas Business and Commerce Code 521.051 provides the private citizen legal protection from businesses by placing the risk onto business stakeholders responsible for monetizing the data by gathering, repackaging, and selling the transportation data to various parties. Within 521.051, data providers are required to obtain the consent of the individual to use their PII to obtain a service or product. This consent is often obtained by formal notifications from the telecommunications providers to their customers, or through user agreements built into their smart phones or smart phone apps.

If a business obtains the consent to use the PII, and shares it beyond the confines of its own control as a product or service to other organizations, the Texas Business and Commerce Code 521.052 requires them to alter the PII that it obtained by consent to make the identifying information unreadable or indecipherable. These Texas laws make it clear that the businesses are liable for any combination of PII that leads to the re-identification of the individual (*43*).

Data Ownership

There are widely differing ideas about who owns, or can have access and control of, transportation and vehicle data at different stages of the lifecycle. Presently, the data belongs to those who collect it. Although some data is related to the private citizen or the private citizen's vehicle, it is often not owned by the private citizen. It is owned by the organization that collects it. The public may view the data as owned by the individual, but this ownership often only extends to privacy rights detailing how companies can use it while also preventing PII from being revealed. Therefore, various groups that collect transportation data have data ownership rights to control, sell, and redistribute that same data often as a result of user agreements (*40*).

Several definitions exist for data ownership and all that it entails (*44*):

> Definition 1: *"Entity that can authorize or deny access to certain data, and is responsible for its governance with regard to accuracy, integrity, and timeliness."*

This definition has potentially complex implications for vehicle data ownership. As a result of the recent federal legislation, vehicle owners also have control of their event data recorders (EDR), yet have little responsibility for the accuracy, integrity, or timeliness of this EDR data.

In this sense, OEMs gather these data and govern the accuracy, integrity and timeliness or rely on third party services to take on this role. They in effect become the owners of this vehicle data, which has been obtained as a result of terms and conditions associated with the vehicle purchase (*45*). Recent federal and state laws establish individual vehicle owners as the owners of event data recorder (EDR) information. The recent Federal Driver Privacy Act of 2015 established individual vehicle owners as also owning EDR data and details how the government is reluctant to impede the flow of data necessary to develop connected and automated vehicle systems: "The term [EDR] should not be interpreted as to burden unnecessarily the development of advanced vehicle safety technologies, including autonomous vehicles. The committee contemplates that the EDR would be discrete from any devices and functions used for the operation of such vehicles" (*46*).

The state of Texas has enacted its own version of the Federal EDR law. Established under the 79th Legislature in 2005, HB 160 provides for the owner to retain possession of the EDR data similar to the Federal law with several important caveats (*47*):

> "Information recorded or transmitted by a recording device may not be retrieved by a person other than the owner of the motor vehicle in which the recording device is installed except:
>
> > (1) on court order;

> (2) with the consent of the owner for any purpose, including for the purpose of diagnosing, servicing, or repairing the motor vehicle;
>
> (3) for the purpose of improving motor vehicle safety, including for medical research on the human body's reaction to motor vehicle accidents, if the identity of the owner or driver of the vehicle is not disclosed in connection with the retrieved information; or
>
> (4) for the purpose of determining the need for or facilitating emergency medical response in the event of a motor vehicle accident."

Federal Law does not provide for the use of EDR data in medical research beyond the owner's consent, and it does not provide for the transmission of this data to a central location beyond the vehicle's boundary. These two additions turn over the anonymized EDR data to medical research and PII to emergency dispatch in the event of the incident, regardless of the vehicle owner's consideration. Texas law does not consider non-EDR telematics data and the transfer of this data within the scope of the law.

The ability to deny access to telematics data is conferred upon both the vehicle owner and the automotive manufacturers. Automotive manufacturers rely on terms and conditions associated with the vehicle purchase agreement to establish their control and access to individual vehicle telematics data. Automotive manufacturers also rely on third party service providers, such as OnStar, to provide assistance services at an extra monthly fee. The vehicle owner may elect to not pay for the OnStar subscription service, which in effect turns off the vehicle telematics data transfer associated with this subscription service. This renders both the automotive manufacturer and the vehicle owner as entities who can authorize or deny access to the set of telematics data within a vehicle. Since a private citizen needs to be able to purchase the car by necessity, they do not really have ownership as to whether they transfer telematics data to the automotive manufacturer.

> Definition 2 (*48*): *"Data ownership is the act of having legal rights and complete control over a single piece or set of data elements. It defines and provides information about the rightful owner of data assets and the acquisition, use and distribution policy implemented by the data owner."*

In this definition, it is not just the sourcing of the data or even possession that defines data ownership. Rather, it is the control over data distribution, acquisition, and use that determine ownership. In the case of data aggregator services, multiple sources of data are collected from all manner of inputs as a result of data sharing agreements with various entities. As stewards of the data, they are responsible for anonymizing the data source before using it to prevent the data from being turned into PII and for establishing user agreements once the data has been processed to ensure no one is using it in ways that expose it to re-identification. The source of the data through user agreements establishes what data aggregators can do with it, but once they have obtained it, they become responsible for governing how it is used, distributed, and acquired.

In both definitions it becomes clear that terms and conditions, and user agreements establish the underlying data ownership clauses or ground rules for how data and information is governed, processed, shared, and applied between individuals, the public, and private sector entities. Figure 5 provides a high level view of how vehicle data flows out from a vehicle and relies extensively on user agreements to facilitate data ownership and governance in the marketplace.

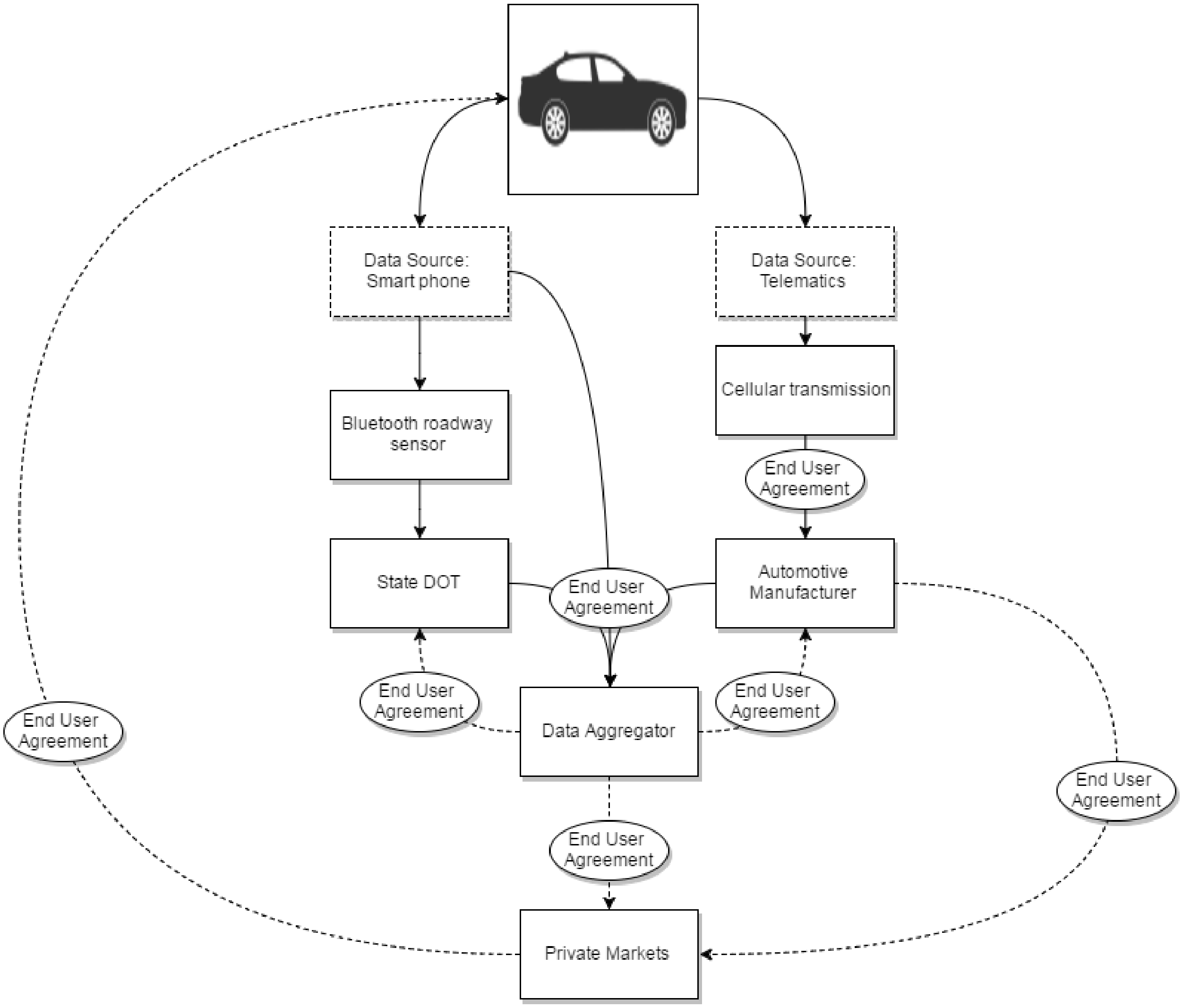


Figure 5. Vehicle Data Transfer and Ownership.

Data aggregators in the private sector procure all manner of vehicle-based and vehicle-oriented data for combination and repackaging into a product or service. For example, data aggregators will procure Bluetooth speed data sets from sensors in traffic intersection signals and along the roadway from the public sector and then match it with telematics data from auto manufacturers and freight fleet management systems. This combined data set is used to create a data product depicting a more detailed real time vehicle speed that can then be purchased by the public and private sector for congestion analysis purposes (*45*). This information may also end up back in the data aggregators own application, which vehicle owners or auto manufacturers may use to assist with navigation around traffic events.

The degree to which the public sector and private sector collect and distribute vehicle data is in flux. As a result of connected vehicle developments, the future is uncertain for how vehicle data will be used and distributed between public and private sector participants.

Liability

Liability is closely tied to ongoing developments in technological innovation, especially as it relates to what is considered PII, or “linked” to PII. The public and private sector are liable in how they govern and protect their data assets from damages associated with data breaches and cyber-attacks, which have far ranging security implications that can potentially affect the daily functions of a transportation network. Organizations responsible for collecting and governing the data from roads, ports, and public transit may be held liable for damages resulting from a data breach or cyber-security attack. The costs include damaged transportation assets, regulatory penalties, and legal costs to minimize the impact of a data security breach on customers, employees, and the transportation asset. There are also questions on the extent to which a governmental organization bears responsibility for the shape and use of the transportation data it provides to third party services and the public.

Federal regulations and lawsuits serve to shape the debate and establish policy governing the extent to which public and private organizations are liable for formatting, sharing, and guarding data related to personally identifiable information.

Vehicle location data is important for transportation agencies because it best accounts for transportation network demands and needs so that agencies can better plan and manage the transportation assets under their care. In Texas, under the 80th legislature HB 2278 in 2007, Business and Commerce Code was established to require businesses to implement reasonable procedures to protect unlawful use or disclosure of PII. The act also requires businesses to destroy PII records that are not scheduled to be retained by the business. This act requires businesses to notify private citizens following the breach of security of PII data within certain cost constraints (a cost $250,000 limit) (*43*).

There are several considerations an organization may take when it comes to ways to mitigate potential transportation data liability:

- Clear and unambiguous terms of use help address liability with regard to data by establishing what a violation of the data use would be (*49*). This may include unauthorized combination of data sets in ways that create increased risk of re-identification of anonymized data.
- Government agencies interested in sharing transportation data through open access platforms for public consumption may benefit from defining “data to be released” to not include information protected by privacy, security, and accessibility laws (*50*).

- Agencies are also liable in how they format and provide data (*51*). For example, information available for public consumption by public agencies must be done in a manner that meets requirements under the Americans with Disabilities Act.
- Liability associated with anonymous geolocation data becoming PII when merged with published data sets in what is known as the mosaic effect (*52*).
- Establishment of network security requirements, employee training, privacy and network policies and procedures, and data breach/cyber-security attack response planning will help in further reducing risks and liability associated with transportation data and access.

Public Perception

Public perception of transportation data management is closely associated with many phases of the data management life cycle. The Pew Research Center found that "68 percent of internet users believe current laws are not good enough in protecting people's privacy online." Pew also found that young adults prioritize privacy issues higher than many of their elders with many taking efforts to protect their privacy by removing their names from tagged photos, and taking steps to mask their identity. Nearly 75 percent of Americans believe it is very important to be in control of their personal information (*53*).

For example, the public perception of connected vehicle technologies have a role in the development of new transportation data sets. Connected vehicle data can include speed, heading, temperature, tire gauge sensors, seat belt engaged sensors, and other internal sensors layered into a vehicles operating system. The majority of drivers consider electronic monitoring of their driving a violation of privacy (*54*). Driving this consideration are concerns about how use of data on travel routes and stops could be embarrassing and harmful if disclosed to third parties with access to the data resulting in a variety of damages including commercial misuse, public corruption, and identity theft.

Given the lack of public trust in data collection, sharing, and security, as noted in several studies and polls (Pew, Politico/MorningConsult poll), there is support for new U.S. Privacy Laws and limits on data retention. Americans favor limits on how long the records of their activity are stored.

Security

Data security refers to the protective measures applied to private data sets in order to prevent unauthorized access to IP addresses, whether it be through computers, databases, websites, mobile devices, or vehicles (*55*). According to the Storage Networking Industry Association (SNIA), storage security represents the convergence of the storage, networking, and security disciplines, technologies, and methodologies for the purpose of protecting and securing digital assets (*56*). Data storage security is also considered as "a wide-ranging area that covers everything from legal compliance, through preparedness for e-discovery requests to user access control and the physical security of data storage" (*57*). It can be a group of parameters and settings that make storage resources available to authorized users and trusted networks.

When data are secure and appropriately regulated, there is greater trust and confidence in its use. Data security issues have received increased attention as a result of data breaches that have become a regular thread in the news. CBS' 60 Minutes has reported that the lack of any overarching federal legislation on data security increases the pressure for individual states and state lawmakers to address data security (*58*). Studies shows that a majority of Americans (64 percent) have personally experienced a major data breach, and relatively large shares of the public lack trust in key institutions – especially the federal government and social media sites – to protect their personal information (*59*). Laws typically cover data security by requiring public and private organizations to apply data security measures such as:

- Breach notification.
- Backups.
- Masking.
- Erasure.
- Encryption.
- Access authentication.
- Clearly defined privacy rights.

By clearly defining data privacy rights, any data breach can be linked to violation of privacy rights and prosecuted accordingly. California enacted the first data breach notification law in 2002, and several states have followed suit. Data security laws covering other topics are also increasing across the states. For example, 31 states have established laws regulating the secure destruction of personal information.

Within these states passing or considering new data security laws, cross-cutting topics emerge like data ownership. For example, Massachusetts passed a law requiring organizations to maintain data security programs with specific requirements that include overseeing third-party service providers, conducting risk assessments, and enforcing violation of security policies. New York State passed a law in 2014 (A.10190) similar to the Massachusetts law with one distinction: it set up separate requirements for data owners and data maintainers or third party services that aggregate and maintain computerized personal information. In addition to requirements for data owners, these third party service providers must also:

- Secure user authentication protocols.
- Secure access control measures that assign unique IDs and passwords to each person with access to systems.

- Encrypt personal information that travels across public networks or is transmitted via wireless.
- Monitor systems for unauthorized use of or access to personal information.
- Encrypt information stored on portable devices.
- Implement appropriate firewall protections and operating system patches.
- Implement security software that receives regular updates.
- Implement security education and training (*58*).

As a result of this distinction of data owner versus data maintainer, in New York third party data service providers, such as INRIX, must follow even more stringent data security requirements to handle PII data sourced from probe data in vehicles.

Data security and breach notifications are based on exposure of PII and defining what PII is. The definition of PII differs across states but is basically an individual's first name or first initial and last name plus one or more of the following data elements:

- Social security number.
- Driver's license number or state-issued identification card number.
- Account number, credit card number or debit card number combined with any security code, access code, PIN or password needed to access an account (*60*).

As denoted in Table 3, data security breach definitions follow the basic tenet of unauthorized acquisition of PII across most states.

Table 3. Data Security Breach Definitions across States (xxv).

Arizona	"Security Breach" means an unauthorized acquisition of and access to unencrypted or unredacted computerized data that materially compromises the security or confidentiality of personal information maintained by a covered entity as part of a database of personal information regarding multiple individuals **and that causes or is reasonably likely to cause substantial economic loss to an individual.**
Texas	"Security Breach" means unauthorized acquisition of **computerized data that compromises the security, confidentiality or integrity of sensitive personal information, including data that is encrypted** if the person accessing the data has the key required to decrypt the data.
California	"Security Breach" means an **unauthorized acquisition of computerized data that compromises the security, confidentiality or integrity of personal information maintained by a covered entity**.

Some states, such as Arizona, also add a qualifier that the PII obtained in the data breach must also cause or potentially cause substantial economic loss to an individual in order to warrant a data breach notification. Given that breach notification laws have different obligations and requirements across states, there is a heightened risk that multistate organizations and third party transportation data service providers will have to contend with potentially conflicting data security obligations in the future as a result of the United States not having an overarching data security framework in place.

It is important to follow cybersecurity-related bills enacted during the 83rd Legislative Session in 2013 that affect how agencies and educational institutions develop and report information security plans (*61*), notably Senate Bills 1597 and 1134. Senate Bill 1597 requires that each state agency submit a security plan to DIR by October 15 of each even-numbered year. Senate Bill 1134 requires that DIR develop strategies and a framework for the securing of cyber infrastructure by state agencies.

Standards and Data Quality

Data standards ensure high quality and high value data and are important in all phases of the data management lifecycle. Quality assurance and quality control (QA/QC) is the process used to discover inconsistencies and other anomalies in the data, as well as performing data cleaning activities to improve data quality. It can be applied to the first-hand data collected by the transportation agencies, or the data purchased from the private companies. Quality ensures the data was collected correctly and could be used to generate meaningful results. While significant data have been generated in recent years, one study suggests that “the utilization and operation of the data is an increasingly difficult task since the data are collected with different levels of accuracy and resolution, and data formats are incompatible. Furthermore, the problem worsens as the amount of data continues to grow. The quality of data in data collection, operation, and management efforts has resulted in the underutilization of data and increased utilization costs” (*62*). This study only addresses questions regarding real-time travel information. There has not been any establishment of data quality standards across the whole transportation system.

Significant human and system resources are consumed in the collection, manipulation, and dissemination of data, so it is essential that the most effective use of public funds is achieved through appropriately directed attention to data quality and the procedures to realize quality. Research is uncovering the need for standards in reporting data to help allow for meaningful comparisons, and exploring the role of big data in informing policy decisions.

Often local, regional, and state agencies work independently to collect the same type of data for different projects. Because there is a lack of formal guidance, it has been challenging to assemble data collected by different agencies into compatible, standardized formats accessible from a single location.

For instance, the MPOs and TxDOT district offices collect and maintain their own pedestrian and bicyclist counts. However, the data are coded and stored in different formats. This results in extra work when the state tries to centralize all the data from different sources and convert them into the same format. A potential solution is to create state standards for all future collection of pedestrian and bicycle data with a reference to the Coding Nonmotorized Station Location Information in the *2016 Traffic Monitoring Guide Format*. In this case, a standardized collection methodology will reduce the amount of time and resources needed for all users to access data, and encourages interagency partnerships.

Improvements in traffic data collection technology have allowed states to improve their data collection processes and to streamline QA/QC procedures. In the field of real-time traffic monitoring and control, data users focus on traffic management and provision of traveler information. Data uses are considered real-time with some agencies also beginning to use historical real-time data to provide additional value to traveler information. Data quality checks are mostly run through field data collection hardware and software. Field hardware and software failures are common. This field uses equipment that could be considered a pre-cursor to automated and connected vehicle systems. For example, many of the sensors and intelligent transportation system components like CCTV, radar, and microwave sensors capable of detecting speeds and transmitting information quickly to traffic management centers showcase some of the difficulties that connected vehicle sensors will encounter in all types of weather and operating conditions present on road systems across a given calendar year (*63*).

In the field of air quality, data collection requirements for mobile source emissions stem from the Clean Air Act Amendments of 1990 (CAAA) and the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA) (*64*). These mobile source emission estimates are based on standards, quality control, and quality assurance associated with planning data collection requirements that have evolved over time with original estimates tied to traffic volume counts on roadways.

Congestion management planning efforts and traffic speed data have recently been standardized and inserted into emission models for use in air quality determinations. Proper use of data is ensured through the use of standards, whether imposed structurally through the design of a data program or in a regulatory manner. One benefit to employing such principles is that data is collected once and used to support multiple purposes, and is of the quality necessary to maximize its use in the decision-making process.

Vehicle probe data provides the vehicle's current position, motion, and time stamp. It is collected from smart phones and sensor-based technologies inside vehicles as they move down a roadway. Vehicle probe data supports government services that help improve road operations, planning, maintenance, and traveler information. The U.S., EU, and Japan entered into a vehicle probe data collection partnership to make an attempt at standardizing probe-data-enabled applications through the Society of Automotive Engineers, ITS SPOT data in Japan, and the Cooperative Awareness Message in Europe. The top three findings from this effort indicated that security requirements, privacy policies (including anonymous data collection and voluntary opt-in applications/services), and data ownership/data rights were the primary challenges to achieving any sort of standardized probe data solution that can work in all three regions (*65*).

Policy Implications

In transportation, data is used to assess alternatives, weigh tradeoffs, evaluate performance, and inform travel behavior. Public support for privacy laws and limits on data retention will impact transportation data management lifecycles. Also, anticipating and understanding Big Data is not only a necessity for innovative solutions to policy problems, but it can also bring about the more efficient allocation of public funds. Passive data collection from probes, GPS, Bluetooth sensors, mobile devices, and cameras can replace traditional travel survey methods, reducing public agency costs. Similarly, insights gained about travel behavior from Big Data sources can enable the more effective use of public funds. For example, analysis of the detailed travel behavior information gathered from GPS and mobile devices can serve to better prioritize traffic management projects.

As transportation organizations work with more stakeholders and external partners to incorporate them into decision making, planning, and operations, there is an increased pressure to also share data. Shared data can help improve decisions since agencies/researchers will be able to obtain a more comprehensive picture of the impacts their decisions have based on contributions of new data sets from a wider variety of sources, both internally and externally. Open sharing of information and the release of information via relevant agreement must be balanced against the need to restrict the availability of classified, proprietary, and sensitive information.

While policy makers should be aware of the opportunities presented by Big Data, it should not be mistaken as a replacement for more traditional research activities. Big Data does not equal whole data. Such things as vehicle-based data sources open new avenues of business development and scientific exploration, and improve shared data set values. The potential to merge newly collected transportation data with older data sets in new and innovative combinations in order to improve future predictions and drive new business solutions is why Big Data has important implications for what data to store and archive, and in what format.

Potential ways to enhance data storage security include data classification and encryption. It is important to understand what data need to be protected and to create a “Data Classification Policy” to classify data based on sensitivity. It is recommended to create a minimum three levels of data classification (e.g., restricted, confidential/private, public). There are many ways to encrypt data, and it should be done before sharing sensitive data over untrusted networks. The key is to use strong encryption and proper key management.

The importance of data in this era of data-driven decision making, the swift increase in the volume of data due to improved collection methods, new uses such as automated and connected vehicles, and increased interest on the part of the public in factors underlying decision making, suggests that policymakers may have an interest in understanding and addressing the quantity, quality, creation, collection, storage, retention, privacy, security, and availability of transportation data across agencies. Data-driven insight can serve to inform policy decisions at all levels, helping to conserve limited public funds and ensure the most efficient and effective use of transportation systems.

References

1 Vandervalk, A. Turning Data into Information for Transport Decision Making. *European Transport Conference 2012*, 2012.

2 Bureau of Transportation Statistics. Planning and Design of Data Collection Systems. In *BTS Statistical Standards Manual October 2005*, August 2005. https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/subject_areas/statistical_policy_and_research/bts_statistical_standards_manual/html/chapter_02.html. Accessed Apr. 09, 2017.

3 Mohaddes, A., and Sweatman, P. *Transportation Research Circular E-C208: Transformational Technologies in Transportation: State of the Activities.* Transportation Research Board, Washington, D.C., 2016. https://dx.doi.org/10.17226/23599. Accessed Apr. 9, 2017.

4 National Conference of State Legislatures (NCSL). Automated License Plate Readers: State Legislation. November 2015. http://www.ncsl.org/research/telecommunications-and-information-technology/2014-state-legislation-related-to-automated-license-plate-recognition-information.aspx. Accessed December 21, 2017.

5 http://onlinepubs.trb.org/onlinepubs/conferences/2013/MPO/Slater.pdf.

6 Bureau of Transportation Statistics. *Guide to Good Statistical Practice in the Transportation Field*. May 2003. https://www.bts.gov/sites/bts.dot.gov/files/legacy/publications/guide_to_good_statistical_practice_in_the_transportation_field/pdf/entire.pdf. Accessed December 21, 2017.

7 Cuellar, R., Bricka, S. G., and Moran, M. M. *Big Data Scan*. TTI/SRP/15/161505-1. Texas A&M Transportation Institute, September 2015. http://static.tti.tamu.edu/tti.tamu.edu/documents/161505-1.pdf. Accessed December 21, 2017.

8 Cloud computing. April 2017. https://en.wikipedia.org/wiki/Cloud_computing. Accessed April 10, 2017.

9 LaChapelle, C. The Cost of Data Storage and Management: Where Is It Headed in 2016?. *The Data Center Journal*, March 10, 2016. http://www.datacenterjournal.com/cost-data-storage-management-headed-2016/. Accessed April 10, 2017.

10 Lei, H., Xing, T., Taylor, J. D., and Zhou, X. Monitoring Travel Time Reliability from the Cloud. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2291, 2012, pp. 35-43. http://dx.doi.org/10.3141/2291-05.

11 Vandervalk, Anita, *Turning Data into Information for Transport Decision Making,* Cambridge Systematics. Association for European Transport and Contributors, Glasgow, UK, 2012. https://aetransport.org/public/downloads/zYJtR/5594-5218a23756fcf.pdf

12 National Household Travel Survey. No Date. https://www.nationalhouseholdtravelsurvey.com/. Accessed April 09, 2017.

13 Cuellar, R., Bricka, S. G., Moran, M. M., *Big Data Scan*, Texas A&M Transportation Institute, Report No. TTI/SRP/15/161505-1, September 2015. https://static.tti.tamu.edu/tti.tamu.edu/documents/161505-1.pdf

14 Miller, Matt. (2015 April 2). Telephone Discussion with Stephen Lockwood, PB Consult.

15 Chen, M. C., Chen, J. L., and Chang, T. W. Android/OSGi-based vehicular network management system. *Computer Communications*, Volume 34, No. 2, 2011, pp. 169-183.

16 Pack, M. L. and Ivanov, N., *NCHRP Synthesis 460: Sharing Operations Data among Agencies, a Synthesis of Highway Practice*. Transportation Research Board, Washington, D.C., 2014. http://dx.doi.org/10.17226/22372.

17 Dell EMC. Dell EMC Glossary: Data Archiving. No date. https://www.emc.com/corporate/glossary/data-archiving.htm. Accessed April 4, 2017.

18 Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT). Information Management > Data Archive. No date. http://www.itscosts.its.dot.gov/ITS/benecost.nsf/SingleLink?OpenForm&Tax=Intelligent+Transportation+Systems+Information+Management+Data+Archive&Location=Cost. Accessed April 7, 2017.

19 PanolaWatchman.com. *TxDOT launches new web application to streamline crash data reporting*. October 18, 2011. http://www.news-journal.com/panola/news/txdot-launches-new-web-application-to-streamline-crash-data-reporting/article_93a91ed6-4fff-5758-8e98-5de97ad6fe28.html. Accessed December 2, 2013.

20 Texas Department of Transportation. CRASH. No Date. http://www.txdot.gov/government/enforcement/crash-system.html. Accessed December 2, 2013.

21 Texas Department of Transportation. Crash Data Analysis and Statistics. 2016. http://www.txdot.gov/government/enforcement/crash-statistics.html. Accessed April 4, 2017.

22 Auer, A., Feese, S., and Lockwood, S. *History of Intelligent Transportation Systems*. FHWA-JPO-16-329. FHWA U.S. Department of Transportation Intelligent Transportation Systems Joint Program Office, May 2016.

23 Turner, S. *Guidelines for Developing ITS Data Archiving Systems*. Report 2117-3. Project Number 0-2127. Texas Department of Transportation and U.S. Department of Transportation, 2001. http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/2127-3.pdf. Accessed April 7, 2017.

24 Archived Data Management Systems: Cross Cutting Study, *Linking Operations and Planning Data*. Publication FHWA-JPO-05-044. FHWA, U.S. Department of Transportation, 2005. https://ntl.bts.gov/lib/jpodocs/repts_te/14128/14128.pdf. Accessed December 21, 2017.

25 Auer, A., Feese, S., and Lockwood, S., *History of Intelligent Transportation Systems*. Booz Allen Hamilton for the U.S. Department of Transportation Intelligent Transportation Systems Joint Program Office. FHWA-JPO-16-329. May 2016. https://rosap.ntl.bts.gov/view/dot/30826

26 Hitachi Data Systems. *The Internet on Wheels and Hitachi, Ltd*. December 2015. https://www.hds.com/en-us/pdf/white-paper/hitachi-white-paper-internet-on-wheels.pdf. Accessed May 30, 2017.

27 McFarland, M. Your Car's Data May Soon Be More Valuable Than the Car Itself. *CNNMoney*, February 7, 2017. http://money.cnn.com/2017/02/07/technology/car-data-value/. Accessed May 31, 2017.

28 McDonald, D. (2015, June 10). *Managing Open Transportation Data at the U.S. Department of Transportation*. June 10, 2015. http://www.datacommunitydc.org/blog/2015/6/managing-open-transportation-data-at-the-us-department-of-transportation. Accessed May 31, 2017.

29 Ongoing Operations. *Data Destruction – Protecting private data when moving to or from a cloud service*. December 10, 2012. https://ongoingoperations.com/2012/12/10/data-destruction-protecting-private-data-cloud-service/. Accessed May 30, 2017.

30 Violino, B. The in-depth guide to data destruction. *CSO*, February 6, 2012. http://www.csoonline.com/article/2130822/it-audit/the-in-depth-guide-to-data-destruction.html. Accessed May 30, 2017.

31 National Conference of State Legislatures (NCSL). Data Disposal Laws. December 2016. http://www.ncsl.org/research/telecommunications-and-information-technology/data-disposal-laws.aspx. Accessed May 30, 2017.

32 Institutional Review Board (IRB). Retention of Research Records and Destruction of Data. No Date. http://www.virginia.edu/vpr/irb/sbs/resources_guide_data_retention.html. Accessed May 30, 2017.

33 McKinsey & Company. (2016, March). Car data: paving the way to value-creating mobility. March 2016. http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/creating-value-from-car-data. Accessed May 31, 2017.

34 Beresford, A.R., and Stajano, F. 2003. Location Privacy in Pervasive Computing. *IEEE Pervasive Computing*, Volume 2, No. 1, 2003, pp. 46–55. http://dx.doi.org/10.1109/MPRV.2003.1186725.

35 de Montjoye, Y. A., Radaelli, L., Singh, V. K., and Pentland, A. S. Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata. *Science*, Vol. 347, No. 6221, 2015, pp. 536–539. http://science.sciencemag.org/content/347/6221/536.full. Accessed December 21, 2017.

36 Markey, E. *Tracking and Hacking: Security and Privacy Gaps Put American Drivers at Risk*. U.S. Senator Edward Markey's Office, Washington, D.C., 2015.

37 Auto Alliance Poll. *Consumers Still Want to be in the Driver's Seat, Self-Driving Cars Raise Concerns*. 2013. http://www.autoalliance.org/INDEX.CFM?OBJECTID=156688B0-CD5D-11E2-8898000C296BA163.

38 American Automobile Association. *The Connected Car: It's Your Vehicle, But Is It Your Data?*. 2014.

http://midatlantic.aaa.com/~/media/Files/Connected%20Car/The%20Connected%20CarIts%20Your%20Vehicle%20But%20is%20it%20Your%20DataTO%20PRINT.ashx.

39 U.S. Department of Transportation (US DOT). *Privacy Impact Assessments*. March 2012. https://www.transportation.gov/individuals/privacy/privacy-impact-assessments. Accessed April 10, 2017.

40 Zmud, J., Tooley, M., and Miller, M. *Data Ownership Issues in a Connected Car Environment: Implications for State and Local Agencies*. TTI/SRP/16/165604-1. Texas A&M Transportation Institute, College Station, T.X., 2016. https://static.tti.tamu.edu/tti.tamu.edu/documents/165604-1.pdf. Accessed December 21, 2017.

41 Kaye, K. The $24 Billion Data Business that Telcos Don't Want to Talk About. *AdAge*, 2015. http://adage.com/article/datadriven-marketing/24-billion-data-business-telcos-discuss/301058/. Accessed December 21, 2017.

42 General Services Administration. *GSA Rules of Behavior for Handling Personally Identifiable Information (PII)*. 2014. https://gsa.gov/portal/getMediaData?mediaId=199847. Accessed December 21, 2017.

43 Texas State Legislature. *Unauthorized Use of Identifying Information*. Business and Commerce Code; Title 11. Personal Identity Information; Subtitle B. Identity Theft; Chapter 521; Subchapter A. General Provisions, 2007. http://www.statutes.legis.state.tx.us/Docs/BC/htm/BC.521.htm. Accessed December 21, 2017.

44 Business Dictionary. *Definition of Data Ownership*. 2017 http://www.businessdictionary.com/definition/data-owner.html. Accessed December 21, 2017.

45 Alliance of Automobile Manufacturers and Association of Global Automakers. *Consumer Privacy Protection Principles: Privacy Principles for Vehicle Technologies and Services*. 2014. https://autoalliance.org/connected-cars/automotive-privacy-2/principles/.

46 U.S. Government Publishing Office. *Senate Report 114-147 – Driver Privacy Act of 2015*. 114th Congress, 2015-2016. https://www.congress.gov/congressional-report/114th-congress/senate-report/147/1. Accessed December 21, 2017.

47 Texas State Legislature. *Section 547.615-Recording Devices*. Title 7. Vehicles and Traffic. Subtitle C. Rules of the Road. Chapter 547. Vehicle Equipment. Subchapter A. General Provisions, 2005. http://www.statutes.legis.state.tx.us/docs/TN/htm/TN.547.htm#547.615. Accessed December 21, 2017.

48 Techopedia. *Data Ownership: Definition: What Does Data Ownership Mean?*. 2017. https://www.techopedia.com/definition/29059/data-ownership. Accessed December 21, 2017.

49 The Hartford Insurance Company. *What is Data Breach Insurance?*. https://www.thehartford.com/data-breach-insurance. Accessed March 8, 2017.

50 Sunlight Foundation. *Open Data Policies and Implementation: Frequently Asked Questions*. https://sunlightfoundation.com/policy/opendatafaq/#liability. Accessed March 8, 2017.

51 Bowser, A., Wiggins, A., and Stevenson, R. D. Data Policies for Public Participation in Scientific Research: A Primer. *Report from the DataONE Public Participation in Scientific Research Working Group. Albuquerque, N.M. (13pages)*. 2013. http://www.birds.cornell.edu/citscitoolkit/toolkit/policy/Bowser%20et%20al%202013%20Data%20Policy%20Guide.pdf. Accessed December 21, 2017.

52 Mazmanian, Adam. *The mosaic effect and big data*. FCW, The Business of Federal Technology. 2014. https://fcw.com/articles/2014/05/13/fose-mosaic.aspx. Accedded February 15, 2018.

53 Olmstead, Kenneth and Smith, Aaron. *Americans and Cybersecurity*. Pew Research Center. 2017. http://www.pewinternet.org/2017/01/26/americans-and-cybersecurity/. Accessed February 15, 2018.

54 Cregger, J., Brugeman, V. S., Wallace, R. *Public Perceptions of Connected Vehicle Technology*. Center for Automated Research, Michigan Department of Transportation, 2012. http://www.cargroup.org/wp-content/uploads/2017/02/PUBLIC-PERCEPTIONS-OF-CONNECTED-VEHICLE-TECHNOLOG.pdf. Accessed December 21, 2017.

55 Techopedia. *Data Security: Definition: What Does Data Security Mean?*. 2017. https://www.techopedia.com/definition/26464/data-security. Accessed December 21, 2017.

56 Hibbard, E. A., and Austin, R. *Storage Security Professional's Guide to Skills and Knowledge*. SNIA, 2008. www.snia.org/ssif. SNIA. Accessed August 18, 2014.

57 Data storage security: What it is and the key components of a storage security strategy. *ComputerWeekly.com*, 2010. http://www.computerweekly.com/feature/Data-storage-security-What-it-is-and-the-key-components-of-a-storage-security-strategy. Accessed April 10, 2017.

58 Lovells, H. Outlook for State Data Security Laws: More than Breach Notification. *IAPP*, 2014. https://iapp.org/news/a/outlook-for-state-data-security-laws-more-than-breach-notification/. Accessed December 21, 2017.

59 Olmstead, K., and Smith, A. Americans and Cybersecurity. *Pew Research Center*, 2017. http://www.pewinternet.org/2017/01/26/americans-and-cybersecurity/. Accessed April 10, 2017.

60 Mintz Levin. *State Data Security Breach Notification Laws*, 2017. https://www.mintz.com/newsletter/2007/PrivSec-DataBreachLaws-02-07/state_data_breach_matrix.pdf. Accessed December 21, 2017.

61 Texas Department of Information Resources (DIR). *Information Security*. No Date. http://dir.texas.gov/View-About-DIR/Information-Security/Landing.aspx. Accessed April 10, 2017.

62 Ahn, K., Rakha, H., and Hill, D. *Data quality white paper*. FHWA-HOP-08-038. US Department of Transportation, Federal Highway Administration. 2008

63 Turner, S. Defining and Measuring Traffic Data Quality White Paper. *Proceedings of the Traffic Data Quality Workshop, Washington, D.C.* 2002. https://ntl.bts.gov/lib/jpodocs/repts_te/13767.html. Accessed December 21, 2017.

64 Karash, K. H., and Schweiger, C. *Identification of Transportation Planning Data Requirements in Federal Legislation*. DOT-T-94-21. U.S. Department of Transportation. FHWA, 1994. https://ntl.bts.gov/DOCS/tmi.html. Accessed December 21, 2017.

65 United States-Japan-European Union Probe Data. FHWA-JPO-14-155. FHWA. U.S. Department of Transportation, No Date. https://www.its.dot.gov/factsheets/pdf/ITS%20JPO_FS_US-Japan_Probe_Data.pdf. Accessed December 21, 2017.