THE UNIVERSITY OF TEXAS AT AUSTIN
CENTER FOR TRANSPORTATION RESEARCH

0-6697-CTR-P1

FREIGHT DATA ARCHITECTURE BUSINESS PROCESS, LOGICAL DATA MODEL, AND PHYSICAL DATA MODEL

Dr. C. Michael Walton
Dan Seedah
Alejandra Cruz-Ross
Dr. William J. O'Brien
Peter La Fountain
Prateek Agarwal
Bharathwaj Sankaran

TxDOT Project 0-6697-CTR: Integration of Data Sources to Optimize Freight Transportation in Texas

JANUARY 2014; PUBLISHED SEPTEMBER 2014

Performing Organization:	Sponsoring Organization:
Center for Transportation Research The University of Texas at Austin 1616 Guadalupe, Suite 4.202 Austin, Texas 78701	Texas Department of Transportation Research and Technology Implementation Office P.O. Box 5080 Austin, Texas 78763-5080
Performed in cooperation with the Texas Department of Transportation and the Federal Highway Administration.	

Table of Contents

INTRODUCTION

The passage of the Moving Ahead for Progress in the 21st Century Act (MAP-21), enacted in 2012, provides guidance and regulation and creates *momentum* to improve the condition and performance of the national freight network, and support investment in freight-related transportation projects. MAP-21's stipulations expand agencies' interest in freight initiatives and modeling within statewide planning efforts, particularly the evaluation of current and future freight transportation capacity necessary to ensure freight mobility. For example, in the case of project authorization, MAP-21 specifically states that a description of how the project will improve the efficient movement of freight can be created through data and information that support quantitative analysis. However, the understanding of freight demand and the evaluation of current and future freight transportation capacity are not only determined by robust models, but are critically contingent on the availability of accurate data. In this regard, insufficient and inferior quality data is the most commonly cited challenge in the development of freight models[1,2,3,4,5,6,7,8].

State departments of transport (DOTs) typically (a) rely on the data compiled and published by federal agencies, such as the Freight Analysis Framework, (b) obtain one of the private commercial sources of data related to freight movements, such as the IHS Global Insight TRANSEARCH database, or (c) collect primary data through interview and surveys of freight stakeholders. Few states, however, collect primary freight data, partly because this can be a costly and time-consuming process. Also, currently few, if any, procedures at the federal, state, or local levels provide specific guidance on the collection of freight data[9, 10].

In November 2011, the Texas Department of Transportation (TxDOT) funded a study to integrate data from multiple sources to optimize freight transportation planning efforts in the state. The Center for Transportation (CTR) study team was commissioned to explore the feasibility of entering into a data-sharing partnership with representatives of the private sector (i.e., shippers, receivers, trucking companies, forwarders, etc.), and obtain sample data that can be used in formulating a strategy for integrating multiple data sources.

To build relationships with the private sector, the study team contacted 493 companies, received 151 responses, and conducted 33 key-person interviews with executive level managers of trucking companies, shippers, airports, logistics companies, rail carriers, and ports. These interviews were preceded by an online survey of 32 companies to establish private-sector willingness to enter into a data-sharing partnership, and assist the study team in understanding their needs and concerns.

This document summarizes the study team's efforts to establish data-sharing partnerships and relay the lessons learned. In addition, it provides information on a prototype freight data architecture and supporting description and specifications that will facilitate the storage and exchange of data through a data sharing partnership with members of Texas's freight community. Final recommendations on who should be responsible for populating and developing the integrated freight system are also made, and list of items to be considered in estimating the cost for developing and maintaining the system are presented.

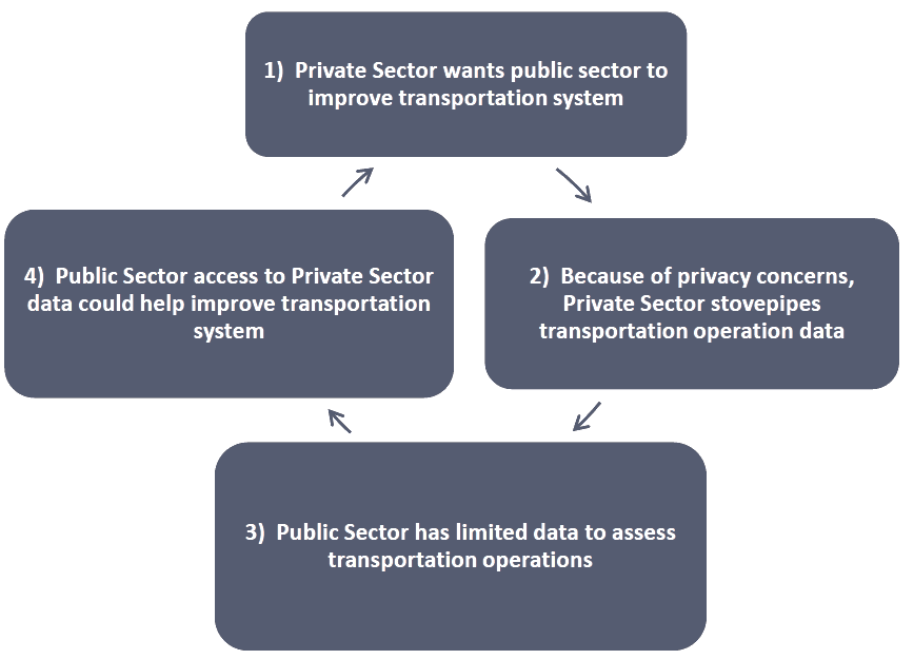
For additional detailed information on the background and supporting work performed, please refer to the Final Report of this study.

WHY DATA PARTNERSHIP?

Traditional primary data collection methods such as roadside intercept surveys, mail-out/mail-back questionnaires, combined telephone-mail-out/mail back questionnaires, driver trip diaries, personal interviews, etc., when done correctly, are regarded as very reliable for obtaining freight data. However, these methods can be costly and generally involve a smaller, more select or targeted sample that may not be appropriate for the population of statewide freight movement databases.

Progress is being made in some states and regions to cost-effectively procure more accurate truck travel data using technological applications instead of the traditional survey methods. In the case of truck data, a number of ITS technologies such as global positioning systems (GPS) are able to collect various truck travel attributes such as routing, time, carrier, and origin and destination information. Current technological innovations provide an opportunity for effective data-sharing partnerships between transportation planning agencies and the private sector like never before.

During the final outreach phase of this project, the study team found that the Transportation Research Board (TRB) was carrying out research on establishing guidelines for data-sharing partnerships between public and private freight stakeholders through its National Cooperative Freight Research Program (NCFRP). The study, *Freight Data Sharing Guidebook*[11], explored some of the barriers to freight data sharing between the private and public sectors. The report establishes as a premise that "planners are only as effective as the quality of their updated information on movement needs." However, data is guarded by firms in the private sector, because it affects their competitive edge in the marketplace. These interactions and interests for data exchange are demonstrated in Figure 1.



Source: Transportation Research Board[11]

FIGURE 1: Establishing the Need for Private Freight Stakeholder Data

In the first instance, private sector stakeholders were found to be reticent to share data because of privacy concerns. This observation was also made in the online survey conducted by the CTR study team. Of the 32 companies that participated in the survey, 22 cited the fear of data being mishandled or being improperly used. If guaranteed that the information will never become public or shared with others, 28 companies cited that they will be more willing to enter into a data-sharing partnership. In the second instance of Figure 1, the TRB report cited that the private sector wishes the public sector could improve the transportation system to allow them to become more competitive. This observation was also made by the CTR study team. When asked what benefits they will seek from a data-sharing partnership, 20 companies answered that addressing current and anticipated transportation issues such as congestion and capacity constraints will be beneficial and 19 companies cited that improving roadway safety and public education on freight will be beneficial.

Based on these observations and previous studies[1, 5], it can be inferred that both public and private sector agencies recognize that effective partnerships are needed to ensure adequate planning and funding of transportation infrastructure at the state and local levels. Despite this recognition, the private sector is reluctant to share data because of privacy concerns, and the transportation planning community continues to struggle to understand the needs of the freight community. Enhanced freight mobility through infrastructure improvements is in the interest of both the private sector and the transportation planning community. In addition, reliable freight data can be valuable to the private sector in informing investment decisions relating to equipment utilization, new markets, and business opportunities[1].

With knowledge of the concerns of the private sector, the CTR study team moved forward with approaching the freight community to establish data-sharing partnerships. This option was analyzed under the hypothesis that a statistically representative sample of Texas freight stakeholders (i.e., shippers, receivers, trucking companies, rail carriers, ports, airports, inland port, logistics companies, freight forwarders, brokers, etc.) can be convinced to enter into a data-sharing partnership with TxDOT. The greatest challenge is getting a foot in the door with firms who are busy, preoccupied, and suspicious of the nature of a pilot project involving data-sharing. Of the 493 companies contacted, 151 expressed an interest in receiving additional information about the project. However, only 33 participated in one-on-one interviews and 3 provided sample data, notwithstanding the lack of a non-disclosure agreement. However, through these efforts, the study team initiated productive relationships with freight stakeholders, and recommends continuing to develop these freight relationships on an ongoing, long-term basis in order to gain trust and secure partnerships. In summary, the following lessons were gleaned from the outreach efforts:

1. Most stakeholders interviewed considered that a partnership would have beneficial outcomes, including addressing any current or anticipated transportation issues such as congestion and capacity constraints, and providing recommendations in the design and development of new infrastructural projects which can impact freight operations.
2. The majority of stakeholders were concerned with the mishandling or improper use of data, the time commitment required in scrubbing and preparing data in-house, and new government regulations and law enforcement measures pertaining to data security.

3. Lightening the information technology (IT) requirements for stakeholders is highly recommended. Stakeholders were more willing to share data when the study team offered to accept data samples in either CSV or Excel formats. In addition, offering to “clean” or scrub the data for them was also welcomed.
4. If guaranteed that the information would never become public, 88% of survey respondents were willing to participate in a data-sharing partnership.
5. None of the respondents interviewed or surveyed are currently participating in a data-sharing partnership.
6. Data variables that stakeholders were willing to share include:
    - Vehicle Type
    - Trip Origin/Destination
    - Number of Trips
    - Load Type (truck load, less than truckload, service)
    - Commodity
    - Mode of Transport,
    - Route Preference (e.g. interstate, toll, etc.),
    - Cargo Weight and Value,
    - Frequency of Trips, and
    - Trip Type (internal, external).
7. Most stakeholders were willing to share data either through an electronic submission via email or via a secured website.
8. A clear non-disclosure contract is required: a written contract would reinforce a sense of trust with freight partners and the sample provided in the *Freight Data Sharing Guidebook* is a good option.
9. Support from trade associations such as Texas Trucking Association (TXTA) was found to be invaluable in the outreach efforts. TXTA published the project information in its weekly newsletter, including a link to the online survey, and invited team members to events, which provided an opportunity to meet potential stakeholders.
10. Participation—secret for some, open for others: Some stakeholders expressed preference for their participation to remain anonymous; others inquired about the possibility of advertising their participation in such data-sharing partnership.
11. Most stakeholders contacted asked to see something more tangible to share with company executives and decision-makers before deciding whether to enter into a data-sharing partnership. Therefore, a built-out demo website (complete with the initial architecture) that demonstrates the integration of public and private datasets could promote confidence in future data partnership efforts.

CONCEPTUAL FREIGHT DATA INTEGRATION ARCHITECTURE

It was determined from the survey of freight stakeholders that electronic file submission (i.e., either by email or directly via a website) is the preferred method for data exchange for any freight data sharing partnership (see Figure 2). It is therefore imperative that any successful freight data sharing architecture have this capability available in addition to any other features such as data security and data privacy that will need to be included. Based on this knowledge and information gleaned during the performance of this study, the study team proposes a conceptual system architecture with the following minimum capabilities as illustrated in Figure 3 and described in detail in subsequent subsections:

1. Integration and use of publicly available data
2. Electronic submission of data by freight data sharing partners
3. Data quality and validation
4. Automated data scrubbing and aggregation
5. Secure data storage and restricted access
6. Value added services through integration into existing TxDOT data systems
7. Data output and analysis tools

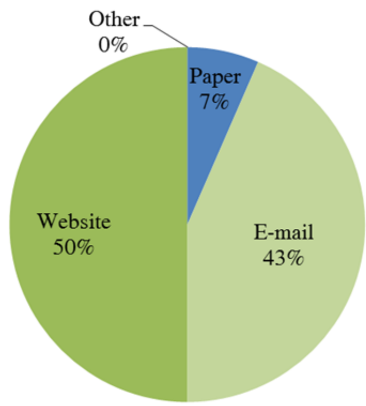


FIGURE 2: Stakeholder Preferred Data Submission Format

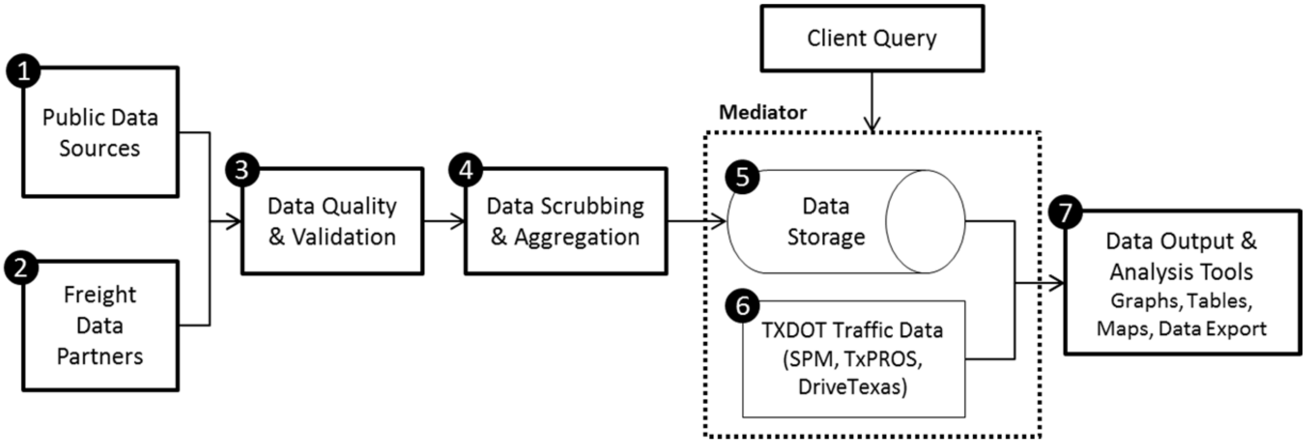


FIGURE 3: Conceptual Architecture for an Integrated Freight Database System

Integration and Use of Publicly Available Data

We determined that a total of 21 public freight databases can be readily utilized in the development of an integrated freight database. The data collected by government agencies and other industry associations usually do not have the level of disaggregation required for county or zip code level planning. Through the use of data fusion and mediation processes, it is possible to integrate these databases into a single system as demonstrated in the prototype system. Data integration will, however, have to be done carefully because of varying naming schemes and reporting methodologies. The mediator architecture section of this tech memo describes in detail a recommended approach for integrating these databases with private sector data.

Electronic Submission of Data by Data Sharing Partners

Electronic data submission involves the submission of data either by e-mail or a web form. Despite electronic data submission being the preferred data exchange medium, phone conversations with potential data sharing partners determined that partners may not have sufficient resources to prepare, examine, or scrub the data into a particular standardized format before submission. Therefore, the receiving agency will need to have in place an infrastructure that facilitates this process. The data sharing partner may only need to submit the data in the format they feel comfortable with and the data exchange system should take care of the rest automatically. The minimal standard that needs to be adhered to by the data sharing partner is the file exchange format (e.g., CSV, Excel) and a partner-specific agreed upon and consistent data layout (e.g., number of columns, variable types).

Not all electronic data will require manual submission. Data from non-intrusive technologies such as radio frequency identification (RFID) tags and geographical information system (GPS) devices can be set up to automatically transmit data to the integrated system. However, privacy concerns will have to be addressed, and data aggregation, scrubbing, and cleaning will still be required before final inclusion into the integrated freight database.

Data Quality and Validation

Ensuring data quality is primary to the success of any data sharing program. Data cleansing or data cleaning is required to ensure that data being stored is accurate, complete, relevant and consistent. Data cleansing can be performed through the provision of standard data dictionary definitions for each data source to be included in the integrated freight database. A system should be in place to detect missing or inaccurate data types and notify the system administrator of any such errors.

Data validity can be checked using the following suggested constraints[12,13,14]:

i) Data-Type Constraints: Values in a particular column must be of a particular data type, e.g., Boolean, numeric (integer or real), date, etc.

ii) Range Constraints: numbers or dates should fall within a certain range set by a maximum or minimum value. This can be performed using the database *check constraint* feature, e.g., setting the figure for average annual daily traffic to not be negative and less than 500,000 for a specific region or area.

iii) NOT NULL Constraints: to prevent null values from being entered into a column

iv) Unique Value Constraints: A field, or a combination of fields, must be unique across a dataset, e.g., using designated roadway names to describe a particular roadway link.

v) Foreign-key Constraints: Also known as referential constraints or referential integrity constraints, these are used in defining required relationships between and within tables, e.g., referencing all NAICS commodity code classifications in multiple tables to a single table containing all NAICS commodity code definitions.

vi) Regular Expression Patterns: These are used in validating text fields if data is required to have a certain pattern, e.g., five digit zip codes (00000).

vii) Cross-field Validation: To ensure that certain conditions that utilize multiple fields must hold, e.g., data uploaded for 2012 cannot contain 2013 data or percentage of truck traffic cannot exceed 100.

viii) Duplicate Data Elimination: elimination of duplicate or already reported data. These will, however, need to be done carefully based on a knowledgeable set of unique value constraints.

In addition to using software to set data constraints, other data quality steps may need to be performed manually by the receiving agency. These steps include checking for data accuracy, data consistency (i.e., comparison with previously reported data), and data uniformity (i.e., units of measurement).

Automated Data Scrubbing and Aggregation

As identified in this study and other related studies[11], the private sector is reluctant to share data for fear of government regulation and industry competition, among other reasons such as resource constraints in making the data available in an acceptable format. Data scrubbing and/or aggregation steps are therefore necessary to address privacy concerns and ensure trust between parties involved in the data sharing partnership. Data scrubbing will remove sensitive data such as company name, customer address, tracking data, specific commodity identifiers, or any other variables that may endanger the competitive advantage of partners. In addition to data scrubbing, data aggregation will consolidate data from multiple sources, making it impossible to trace original sources; aggregation is useful for reporting network- or zone-level information. For data for which scrubbing and aggregation may be insufficient, noise infusion can be applied. Commonly used by the U.S. Census Bureau, noise infusion is a method of disclosure avoidance in which values are perturbed prior to reporting by applying a random noise multiplier to the originally reported magnitude data[15,16].

Secure Data Storage and Restricted Access

Essential to data sharing programs is the ability for the receiving agency to securely store and restrict access to data received from partners. Options include the use of data encryption technologies such Secure Sockets Layer (SSL) encryption during the data submission process and 256-bit Advanced Encryption Standard (AES) encryption for data storage.

In addition, access to raw data should be restricted to users at a level sufficient to examine data integrity but still protect the privacy of partners involved in the data partnership program. Access to output data should also be restricted at different usage levels. For example, agency staff can have access to samples of the prior scrubbed data and the general public can be restricted to summary data aggregated from multiple databases. Exact user policies and roles can be further defined in detail should TxDOT choose to implement the integrated freight database.

Value-added Services through Integration into Existing TxDOT Data Systems

TxDOT currently has available a suite of tools that, when linked into the integrated freight database, will add additional value to the database. Examples of identified TxDOT data tools include the following:

i) TxDOT's Statewide Planning Map (SPM): This contains relevant information on TxDOT on-system roadways. Example data reported by the SPM include current and historic traffic counts, Texas trunk system routes, roadway speed limits, geopolitical boundaries, imagery, and other geospatial and biogeographic data such as watersheds, aquifers, and vegetation [17].

ii) DriveTexas.org Road Construction and Incident Reporting System: DriveTexas.org is another online TxDOT mapping system that can complement the integrated freight database system by providing live traffic feed data. Still in development, the system provides live information on roadway accidents, closures, construction work zones, damages, ice/snow conditions, and other condition types such as special events and parades[18].

iii) Texas Permitting & Routing Optimization System (TxPROS): TxPROS is an online permitting and GIS-based mapping system capable of producing information on the routing of oversize and overweight (OS/OW) loads permitting system. TxPROS data can complement the integrated freight database system by providing routing information, number of permits, vehicle sizes and weights, and frequency of OS/OW trips on various roadways [19].

iv) TxDOT's Statewide Analysis Model (SAM): SAM is a state-of-the-practice multimodal travel model that provides highway traffic forecasts for highway passenger travel and freight transport, intercity and high speed passenger rail ridership, freight rail tonnage and train forecasts, and forecasts of air passenger travel to and from Texas airports. The most recent version, SAM-V3, provides travel forecasts at a level of detail suitable for use in comparative analyses of large scale transportation corridor projects and other large scale investments. The model can also be used to perform analyses of the transportation outcomes and economic impacts of state-level transportation, land use, and economic policy decisions and strategies. It is the primary tool used by TxDOT for evaluating large intercity transportation projects through the state. SAM-V3 already incorporates publicly available data into its models and compatible to proprietary data sources such as Wood and Poole employment forecasts and TRANSEARCH/Global Insight data[20].

Data Output and Analysis Tools

During the workshops conducted as part of this project—which solicited feedback from potential users of the integrated freight database in TxDOT's district and regional offices, MPOs, and city offices—the most popular freight database features requested by participants were

i) a web-based platform with graphical user interface and GIS,

ii) the ability to determine the most effective or shortest routes to transport goods in the State of Texas,

iii) disaggregation of the data at the county- or corridor-level,

iv) show traffic generators (current and future), changes in modes, commodity information (weigh out, cube out), loads locations, OS/OW loads, and networks used, and

v) the ability to determine traffic flow and percentage of truck traffic

The final requirement of this system should incorporate the above-requested features in addition to the following:

vi) Avoid combining data from multiple databases if integration is not possible. Instead, data from each database should be shown separately[21].

vii) Provide users with a summarized view of the data as well as the ability to download the complete queried data.

viii) Display information as graphs and tables where possible.

PROPOSED MEDIATOR ARCHITECTURE

The mediator architecture proposed by the study team and shown in Figure 4 provides a strategy that maintains the integrity of datasets[i], facilitates future updates, and speeds up the processing time of user queries. This architecture was utilized in the development of the prototype system. Datasets used in the development of the prototype system include the Freight Analysis Framework (FAF); Commodity Flow Survey (CFS); TxDOT Highway Performance Monitoring System (HPMS); three private datasets; and National Corridors Analysis and Speed Tool (NCAST).

[i] The term *dataset* is used henceforth to refer to a single freight database table used in the development of the prototype integrated system. A single dataset is made up of columns that represent variables in that dataset and rows that represent records contained in the dataset.

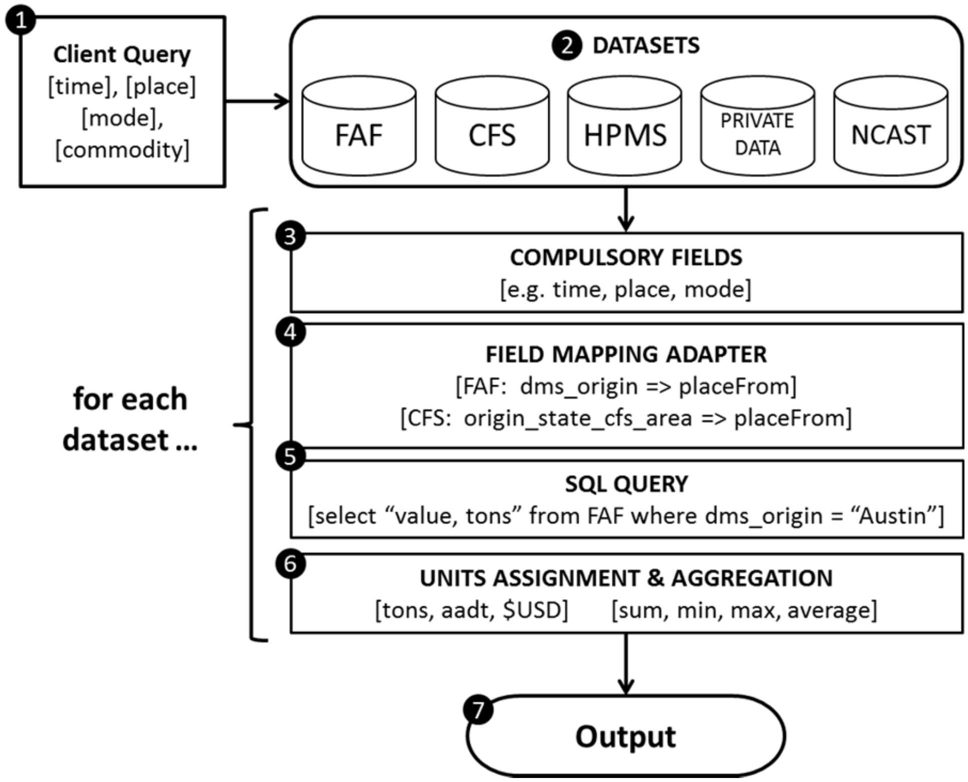


FIGURE 4: Proposed Mediator Architecture

Client Queries

The study team recommends that client queries should have the following minimum options available to users as tested in the prototype system (see Figure 5):

1. **Time –** For the prototype system, this was a date field that provided users with the option of querying data *from* a specific time frame or *limited to* a specific time period, e.g., "show all records from January 2004" or "show all records from January 2004 to August 2012".

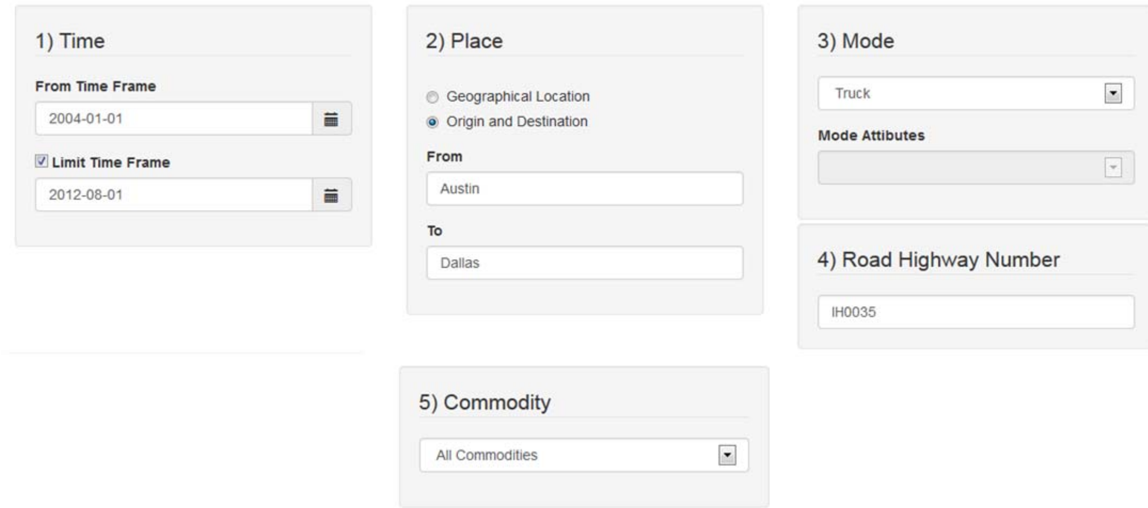


FIGURE 5: Example screenshot of minimum available query options

2. **Place –** The place option is broken down into two main options:
    a. Geographical Location – when a user chooses to search for data only for a specific geographical location, e.g., "show all inbound and outbound records relating to the city of Austin".
    b. Origin and Destination – when a user chooses to search for data relating to freight movement from an origin to a destination, e.g., "show all records for which origin equals Austin and destination equals Dallas".

3. **Mode –** Users should have the capability to filter records based on a specific mode of transport, e.g., air, marine, truck, pipeline, or rail. An additional option is the ability to select features unique to a specific mode, e.g., highway name for trucks, number of rail cars for rail, number of vessels for marine, etc.

4. **Commodity –** Users should have the option of filtering between multiple commodities.

Datasets

In order to keep the integrity of the datasets and facilitate future updates of existing datasets, the study team proposes that the data dictionary structure of datasets be kept the same as the original. Field mapping adapters are then linked to these datasets when users submit queries.

For datasets referencing similar lookup tables, a single lookup table is recommended. For example, both the FAF and the CFS use the Standard Classification of Transported Goods (SCTG) classification system. Therefore, linking both datasets to a single SCTG table is much more efficient than having duplicate tables (see Figure 6).

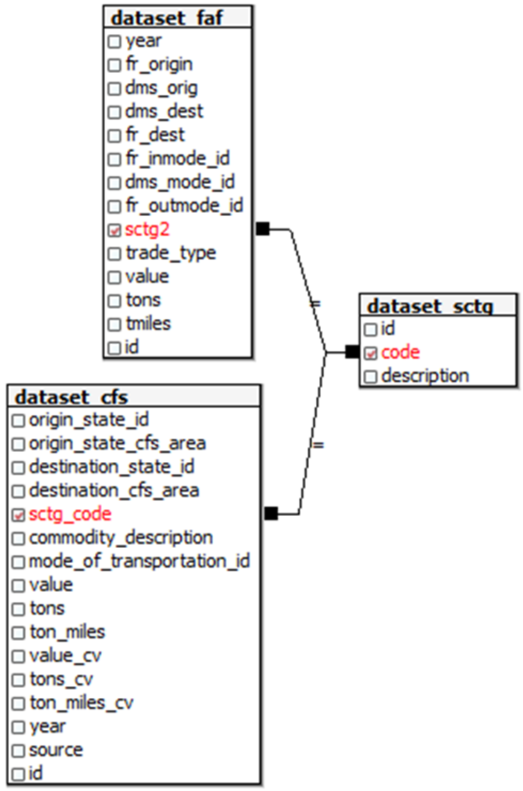


FIGURE 6: Example showing FAF and CFS datasets referencing a common SCTG lookup table

Compulsory Fields

The compulsory fields' layer provides an additional layer for speeding up query processing time. The purpose of this layer is to prevent the system from querying data tables for which the minimum required search options are not specified by the user. For example, for most datasets time and place are required. Therefore, if the user fails to specify a time option during the querying phase, then the system will alert the user that "time" is a required input to query a certain dataset. Another example, as illustrated in Figure 7 where the user specifies only a time frame ("timeFrom" and "timeTo"), the system notifies the user that in order to query the FAF and CFS datasets, the "placeFrom" variable is required, and in order to query the HPMS and NCAST datasets, the roadway highway number (or "link") is required.

Input:

Query Search for: **timeFrom**:2004-01-01 **timeTo**:2012-08-01

Output

FAF : **Needed at least timeFrom, placeFrom**

CFS : **Needed at least timeFrom, placeFrom**

HPMS : **Needed at least link**

NCAST : **Needed at least timeFrom, link**

FIGURE 7: Compulsory Fields Query Input and Output Illustration

The compulsory field rules can be stored in a data table was illustrated in Figure 8, where *dataset_id* is a reference to the datasets being used in the system.

id [PK] serial	dataset_name character varying(256)	is_active boolean
1	dataset_faf	TRUE
2	dataset_ncast	TRUE
3	dataset_cfs	TRUE
4	dataset_hpms	TRUE
5	dataset_pvtdata1	TRUE
6	dataset_pvtdata2	TRUE
7	dataset_pvtdata3	TRUE

id [PK] serial	dataset_id integer	compulsory_field character varying(256)	is_active boolean
1	1	timeFrom	TRUE
2	3	timeFrom	TRUE
3	2	link	TRUE
4	2	timeFrom	TRUE
5	4	placeFrom	FALSE
7	1	placeFrom	TRUE
8	3	placeFrom	TRUE
10	4	link	TRUE

FIGURE 8: Compulsory Field Table

Field Mapping Adapter

The field mapping adapter relates the user-specified query fields to the original fields for each dataset. The structure of the field mapping adapter is shown in Figure 9 and as follows:

i) *dataset_id* (Foreign key): A reference to the existing dataset as shown in Figure 8

ii) *original_field* (Text): The original field name as specified in the dataset

iii) *mapping_field* (Text): The user specified field variable, e.g., timeFrom, timeTo, commodity, mode, placeFrom, placeTo

iv) *is_active* (Boolean): Determines if the field is still available in the dataset

v) *conversion_factor* (Numeric): Utilized if fields are known to report similar values but in different units, e.g., dollars and million-dollars.

vi) *human_readable_name* (Text): This field displays the human-readable text of a field name, e.g., “dms_orig” in FAF will have a human-readable name of “Domestic Origin FAF Area”.

vii) *unit_name* (Text): Field unit of measurement to be used in reporting and conversions

viii) *is_summable*, *is_averageable*, *is_minmax* (Boolean fields): These fields are used in the units assignment and aggregation stage

ix) *field_type* (Text): This variable is used in specifying whether a field is an input field or an output field, i.e., if an input field, then the field can be used in querying the dataset and if any output field then the field is only useful for showing querying results.

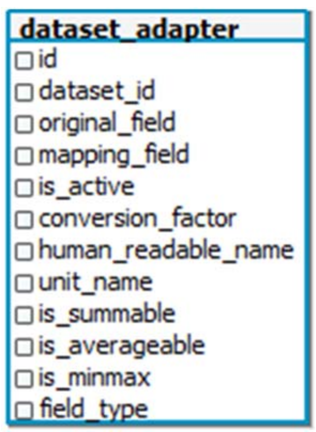


FIGURE 9: Field Mapping Adapter

Figure 10 shows sample records of the field mapping adapter table.

id [PK] s	datas intege	original_field character varying(256)	mapping_field character varying	is_active boolean	conversion double pre	human_readable_ character varying(	unit_name character vary	is_summable boolean	is_average boolean	is_minmax boolean	field_type character va
10	3	mode_of_transportation__p	mode	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
11	3	sctg_code	commodity	TRUE	1	Commodity	''	FALSE	FALSE	FALSE	IN
12	3	commodity_description	commodity	TRUE	1	Commodity	''	FALSE	FALSE	FALSE	IN
13	3	tons	commodity	TRUE	1	''	tons	TRUE	FALSE	FALSE	OUT
14	3	value	commodity	TRUE	1	''	milliondolla	TRUE	FALSE	FALSE	OUT
15	3	ton_miles	commodity	TRUE	1	''	''	FALSE	TRUE	TRUE	OUT
16	3	value_cv	commodity	TRUE	1	''	''	FALSE	FALSE	FALSE	''
17	3	tons_cv	commodity	TRUE	1	''	''	FALSE	FALSE	FALSE	''
18	3	ton_miles_cv	commodity	TRUE	1	''	''	FALSE	FALSE	FALSE	''
19	1	fr_origin__area_name	place	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
20	1	fr_origin__area_name	placeFrom	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
21	1	dms_orig__area_name	place	TRUE	1	Domestic Origin	''	FALSE	FALSE	FALSE	IN
22	1	dms_orig__area_name	placeFrom	TRUE	1	Domestic Origin	''	FALSE	FALSE	FALSE	IN
23	1	dms_dest__area_name	placeTo	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
24	1	fr_dest__area_name	placeTo	FALSE	1	''	''	FALSE	FALSE	FALSE	IN
25	1	fr_inmode__mode__parent	mode	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
26	1	dms_mode__mode__parent	mode	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
27	1	fr_outmode__mode__parent	mode	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
28	1	year__gte	timeFrom	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
29	1	year__lte	timeTo	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
30	1	sctg2	commodity	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
31	1	tons	commodity	TRUE	1	''	tons	TRUE	FALSE	FALSE	OUT
32	1	value	commodity	TRUE	1	''	milliondolla	TRUE	FALSE	FALSE	OUT
33	1	tmiles	commodity	TRUE	1	''	''	FALSE	TRUE	TRUE	OUT
34	1	trade_type	commodity	TRUE	1	''	''	FALSE	FALSE	FALSE	INOUT
35	2	object_id	unclassifiables	TRUE	1	''	''	FALSE	FALSE	FALSE	''
36	2	cal_date__gte	timeFrom	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
37	2	cal_date__lte	timeTo	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
38	2	timezone	time	TRUE	1	''	''	FALSE	FALSE	FALSE	''
39	2	state__name	place	TRUE	1	''	''	FALSE	FALSE	FALSE	IN
40	2	hwy_link	link	TRUE	1	''	''	FALSE	FALSE	FALSE	INOUT
41	2	hwy_num	link	TRUE	1	''	''	FALSE	FALSE	FALSE	''
42	2	mile_seg	link	TRUE	1	''	''	FALSE	FALSE	FALSE	''
43	2	hwy_dir	link	TRUE	1	''	mph	FALSE	FALSE	FALSE	OUT
44	2	seg_id	link	TRUE	1	''	''	FALSE	FALSE	FALSE	''
45	2	length	link	TRUE	1	''	''	FALSE	FALSE	FALSE	''
46	2	am_peak	link	TRUE	1	''	mph	FALSE	FALSE	TRUE	OUT
47	2	midday_peak	link	TRUE	1	''	mph	FALSE	FALSE	TRUE	OUT

FIGURE 10: Field Mapping Adapter Table Sample

SQL Query

The SQL query layer utilizes the mapping scheme developed in the field mapping adapter above. Dataset fields that map to the user-specified variables (see *mapping_field* column) are used in querying each dataset and fields with a field type equal to "OUT" form the query results.

Under the current setup, if two dataset fields map to the same *mapping_field* and both are of the field type "IN", both of the fields will be searched for as both fields can contain different values. For example, should the user search for "Texas" using the "placeFrom" variable, only the "origin_state__name" field under CFS will return results even though both the "origin_state__name" and "origin_state_cfs_area" both map to the "placeFrom" field.

Units Assignment and Aggregation

The units assignment and aggregation layer is used in performing simple arithmetic operations to be used in providing summary data. Arithmetic operations include summation and minimum and maximum value determinations.

Output Screen

As discussed earlier, final output from the database can be in the form of

i) summary data as discussed in the units assignment and aggregation section,

ii) ability to download the complete queried data, and

iii) ability to display data in the form of graphs, tables, or online maps for visualization purposes.

COST ESTIMATION

Following the processes earlier outlined in the conceptual architecture for an integrated freight database system, the study team proposed a list of items to be considered in estimating the cost for developing and maintaining the system. The task breakdowns and associated unit cost measurements are shown in Table 1. These estimates are based on the study team's experiences during the performance of the study and in the development phase of the prototype database.

Table 1: Integrated Freight Database System Architecture Unit Cost Estimates

Tasks	Unit Costs
1. Project Initialization and Setup	
1.1. Finalize agreements on design specifications and needs with TxDOT	200 man-hours
2. Integrate and use publicly available data	
2.1. Inventory data dictionary elements from publicly available databases	300 man-hours
2.2. Design and test theoretical field mapping adapter that connects to each data dictionary element	200 man-hours
3. Engage private sector stakeholders	
3.1. Correspondence with existing and new freight data partners	40 hours/week
4. Electronic submission of data by freight data sharing partners	
4.1. Design initial system administrator interface	300 man-hours
4.2. Develop and test electronic form submission including secure web connections	200 man-hours
5. Data quality and validation	
5.1. Setup automated data quality checks	400 man-hours
5.2. Conduct manual data quality checks	40 hours/week
6. Automated data scrubbing and aggregation	
6.1. Develop universal data scrubbing adapter	400 man-hours
6.2. Develop universal data aggregation adapter	800 man-hours
6.3. Develop partner specific data scrubber, data aggregator, and field mapping adapters	40 man-hours per data partner
7. Secure data storage and restricted access	
7.1. Acquire secure data warehouse and web servers	Lump sum
7.2. Develop restricted user management system	200 man-hours
7.3. Develop interface to store publicly available data	400 man-hours
7.4. Develop interface to store private sector data	400 man-hours
7.5. Integrate existing TxDOT data centers	300 man-hours per data center
8. Data output and analysis tools	
8.1. Develop and test client-side user interfaces including GIS map integration and graphs	700 man-hours
9. Training	
9.1. Develop training manuals for administrators and general users	300 man-hours
9.2. Conduct training sessions for administrators and general users	40 man-hours per session
10. System Maintenance	
10.1. Dedicate administrators to maintain and update system	80 hours/week

RECOMMENDATIONS

Similar to guidelines outlined in *Freight Data Sharing Guidebook*, the study team recommends a rigorous outreach and follow-up effort to sustain the success of any freight data sharing partnership. The study team recommends that TxDOT use a trusted third party such as an academic institution or a consultant to spearhead the implementation and development of the integrated freight data system. Validation of this sort of partnership was evidenced in the study where stakeholders were more inclined to communicate further with the study team based on their affiliation and trust of the institution. A similar observation was made in the American Transportation Research Institute (ATRI) Freight Performance Measures project involving the Department of Transportation and the Federal Highway Administration to develop a strategic freight performance tool for identifying and assessing truck mobility issues on our nation's highways[22]. As described in *Freight Data Sharing Guidebook*, ATRI is considered a trusted third party by the trucking institution and its relationship with the trucking industry allowed sensitive data to be collected and cleansed before being used in the project. Other observations include Washington State's *GPS Truck Data Performance Measures Program*[23], the *Alabama Statewide Freight Study and Action Plan* involving the University of Alabama at Huntsville and the consulting firm of J. R. Wilburn and Associates[24], and the Texas Transportation Institute's work on using RFID readers to measure wait times at U.S.-Mexico border port of entries[25].

In addition to TxDOT partnering with a trusted third party, the agency should be involved in promoting and advocating the initiative through its relationships with trade associations and industry experts. For example, through TxDOT's Freight Advisory Committee (FAC), private-sector partner organizations can be invited to participate in the data sharing partnership. The FAC provides a convenient avenue for both the public and private sector agencies to define and articulate the purpose, benefits, challenges, and concerns of data sharing. Issues relating to privacy concerns and how the data can be better utilized in improving upon TxDOT's planning efforts can be adequately addressed.

Also recommended by *Freight Data Sharing Guidebook* is the frequent communication to public and private-sector stakeholders of the benefits garnered as a direct result of the data partnership efforts. TxDOT should follow up on planning programs or efforts being undertaken to address transportation infrastructure issues identified as a result of the data being provided. Partners should also be assured that the data being collected is used for the purposes for which it is intended and will not be used in any way that harms the operations of the partners, such as through regulatory processes. A clear non-disclosure contract is required to reinforce a sense of trust with freight partners and the sample provided in the *Freight Data Sharing Guidebook* is a good option.

Finally, adequate funding should be allocated to sustain the program and cover cost of operations. Any data sharing partnership will require a long-term commitment from TxDOT and the partners and this can be demonstrated through allocation of adequate funding for the program. Adequate funding also ensures the collection of updated data thus making the program more relevant in addressing freight related issues being faced by TxDOT. A lack of commitment may result in partners not renewing agreements with the agency as they may sometimes not experience any direct benefits from participating in the program.

---

[1] Transportation Research Board (2003). "A Concept for a National Freight Data Program". Committee on Freight Transportation Data: A Framework for Development. Special Report 276.

[2] Mani, Akshay, and Prozzi, Jolanda (2004). "State-Of-The-Practice In Freight Data: A Review of Available Freight Data in the U.S.". Report No. 0-4713-P2. Center for Transportation Research. The University of Texas at Austin

[3] Cambridge Systematics, Inc. and GeoStats, LLP. (2010). "Freight-demand Modeling to Support Public-sector Decision Making". Transportation Research Board. Project NCFRP 8.

[4] Tok, Andre, Miyuan Zhao, Joseph YJ Chow, Stephen Ritchie, and Dmitri Arkhipov. 2011. Online Data Repository for Statewide Freight Planning and Analysis. *Transportation Research Record: Journal of the Transportation Research Board* 2246 (1), 121-129.

[5] Prozzi, Jolanda, Seedah, Dan, Carrion, Migdalia, Perrine, Ken, Hutson, Nathan, Bhat, Chandra, Walton, Michael. (2011). "Freight Planning for Texas – Expanding the Dialogue". Report No. 0-6297-1. Center for Transportation Research. The University of Texas at Austin.

[6] Quiroga, Cesar, Koncz, Nicholas, Kraus, Edgar, Villa, Juan. Warner, Jeffrey, Li, Yingfeng, Winterich, David, Trego, Todd, Short, Jeffrey, and Ogard, Elizabeth. 2011. Guidance for Developing a Freight Transportation Data Architecture. Transportation Research Board. Report NCFRP 9

[7] Rhodes, Suzann S., Berndt, Mark, Bingham, Paul, Bryan, Joe, Cherrett, Thomas J., Plumeau, Peter, and Weisbrod, Roberta. 2012. Guidebook for Understanding Urban Goods Movement. Transportation Research Board. No. Project NCFRP 15A.

[8] Chase, Keith M., Patrick Anater, and Thomas Phelan. 2013. Freight Demand Modeling and Data Improvement Strategic Plan. SHRP 2 Report S2-C20-RW-2.

[9] Cambridge Systematics 1997. A Guidebook for Forecasting Freight Transportation Demand. Transportation Research Board. No. Project NCFRP-388.

[10] Prozzi, Jolanda, Wong, Carleton, Harrison, Rob. 2004. Texas Truck Data Collection Guidebook. Center for Transportation Research. The University of Texas at Austin.

[11] Cambridge Systematics et al. (2013), Freight Data Sharing Guidebook. National Cooperative Research Program of the Transportation Research Board. June. Report 25.

[12] Oracle. "Maintaining Data Integrity in Database Applications." Oracle Database Advanced Application Developer's Guide. http://docs.oracle.com/cd/E11882_01/appdev.112/e25518/adfns_constraints.htm (accessed December 4, 2013).

[13] IBM. "Types of constraints." IBM DB2 9.5 Information Center for Linux, UNIX, and Windows. http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=/com.ibm.db2.luw.admin.dbobj.doc/doc/c0020149.html accessed December 4, 2013).

[14] Wikipedia contributors, "Data cleansing," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Data_cleansing&oldid=584731700 (accessed December 4, 2013).

[15] U.S. Census Bureau, County Business Patterns - How the Data are Collected (Coverage and Methodology) http://www.census.gov/econ/cbp/methodology.htm

[16] Evans, Timothy, Laura Zayatz, and John Slanta. "Using Noise for Disclosure Limitation of Establishment Tabular Data." Proceedings of Bureau of the Census 1996 Annual Research Conference. US Bureau of the Census, Washington, DC. 1996.

[17] TxDOT (2013), Statewide Planning Map, http://www.txdot.gov/apps/statewide_mapping/StatewidePlanningMap.html

[18] TxDOT (2013), DriveTexas, http://drivetexas.org

[19] Texas Department of Motor Vehicles (2013), TxPROS Information & Tutorials, http://txdmv.gov/motor-carriers/oversize-overweight-permits/txpros-info-and-tutorials

[20] TxDOT (2013), Texas Statewide Analysis Model – Third Version (SAM-V3) – Overview, Published by the Texas Department of Transportation

[21] An ongoing project by the study team to identify differences in data element definitions and methods for bridging these differences is currently underway. Additional information on this project is available at http://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=3537

[22] American Transportation Research Institute, National Corridors Analysis and Speed Tool (N-CAST) Available at http://atri-online.org/n-cast/

[23] McCormack, Edward, Wenjuan Zhao, and Dale Tabat. "GPS Truck Data Performance Measures Program in Washington State." (2011).

[24]Anderson, Michael, and Gregory Harris. "Development of Statewide Freight Plan for Alabama Using Integrated Freight Planning Framework." Transportation Research E-Circular E-C158 (2011).

[25] Texas Transportation Institute, Using RFID Readers to Measure Wait Times at the U.S.-Mexico Border. Available at http://tti.tamu.edu/2013/03/01/using-rfid-readers-to-measure-wait-times-at-the-u-s-mexico-border/