1 Capability

The **Safety Data Quality** capability is:

> Establish data quality measures (timeliness, accuracy, and integrity), especially for those data elements used in determining ratings or making decisions. Regularly check data used in Commercial Vehicle Information Systems and Networks (CVISN) processes for quality; purge stale data; repair errors.

2 Working Group Recommendations

The Enhanced Safety Information Sharing Working Group offers these summary recommendations related to this capability:

- Data quality concerns cross all the capability areas and involve a wide array of source systems operated by many agencies and organizations. Data quality improvement should be part of every solution considered in Expanded CVISN.
- The working group recommends one option related to this capability: Standardize information sharing.
- One activity related to standardizing information sharing is proposed for near-term funding. Suggested steps include:
  - Select a high-priority business process where data quality issues have a high impact. The working group suggests these processes as possible candidates: e-screening enrollment, e-screening bypass, selecting carrier/vehicle/driver for inspection, selecting carrier for compliance review.
  - Establish business and technical teams to tackle: review of lessons learned; alignment of business processes; universal data dictionary; common identifiers; definition of standards for information sharing that include structure and protocols; improve constraint checking.
  - Define requirements for new or modified systems and business processes.
  - Analyze technical alternatives for implementing the changes.
  - Choose alternatives and assign responsibility and resources to implement.
  - Implement, test and deploy the chosen alternatives.
  - Measure the resulting improvement in data quality.
  - If quality improvement is demonstrated, return to the first step and iterate. If no improvement is demonstrated, rethink the alternatives.
- Members of the Enhanced Safety Information Sharing Working Group should be invited to participate on the teams. The working group should at least serve in an advisory capacity to the teams' efforts.
- Efforts to standardize information sharing should be coordinated with related initiatives both within and outside the transportation realm.

3 Concept of Operations

The term concept of operations (ConOps) means operational attributes of the system from the operators' and users' views. The ConOps allows for the use of a variety of technologies. There may be potential benefits to be gained by using some sophisticated technologies, but only if the technologies are part of a well-conceived and vetted set of practices, are thoroughly understood and tested, and are implemented and used correctly. This section summarizes the proposed concept of operations.

Existing systems contain much of the information needed to achieve the goals of the Expanded CVISN initiative. To increase information sharing, expand, merge, establish interfaces between, or enhance existing **information management systems** [e.g., Motor Carrier Management Information System (MCMIS), Commercial Driver's License Information System (CDLIS), Safety and Fitness Electronic Records (SAFER), Commercial Vehicle Information Exchange Window (CVIEW), Performance and Registration Information Systems Management (PRISM), International Registration Plan (IRP) and International Fuel Tax Agreement (IFTA) clearinghouses] to include:

- Role-based access to services using single sign-on
- Open standards for information sharing
- Improved and flexible user interfaces (e.g., provide default look and feel based on user's role; allow user to tailor)
- Standardization around a small number of standards. This gives each state the flexibility to work within its overall statewide architecture, but still encourages commonality among states' systems and approaches.
- Collection of data once and frequent reuse (e.g., collect census data from a carrier and re-use that data from a single source whenever it's needed)
- Consistent level of service regardless of time-of-day or day-of-year
- Improved access to data about all commercial drivers
- More timely and complete IRP and IFTA data in snapshots
- Consistent identification of carrier, driver, vehicle, and cargo
- Association of entities that are related during a trip (e.g., John Driver working for Carrier XYZ driving vehicle with plate 1234567 registered in Maryland hauling trailer with plate 8901234 registered in Delaware)
- Electronic security device event data (to track the status of and activities related to a security device attached to the container and/or trailer)
- Integrate with or link to asset tracking, arrival scheduling, and other vehicle, port and freight information systems [e.g., Freight Information Real-Time Systems for Transport (FIRST), electronic freight manifest, State On-Line Enforcement System (STOLEN)].
- Access to up-to-date credentialing information [e.g., oversize/overweight (OS/OW) permits].

To improve the quality of information and to improve access, develop, expand, merge, or enhance **data collection and reporting systems** used in the field [e.g., ASPEN, Carrier Automated Performance Review Information (CAPRI)] to include:

- Open standards for data collection and reporting
- Access to driver snapshots
- Out-of-service (OOS) processing
- Uniform citation reporting
- Uniform crash reporting
- Hours of service compliance evaluation
- Vehicle and cargo security checks
- Heavy duty diesel (HDD) emissions inspections
- Interface with electronic on-board systems
- Wireless technology.

Look for successes within innovative programs and build on or adapt their business models for broader use. Categories of programs/systems to review include:

- Electronic toll collection systems (e.g., E-ZPass)
- Electronic credentialing systems for multiple credentials [e.g., One-Stop Credentialing and Registration (OSCAR)]
- Regional data-sharing systems [e.g., Extensible CVIEW (xCVIEW)]
- Roadside information reporting systems (e.g., ASPEN)
- Port scheduling/access programs (e.g., PierPass)
- Freight security improvement programs [e.g., Operation Safe Commerce (OSC)]
- Cross-program technical interchange (e.g., CVISN/PRISM)
- Border-crossing improvement programs [e.g., Free and Secure Trade (FAST)]
- Data challenge and correction (e.g., DataQs).

Review and build on technology lessons learned. Categories of programs/initiatives to review include:

- Recent operational tests [e.g., Federal Motor Carrier Safety Administration’s (FMCSA’s) Hazardous Materials (HazMat) Op Test]
- Intelligent Transportation Systems (ITS) initiatives [e.g., Vehicle Infrastructure Integration (VII)]
- Applications and uses of standards [e.g., Dedicated Short Range Communication (DSRC) standards]

- Technology transfer opportunities [e.g., Federal Rail Administration's (FRA's) railroad track status reporting]
- Commercial Vehicle Operations (CVO) infrastructure deployments (e.g., e-screening)
- E-credentialing deployments (e.g., Core CVISN Web credentialing)
- Broader transportation infrastructure deployments (e.g., e-toll collection)
- Data sharing models (e.g., CDLIS).

4 Requirements

Discussions with the members of the Enhanced Safety Information Sharing Working Group established by FMCSA via the ITS/CVO 2005 Deployment Showcase seeded the requirements stated in this section. Subsequent review by members of the working group finalized the requirements.

To clarify what we mean by "safety data," this definition is suggested: Safety data includes all information used to:

- Identify a carrier, vehicle, driver, shipper, or cargo.
- Evaluate compliance with all commercial motor vehicle regulations.
- Compute safety assessment.

The focus should be on making sure the originating and authoritative sources have high-quality data.

It is difficult to match data from different sources when the sources identify the same entity differently. So, one set of requirements for improving safety data quality focuses on common identifiers:

- Adopt a common identifier for the motor carrier. Ideally, the approach should be relevant both to interstate and intrastate operations. US Department of Transportation (USDOT) number is preferred.
- Whenever information about a motor carrier is recorded, include the common identifier.
- Adopt a common identifier for the vehicle.
- Whenever information about a vehicle is recorded, include the common identifier.
- Adopt a common identifier for the driver.
- Whenever information about a driver is recorded, include the common identifier.

- See the table below, extracted from the CVISN Architecture's *Recommendations for Primary Identifiers* white paper, Baseline Version 1, October 2002 (paper can be downloaded from http://cvisn.fmcsa.dot.gov/downdocs/cvisndocs/whitepapers/identifiers_v1.doc).
  - Note that there are two primary identifiers for vehicles: the Vehicle Identification Number (VIN) assigned by the manufacturer, and the Vehicle Plate ID assigned by the jurisdiction in which the vehicle is registered. Both identifiers are necessary to accommodate existing systems based on each. The VIN is unique and assigned to the vehicle permanently. A vehicle plate ID may be assigned to more than one vehicle, and a single vehicle may have different vehicle plate IDs over the course of its lifetime.
  - Additions from working group members are shown in red text

Entity	Identifier Name	Identifier Segments	Number of Characters
Motor Carrier	Primary Carrier ID		
	For ***interstate*** carrier:	Carrier-Specific Identifier (alphanumeric); must be USDOT number +	12 (max)
	e.g.,	Carrier Terminal ID designated by carrier (alphanumeric) (optional) +	4 (max)
	12345 A001 (note that '12345' must be the carrier's USDOT # ; the terminal ID 'A001' is optional)		
		CVO Company Type (optional)	TBD
	For ***intrastate*** carrier:	Country Code (alphanumeric); the allowable codes will be defined in the FHWA Code Directory +	2
	e.g.,	Jurisdiction (state or province) Code (alphanumeric); the allowable codes will be defined in the FHWA Code Directory +	2
	US CA 123A45689 1234 (note that the terminal ID '1234' is optional)		
		Carrier-Specific Identifier; if carrier is intrastate and has a USDOT number, must be USDOT number; for state-specific IDs, the Carrier-Specific Identifier may include a prefix to clarify the agency/source of the identifier) +	12 (max)
		Carrier Terminal ID designated by carrier (alphanumeric) (optional)	4 (max)
		CVO Company Type	TBD

Entity	Identifier Name	Identifier Segments	Number of Characters
	For all carriers: Federal Taxpayer Identification Number e.g., E 123456789 Note: Open issue regarding Mexican and Canadian carriers. See section 8.1.	Type (alphanumeric); S for Social Security Number, E for Employer Identification number + Tax ID Number (alphanumeric)	1 9
Vehicle	Vehicle Identification Number e.g., 1FDKE30F8SHB33184 and Vehicle Plate ID e.g., US CA 12345664820M	VIN assigned by manufacturer (alphanumeric) Country code (alphanumeric); the allowable codes will be defined in the FHWA Code Directory + Jurisdiction (state or province) code (alphanumeric); the allowable codes will be defined in the FHWA Code Directory + License plate ID (alphanumeric) + Vehicle class code (alphanumeric) (optional); allowable codes set by each jurisdiction + Vehicle sub-class code (alphanumeric) (optional); allowable codes set by each jurisdiction	30 (max) 2 2 12 (max) TBD (max) TBD (max)
Transponder	Transponder ID e.g., 0 123456789 or 1 9999 232323	segments shown below Transponder ID Definition Flag (0=current; 1=IEEE 1455-1999) + *If Transponder ID Definition Flag = current*, then the other segment is: Transponder Serial Number assigned by manufacturer *If Transponder ID Definition Flag = IEEE 1455-1999*, then the other segments are: Manufacturer Identifier +	10 (max) 1 (1 bit) 8 (32-bit hexadecimal value) 4 (16 bits hexadecimal value)

Entity	Identifier Name	Identifier Segments	Number of Characters
		Transponder Serial Number assigned by manufacturer	5 (20 bits hexadecimal value)
Driver	Driver Unique ID e.g., US MD B99999999999A	Country code (alphanumeric); the allowable country codes will be defined in the FHWA Code Directory + Jurisdiction (state or province) code (alphanumeric); the allowable subdivision codes will be defined in the FHWA Code Directory + Driver specific identifier (driver license number) assigned by jurisdiction (alphanumeric)	2 2 16 (max)
Shipment	Shipment Unique ID e.g., 123456789776655443322	Shipper ID. DUNS number suggested as a candidate (alphanumeric) + Bill of Lading number assigned by the shipper identified above (numeric)	9 (suggested) 12 (max)
Trip	Trip/Load Number e.g., 123456789761231	Carrier DUNS number as assigned by Dun and Bradstreet (numeric) + Trip unique number as assigned by carrier (numeric)	9 6
Container	Container Unique ID e.g., SUDU3070079	Suggested as a candidate: Container number marked on side (in accordance with ISO 6346) (alphanumeric)	11 (suggested)

Many activities collect information about commercial vehicle operations. It is not always straightforward to identify the appropriate carrier or shipper to associate with the vehicle, driver, trailer, and cargo. To improve the quality of the association of entities this capability should include these aspects:

- Clarify the rules for associating carriers and shippers with vehicles, drivers, trailers, and cargo in a variety of commercial vehicle operations including credentialing, roadside inspection, crash events, citations, and convictions. Account for the wide variety of possible business models [e.g., owner-operators under contract to a larger carrier, multiple drivers per trip, less-than-truckload (LTL), cargo that changes hands during shipment].
- Standardize the rules across jurisdictions.
- Make the rules more readily available to the personnel who participate in those activities.
- Increase and improve training for those who must use the rules to establish data associations in data collection activities.

Data definitions differ across business functions (e.g., registration, insurance) and across jurisdictions. To improve data quality, these differences must be addressed:

- Standardize definitions for all CVO data elements that are shared among the state, federal, and industry communities.
- Systems that serve as repositories for information should check incoming data for conformance with established standards and reject data that do not comply.

Existing standards for sharing safety information (e.g., snapshots, inspection reports) should be revised and extended to include the standard identifiers and universal data dictionary data elements. The standards should specify structure and protocols for sharing information. If possible, the standards should be published as open standards to facilitate adoption by stakeholders.

Data collection systems must provide timely updates of information to centralized information systems.

- Criteria for “timeliness” may differ based on the data being shared. For example, initial credentialing activities and OOS status should be reported quickly.
- Timeliness should be measured as end-to-end from the source to the user. All systems that participate in data sharing should be measured.
- In some cases, it may be appropriate to carry a “purge date/time” along with a data element.

Not all roadside systems have access to the data they require. In particular, this capability should address the requirement for roadside systems (e.g., PrePass) to obtain raw crash data.

Crash data are not always accurate or reported correctly. This capability should address these requirements:

- Clarify the rules for reporting crashes to FMCSA.
- Ensure that the interpretation of the rules is the same across jurisdictions.
- Make the rules more readily available to the personnel who participate in crash reporting activities.
- Increase and improve training for those who report crash data.
- In the system(s) used to report crash data, build in cross-checking to validate information provided by the driver.
- Send the crash report to the carrier associated with the crash. Provide a point of contact and a system for resolving errors in the report.

Additional outreach and assistance should be provided to industry for using the existing tools to access and review their own records. The online tools should make it harder to enter errors and easier to correct them. For instance, it is very easy to enter an erroneous name; error checks on names should be added.

5 Potential Solution Alternatives

In Draft 1 of this report, several potential solution options for different aspects of the **Safety Data Quality** capability were identified. The working group decided that it would make sense to combine several of the original options to tackle the data quality issue more effectively:

- Recommended Option 1: Standardize information sharing (Alignment of business processes + Universal data dictionary + Common identifiers + Structure + Protocols + Improved constraint checking)

This option was viewed as the highest priority solution for improving data quality.

The working group placed the remaining solution options at a lower priority:

- Enterprise approaches
- Training
- Data push… and push-back.

For the lower-priority options, the descriptions are included below, but no further analysis will be provided in subsequent sections.

5.1 Recommended Option 1: Standardize information sharing

Working group members stated that the most important data quality improvement that CVISN can address is the standardization of information sharing. This involves five key aspects:

- **Align business processes**. Promote a comprehensive review by business process owners to identify mismatches across agencies and systems. Review lessons learned. Assign business process conflicts to subcommittees for resolution by business process experts.
- **Adopt a universal data dictionary**. Develop/assemble a comprehensive data dictionary. Include syntax and format constraints and add full semantics that explain the intended business use of the data elements for each destination system and user type. Span at least all CVISN systems and legacy systems that are interfaced to SAFER via CVIEW or equivalent, ASPEN or equivalent, MCMIS, Licensing and Insurance (L&I), Analysis & Information (A&I), PRISM, SAFETYNET – any system that manages data that are shared with federal, state, or industry stakeholders or their systems. Consider the possible impacts on non-CVO systems as well.
- **Adopt universal common identifiers**. The CVISN Architecture recommends that the stakeholder community adopt standard primary identifiers for carrier, vehicle, transponder, driver, shipment, and international trip in all data exchanges. The working group suggested additions and updates to the identifiers (see Section 4, Requirements) based on experience with Core CVISN implementation.
- **Publish open standards to specify structure and protocols for safety information sharing.** Extend and update existing interface control documents and open standards to reflect the common identifiers, universal data dictionary elements, message structures, and information sharing protocols. Include any new elements required. If possible, publish the information as open standards.
- **Improve constraint checking**. Using the ASPEN model, change the emphasis of MCMIS, SAFER, and other data management systems' data acceptance policies from accommodation to constraint checking. Provide advance notice of new constraints. Provide consulting support for technical solutions and business process issues to states that would otherwise be inclined to ignore the constraints and stop sharing data.

This option would involve the following activities:

1. Identify a small set of specific data elements with quality problems that do not seem to involve significant institutional issues. Develop a measurement of the current state of data quality for the selected data elements.
2. Assemble an energetic team of experts on the business processes that create and consume the identified data elements. The team will analyze the data quality issues in light of actual business use to determine whether business process clarification or modification would be an appropriate solution to the quality problems, or whether technical change is warranted.
3. Assign a technical team to consult with the business process team and capture their conclusions as system requirements where appropriate. Assess whether rules for

transforming data are required. Concurrently, the technical team will create or update existing data dictionary documentation that specifies, for each identified data element, the data format, the acceptable range of values, the expected update frequency, the authoritative source and the intended use of the data element. Also document any issues with uniqueness or availability of common identifiers for the intended business use. The teams will also recommend structure and protocol standards for information sharing.

4. Analyze technical options for implementing the system requirements, including a "nothing-is-off-the-table" set of options for what system(s) and/or process(es) could be modified (or added) to resolve the quality problems.
5. Choose a solution option and assign responsibility and resources for implementing it.
6. Implement, test and deploy the chosen solution.
7. Measure the resulting improvement in data quality.
8. If quality improvement is demonstrated, return to step 1 and iterate. If no improvement is demonstrated, rethink the alternatives.

The difficult part about solving data quality problems is managing the effort as a whole. While "the data quality problem" is broadly based on institutional issues at its core, a culture of quality improvement can be fostered, and a methodology can be implemented that overcomes the tendency to cynicism and despair that such a grand problem evokes. Key to step 1 is to tackle a set of problems that are real but manageable, avoiding unrealistic commitments to solve anything universally.

For this solution option, the architecture and possible impacts on federal, state, and industry systems/business processes will emerge from the activities summarized above. Significant impacts are to be expected for some processes and systems. When the changes have been made, federal, state, and industry systems should be able to share safety data more reliably. Analysis must carefully balance the costs and value of data correctness and consistency versus data completeness. The assessment of costs and value must be made case-by-case, for specific business objectives, based on practicality rather than philosophical principles. It isn't feasible to provide a quantitative estimate of costs and benefits in advance of each iteration's step 4, but it should not be hard to estimate a level of effort to reach step 4, and then provide a proposed scope, benefits and cost statement as a result of step 5. Even within step 5, an incremental approach is appropriate, where there are decision points to insure that increasingly complex iterations of the specifications (from functional and technical requirements through conceptual design through a detailed implementation plan) are in line with the resource expectations.

To select appropriate data elements for data quality improvement, objective and repeatable measurements of chosen aspects of data quality are needed. Data quality measurements may be broken down into measurements of the quality of the data model, the data values, the data domains, and the data presentation. In addition, the quality of information policy may be measured. Both static and dynamic measurements are needed to assess the quality of existing data and to trace quality through the working systems. The measurements must be tailored to the perceived data quality problems of greatest significance.

Data domains are a way to characterize the structure of data elements, and then to relate sets of data elements to one another through their structural characteristics. The relationships are called data mappings. Consider social security numbers (SSNs) as an example. At the simplest level, they could be characterized as strings of 11 characters, but there's much more structure that can be used to manage the quality of social security numbers in a database system. In particular, the structural description could be enhanced to say that social security numbers use only digits 0-9 and hyphens, and further by saying that they occur as three digits followed by a hyphen, then two digits, and another hyphen followed by four digits. Finally, the structural description can use the Social Security Administration's reference data to restrict the three numeric fields according to rules spelling out what ranges of numbers have actually been issued as of a given date. The increasing levels of detail can be considered as a mapping from one domain to another, more precise domain, in this case being subsets of the higher-level domain.

On the other hand, however, increased structure that is not carefully managed can lead to difficulties when confronted with business change due to regulatory or legislative actions, or when expanding the set of jurisdictions that participate. A hyper-structured SSN checking scheme, for example, is completely appropriate from a technical perspective. However, if there were strong opinions that the scheme could change greatly over time, perhaps a less-structured approach would be more practical from the standpoint of maintenance of the federal and state systems that use and/or provide this data.

In much the same way, CVISN could gain significant data quality improvement by exploiting the deeper structure of the VIN. Similarly, license plate numbers have important structure in all jurisdictions, but it might be hard to maintain the domain mappings effectively for use in the SAFER database. This aspect of the quality of data domains is identified as "stewardship" of the domains and mappings.

Well-defined data domains and mappings are a tremendous aid to data quality management, whether to design constraints on input data, to evaluate the quality of an existing set of data, or for more complex measures and procedures. The uniformity of their use is a determining factor in their value to the data quality initiative.

David Loshin suggests candidate measurements for data quality. Material below was based on Loshin's book Enterprise Knowledge Management: The Data Quality Approach (2001, Morgan Kaufman/Academic Press, ISBN 0-12-455840-2).

The quality measures for data models include:

- Clarity of Definition
  - Degree to which there is a defined naming convention
  - Degree to which the named objects conform to the convention
  - If convention is lacking... number of different names used for attributes holding the same kind of data
  - Assign a score to each table and attribute name qualifying how well it describes the data it holds

- Comprehensiveness and Flexibility
  - Number of requests over a period of time for adding new tables
  - ... for adding new attributes
  - Degree of use of auxiliary databases for CVISN-related data
  - Number of times attributes are overloaded
  - Amount of code/time required to implement new functionality
- Robustness
  - Number of changes within a specified period to the data model other than new tables or columns, e.g., size or type change for an attribute, relations between tables, keys modified, ...
- Essentialness
  - Count the number of data elements, tables and attributes that are never read
  - ... never written
  - Count the number of redundant copies of the same data in CVISN systems
- Semantic Consistency
  - Degree to which similarly named attributes make use of the same data domain
  - Degree to which data attributes using the same domain are similarly named
  - Score data attribute definitions in terms of consistency across CVISN data resources.

Additional categories for data model measurements include:

- Attribute granularity
- Precision of domains
- Homogeneity
- Naturalness
- Identifiability
- Obtainability
- Relevance
- Simplicity/complexity
- Structural consistency.

The quality measures for data values include:

- Accuracy
  - Compare sampled values with...
    - a database of record
    - a similar, corroborative set of values from another table
    - dynamically computed values
    - results of manual work-flow
    - values supplied by irritated customers

  - Measure as percentage of correct values
  - Measure to include a distance function: how far off are the values?
- Null Values
  - Count the number of null-valued attributes
  - Mine for information: why are they null?
    - unavailable from authoritative source
    - not applicable for this entity
    - no allowed value that properly represents attribute value
    - just plain missing
- Completeness
  - Mandatory attributes: count records where attributes that are mandatory in transactions are lacking in tables
  - Optional attributes: count records where optional attributes are present
  - Inapplicable attributes: count records where some attribute is actually irrelevant or meaningless for the corresponding entity (e.g., maiden name for a male)
- Consistency
  - (this is a potentially useful measure that will take some work to tailor it for CVISN data sets – need to identify records to be compared; see Semantic Consistency above)
- Currency/Timeliness
  - Percentage of records that are up-to-date with respect to an identified external source
  - Percentage of records that are updated within their expected interval
  - Time lag between identified authoritative source update and CVISN system update (SAFER, CVIEW, ASPEN, roadside/screening system).

Measurement of the quality of data domains and mappings considers the following:

- The degree of agreement across the enterprise on the usage of defined domains
- The degree to which responsibility has been assigned and executed for stewardship of the domains and mappings
- The degree of ubiquity of use of identified reference data associated with the domains and mappings.

Measures of data presentation generally require dialog with users. Work will be required to tailor measurements to the CVISN data user community, but the reference provides guidance in these Quality Measures for Data Presentation categories:

- Appropriateness
  - Motivating example: if there are many user requests that result in changes to the data model or to the data presentation layer, that would indicate a low level of appropriateness.
- Correct Interpretation

- Flexibility
- Format Precision
- Portability
- Representation Consistency
- Representation of Null Values.

5.2 Low-priority option: Enterprise approaches

For the data quality problems that start at the top, only an enterprise approach can help. Such solutions cannot be funded or controlled by a program the size of CVISN, but support can be offered for those enterprises able to launch their own initiatives of broad scope.

- Develop training and marketing materials promoting data quality to Chief Information Officers (CIOs)
  - Lessons learned in the FMCSA Creating Opportunities, Methods, and Processes to Secure Safety (COMPASS) initiative
  - Benefits to government of developing a Data Quality Plan
  - Lessons and benefits from DataQs
- Consider broadening the reach of DataQs into state systems in return for executive commitment to data quality improvement initiatives.

5.3 Low-priority option: Data push... and push-back

The stakeholder most highly motivated to improve data quality is the motor carrier affected. The FMCSA DataQs system provides a model for customer-initiated data correction. A difficulty with any such approach, however, is that most of the customers are unaware the data even exists until they are caught up in a problem, which might have been preventable. This calls for an initiative to improve carrier awareness of relevant data sources and to ease the burden of successful data challenge across the whole spectrum of CVISN systems.

- Develop a subscription service for the most important data elements
  - Include FMCSA-defined “Influential Information,” similar data used by state regulators, and other data elements to be identified by stakeholders
  - Deliver data to carriers, registrants, owners and third-parties designated by them
  - Support direct “jump” to data challenge system without need for secondary log-in and data entry
- Launch an associated marketing program
  - Educate carriers on the data sources, common reasons for errors, potential impact on their business.

5.4 Low-priority option: Training

Data entry is the source of pernicious errors because it becomes "authoritative" with the click of the "enter" key. Without a solution in hand for common identifiers, with varying terminology interpretations built into regulations across jurisdictions, and with data entry validation dependent on enterprise solutions, training is a near-term approach that can help.

- Capturing correct USDOT numbers
  - Reinforce/reinvent the training methods and materials
  - Keep trying to get to all the folks who put a USDOT number on a form or into a system, wherever and whatever for
  - Specific requirements and scope for this training initiative could be developed by the "USDOT Number Team" recommended above
- Crash report terminology
  - Investigate differences among agencies and jurisdictions in use of similar terminology for different things and different terminology for similar things
  - Develop a "decoder ring" and promulgate.

6 Cost-Benefit Analysis

The following table provides a high-level cost-benefit analysis for the proposed solution option. The cost figures are rough estimates provided by working group members.

- Low means less than $100K
- High means more than $1M
- Medium is everything in between.

Option	Pro	Con	Cost
1 (Standardize information sharing)	All: Improved data quality. Federal: Build on existing standards. Opportunity to integrate with and leverage COMPASS initiative. State: More likely that information shared will be used correctly. New CVISN implementers could start with standards. Industry: More likely that data assigned to carrier will be correct.	All: Significant effort required to standardize information sharing. Federal: --- State: May require unbudgeted changes to legacy systems and/or legacy system interfaces. Industry: ---	Federal: High, but COMPASS should be handling most of this, so cost to CVISN may be low. State: Medium to High, depending on state. Industry: Will vary depending on the extent to which the carrier's systems use the data.

7 Business Case

Data quality concerns cross all the capability areas and involve a wide array of source systems operated by many agencies and organizations. Many of the current data quality problems can be traced back to accommodations made for source systems' variability during early efforts at building initial participation in CVISN. Another problem is the variability of non-CVISN state systems that interact with CVISN systems. There is still a realistic anxiety that if the CVISN program imposes rigorous constraints on data sent to SAFER, some states will simply drop out of the data sharing process. This source of data quality problems requires a careful balancing act between rigor and accommodation.

Another major cause of current data quality problems is a lack of clear guidance to developers of legacy system interfaces on the meaning and intended usage of data elements exchanged via the SAFER system. This source of data quality problems should be relatively easy and inexpensive to address within the CVISN information technology community.

Proposed solutions to these two categories of data quality problems may reveal that conflicting business processes are at the heart of the matter for particular data elements, and these conflicts may not be easy or inexpensive to address. One commonly cited category of such business process conflicts is the use of different identifiers for different credentialing processes for essentially the same entity (the motor carrier under IRP and IFTA, for example). Data quality problems due to conflicting business processes can be resolved with time and good will, and identifying them is of value in its own right.

Also contributing to data quality problems is the common use of multiple-point, manual data entry without adequate front-end and back-end data validation. The lack of adequate data validation stems largely from an inability to cross-reference multiple source systems, as well as from outmoded data entry processes. The FMCSA COMPASS initiative provides a model for addressing these problems through an enterprise-wide approach, but it may be unrealistic to expect state agencies to mount similar efforts due to budgetary constraints, lack of executive commitment, and other institutional barriers.

On-going systemic difficulties, with implications beyond the reach of information technology, are the causes of the great majority of data challenges processed by FMCSA. The most frequently cited reason for challenge to data in crash and inspection reports is the assignment of incorrect USDOT numbers. Incorrect USDOT number assignments might be reduced through training, especially for inspection reports, but the breadth of the population that might submit crash reports limits the potential for improvement via training in that area. Changes to the data challenge process could help: carriers could be required to assist in identifying the correct USDOT number before their challenge will result in removal of an inspection or crash report from their own. Dramatic improvement in correct carrier/vehicle association probably cannot be achieved without a fundamental change in the way motor carriers are identified and vehicles are associated with them.

When restricted to crash report data challenges only, the most-cited reason is that the crash was not reportable. The frequency of these challenges is probably due to differing definitions of

terminology used in crash reporting across jurisdictions and enforcement agencies, with added complications from the crash environment. In the midst of responding to a crash, for example, a first-responder is not likely to regard as top priorities the capture of correct USDOT numbers and proper interpretations of "commercial motor vehicle," "towed," and "disabled." A related difficulty is that some agencies and system operators are reluctant to use crash data that does not support the exclusion of not-at-fault crashes.

Stakeholders are motivated to address quality issues. All Expanded CVISN working groups raised the issue. Efforts already underway at FMCSA, in other federal agencies, and throughout the CVISN community provide momentum to address data quality now.

8 Issues

8.1 Institutional Issues

The deployment approach for improving safety data quality must balance the desire for wide participation among states with the need for consistent and reliable data. In each state, a limited number of enforcement personnel are assigned to handle commercial motor vehicles. Part of the problem with safety data quality is assigning the correct carrier to each reportable crash. At least in some cases, the assignment of the responsible carrier for a crash is made by an officer who interacts with commercial vehicles only infrequently. Can state personnel accurately and reliably make that determination?

For data quality problems that are traceable to business process differences, the hurdles may be very challenging. For example, many business processes are driven by definitions contained in state legislation, and may be very slow or nearly impossible to change.

Other data quality problems are traceable to weaknesses in data entry methods, including lack of data validation capability at the time of data entry. This fundamental problem may only be addressable through enterprise-wide approaches that cannot be afforded or controlled by agencies responsible for motor carrier regulation.

One of the standard identifier recommendations developed by the working group is that both the taxpayer identification number and the USDOT number be used to identify a carrier. The motivation for this recommendation is to be able to tie the carrier's tax data (e.g., IFTA status) to the carrier's safety record. Not all Canadian or Mexican carriers will have a US taxpayer identification number. To operate in the United States, Canadian and Mexican carriers are required to get a USDOT number. As part of that process, the carrier is to supply their US taxpayer identification number, if they have one. The carrier will have a US taxpayer identification number if the carrier pays US taxes.

For old-style VINs, duplication across manufacturers may pose a (remote but potential) issue in assigning a unique identifier to a vehicle.

8.2 Technical Issues

Data are currently replicated in many systems. Correcting an error in one system must ripple through all other systems that hold the same data, and the correction must persist. Anyone who has ever tried to correct a personal credit report knows how time-consuming and difficult the process is. To make it more likely to achieve and maintain a high level of quality, data replication should be reduced. Reducing replication means that information that is legitimately needed by multiple systems (e.g., identity and census data) must be standardized and shared efficiently from the authoritative source. The VIN is 17 characters long and prone to data entry errors. If the VIN could be scanned and entered automatically, errors could be drastically reduced.

9 Deployment Strategy

In deploying the **Safety Data Quality** capability, several aspects should be considered:

Improve data quality and integrity:

- Establish a consistent set of data elements that are common across information systems and analysis applications.
- Expand the use of standard identifiers for entities visible at the roadside (carrier, vehicle, driver, cargo, chassis) to link related information.
- Make information collection, access, and use consistent across interstate, foreign, and intrastate operations.
- Capture data electronically as close to the source as possible; once information is available electronically, it should be re-used instead of re-entered manually.
- Expand standard procedures and tools for reviewing, detecting problems in, and correcting errors in publicly-held data.
- Expand the use of on-line tools that provide industry with the ability to challenge and correct their own census, inspection, crash, and citation information.
- Control access to sensitive information.

Work together and share lessons learned:

- Work with stakeholders to define and deploy common data elements and interoperable business processes for all areas of CVISN expansion.
- Establish standardized terminology and common requirements for data collection, access, quality checks, and making corrections.
- Coordinate standards-related activities with appropriate standards development organizations.

- Actively solicit lessons learned from "early adopters" of CVISN and Expanded CVISN concepts, and determine how to apply those lessons more broadly.
- Actively engage stakeholders in identifying priorities, proposing solutions, and participating in prototype projects.
- Proactively reach out to stakeholders who may be affected by changes to systems or processes that are under discussion.
- Learn from other ITS activities about solutions applicable to CVO.

Deploy targeted solutions incrementally:

- Select information-sharing options based on users' needs and available technology (e.g., proactive data-provider "data push" versus user-initiated "data query").
- Prototype proposed solutions and link to existing capabilities.
- Consider small-scale solutions that can be expanded or serve as models for national deployment.
- Build in metrics to assess real improvements.
- Provide access to on-line analysis tools.
- Provide an approach that allows states to improve the quality of data sent to aggregation sources while continuing to maintain interaction with other state systems that may insist upon "lower quality" or "nonstandard" data.

Use appropriate technology to improve operations:

- Equip commercial vehicles with standard DSRC and other technologies, enabling a multitude of safety, security and productivity applications.
- Deploy interoperable technologies to support CVISN and other related CVO activities.
- As products become available, consider 5.9 GHz DSRC as an enabling technology for roadside-to-vehicle, vehicle-to-roadside, and vehicle-to-vehicle data exchange.
- Equip cargo containers and trailers with standard electronic security devices (ESDs).
- Expand the use and capabilities of portable and remote sensors to monitor environmental, facility, road and vehicle conditions and provide data to interested stakeholders.
- Apply new and emerging wireless capabilities [e.g., Bluetooth, Wireless Fidelity (Wi-Fi), Global Systems for Mobile Communications (GSM)] and onboard technologies to improve on-road and roadside operations and reduce costs.

The working group recommends a series of activities related to the Safety Data Quality capability as defined in Section 5.1.

9.1 Standardize information sharing

It is likely that data quality problems start with business processes, and it is certain that priority for solving data quality problems should be assigned based on the business value of the potential quality improvement.

Analysis of technical options must be flexible, wide-ranging and creative. In a program that crosses as many institutional boundaries as CVISN, attitudes and expectations can prematurely rule out solutions that would be optimal if they were viewed from a whole-system perspective.

Although the words may seem to imply otherwise, the technical options must include non-technical solutions. For example, the analyst team could conclude that a specific data quality problem is harder to solve than the data element is worth, and the best solution could be to abandon the data element in favor of business process modifications.

Technical documentation should take full advantage of existing data dictionaries and similar resources. The focus needs to be on the particular data elements involved in the data quality solution selected for implementation. Most CVISN data elements are already documented, often in multiple places. This data quality improvement methodology will drag the documentation along with it, patching one data element at a time rather than making a large investment in creating or updating whole document sets.

Progress is reportedly strong on the Department of Justice data dictionary program Global Justice XML (eXtensible Markup Language) Data Model. Among the capabilities listed for their approach is that it “Stores and maps data element requirements; this enables tracing and tracking to source data components and measurement by source data requirements.” In addition it “maintains metadata for XML data dictionary registry” and “provides search filters, maintenance forms, and tools.”

Especially considering the concerns expressed about integration of justice-related driver data, the Justice XML data model should be seriously considered. [Reference: http://it.ojp.gov/gjxdm/] The Justice XML Web site reports that American Association of Motor Vehicle Administrators (AAMVA) is a user of the system for “Standardization of Driver's License records across National Law Enforcement Telecommunication System (NLETS) – The International Justice & Public Safety Information Sharing Network.” Resource materials are available free of charge (http://it.ojp.gov/jxdm/), and technical assistance is available (http://it.ojp.gov/topic.jsp?topic_id=192).

For certain identified data quality problems, where data elements are used across state and federal systems, optimal technical solutions may require coordination with the FMCSA COMPASS team, especially DataQs.

Technical approaches based on constraint-checking must be coordinated with pending change requests and related on-going technical activities of the SAFER and PRISM support teams.

The working group recommends that a project be initiated to implement the process identified in Option 1, starting with a focused effort on a high-priority business process where data quality issues have a high impact. Two teams should be assembled: a business process review team and a technical team. As described in Section 5.1, the teams should be assigned different tasks but must interact with each other. The business process review team should include stakeholders from government, enforcement, industry, and service bureaus. The technical team should include stakeholders from government, enforcement, industry, service bureaus, planners, and those who would be involved in implementing the kinds of changes identified in this report. Representatives from appropriate standards development organizations should be consulted as the requirements for safety information sharing are clarified. Members of the Enhanced Safety Information Sharing Working Group should be invited to participate on the teams. The working group should at least serve in an advisory capacity to the teams' efforts.

An outline of the steps for the project follows:

- Select a high-priority business process where data quality issues have a high impact. The working group suggests these processes as possible candidates: e-screening enrollment, e-screening bypass, selecting carrier/vehicle/driver for inspection, selecting carrier for compliance review.
- Establish business and technical teams to tackle: review of lessons learned; alignment of business processes; universal data dictionary; common identifiers; definition of standards for information sharing that include structure and protocols; improve constraint checking.
- Define requirements for new or modified systems and business processes.
- Analyze technical alternatives for implementing the changes.
- Choose alternatives and assign responsibility and resources to implement.
- Implement, test and deploy the chosen alternatives.
- Measure the resulting improvement in data quality.
- If quality improvement is demonstrated, return to step 1 and iterate. If no improvement is demonstrated, rethink the alternatives.

Idaho, Maryland and Wisconsin expressed interest in participating in this activity.