U.S. Department of Transportation
Federal Highway Administration
Research and Innovative Technology Administration

State-of-the-Practice and Lessons Learned on Implementing Open Data and Open Source Policies

www.its.dot.gov/index.htm

Final Report — May 2012
FHWA-JPO-12-030

Produced by the John A. Volpe National Transportation Systems Center
U.S. Department of Transportation
Research and Innovative Technology Administration
ITS Joint Program Office

Notice

Acknowledgements

The Volpe Center team would like to acknowledge the leadership of Walter During, P.E., of the Office of Transportation Management (HOTM) within the Office of Operations, Federal Highway Administration, U.S. Department of Transportation, in providing the guidance necessary to conduct the review and analysis of lessons learned that form the basis for this document.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-JPO-12-030		

4. Title and Subtitle	5. Report Date
State-of-the-Practice and Lessons Learned on Implementing Open Data and Open Source Policies	May 2012
	6. Performing Organization Code

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. Performing Organization Report No.
Aviva Brecher, Matt Cuddy, Josh Hassol, and Suzanne Sloan	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. Work Unit No. (TRAIS)
U.S. Department of Transportation Research and Innovative Technology Administration John A. Volpe National Transportation Systems Center Cambridge, MA 02142	11. Contract or Grant No. HW4A3

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Highway Administration (FHWA) Office of Advanced Travel Management U.S. Department of Transportation 1200 New Jersey Ave., S.E. Washington, D.C. 20590	Policy Analysis, 2011-2012
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report describes the current government, academic, and private sector practices associated with open data and open source application development. These practices are identified; and the potential uses with the ITS Program's Data Capture and Management (DCM) Program and Dynamic Mobility Applications (DMA) Program are noted. The report attempts to present examples and lessons learned in a manner targeted at comprehensively addressing the critical policy and institutional issues that are analyzed and addressed in a series of separate reports. The practices herein provide a basis for determining the most appropriate or new policies for an open data and open source approach for the DCM and DMA programs.

17. Key Words	18. Distribution Statement
Open data and open data environments, open source portals, dynamic mobility applications, core system, vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), connected vehicle environment, privacy, liability, user access controls, meta-data classification, standards, user registration, project forking, governance, intellectual property, data capture and management, policy and institutional issues, data capture and management, dynamic mobility applications, data quality assurance	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
None	None	77	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Table of Contents

Executive Summary

This report presents the results of a review of current practices in open data and open source in order to develop working definitions and construct useful lessons that may be applied to the U.S. Department of Transportation's (USDOT) Data Capture and Management (DCM) and Dynamic Mobility Applications (DMA) programs. These programs are providing foundational research for the Intelligent Transportation Systems (ITS) Program's connected vehicle environment. Their focus is to address technical and policy issues associated with technological innovations in the capture and management of real-time transportation data and the innovative uses of those data for transformative mobility and system management applications.

The report covers the following areas, all of which are defined in the Introduction. For each state-of-the-practice area a definition and state-of-the-practice examples are provided, along with a recommended policy option.

Metadata

Definition: The National Information Standards Organization (NISO) defines metadata as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource."[1] Often metadata are called "data about data." These data may be textual, numeric, spatial, verbal, or visual. For example, metadata have been used in the traditional library cataloging system where a library catalog card displays a book's title, author, subject matter, brief synopsis, and alpha-numeric identification. Such data help classify, aggregate, identify, and locate a particular book. Using metadata to organize the wealth of information that could potentially be archived from ITS systems helps users understand which data apply to them and whether those data are appropriate for their needs.

Recommended Policy Option: Metadata standards minimize the potential for inconsistency caused by creation or modification of metadata by numerous participants. Adoption of the ASTM International Standard Practice for Metadata to Support Archived Data Management Systems (E 2468-05),[2] as has been done for the ITS/JPO Research Data Exchange, is logical to support both research on and operationalization of metadata to support DMA/DCM applications. Furthermore, use of a modified Dublin Core[3] as the metadata schema, as the National Transportation Library (NTL) does, is recommended. Dublin Core is a well-refined, widely used, and effective approach to developing metadata. Both of these elements have widespread industry acceptance, which will a smooth transition to private sector development of DCM/DMA applications.

[1] NISO. Understanding Metadata, http://www.niso.org/publications/press/UnderstandingMetadata.pdf

2 http://www.standards.its.dot.gov/fact_sheet.asp?f=73

[3] http://dublincore.org/

Security

Definition: Security encompasses two aspects:

- Data security: Ensuring that data are kept safe from corruption and that access to them is suitably controlled.
- Data Environments and System security: The use of software, hardware, and procedural methods to protect data systems from external threats.

Security measures that are built into applications, in conjunction with a sound security routine, minimize the likelihood that hackers will be able to manipulate applications to access, steal, modify, or delete sensitive data. They form the foundation of a secure operating environment and help protect personal data.

Recommended Policy Option: A robust policy framework and guidelines exist to ensure security of data sets, data environments, and the hardware and software associated with both. In particular, ongoing adherence to the Federal Information Security Management Act (FISMA), [4] as well as data security guidelines from the National Institute of Standards and Technology (NIST) can effectively address security concerns.

Privacy

Definition: Privacy policies predominantly reflect policies for the protection, handling, and use of personally-identifiable information (PII). For the DMA and DCM programs, a challenge is the protection of data within an environment based on open data policies. The primary PII associated with the dynamic mobility applications is locational data, but a few of the applications require account or financial information, may include information that is highly competitive, or may contain medical information that might be linked to an individual.

Recommended Policy Option: Developing a privacy policy requires development of Fair Information Practices (FIPs). A recent NIST publication[5] provides a well-organized "toolbox" of approaches for mitigating privacy risks and which are based on best practices around the world. Included in these practices are Privacy Enhancing Technologies which support designing privacy into systems, technologies, and applications. Also, because data privacy is highly interrelated with data security (i.e., an organization cannot have data privacy without first establishing a solid data security foundation), policies, and guidelines for security are a key element of a privacy policy.

User Access Policies and Controls

Definition: User Access policies specify who can access, use, and contribute to a website. User Access is often defined in legal documents called "User Access Agreements" (also sometimes referred to as "User Agreements," "Terms of Service" or "Terms of Use"). These documents set the terms and conditions under which users are permitted to access and use

[4] http://csrc.nist.gov/groups/SMA/fisma/overview.html
[5] http://csrc.nist.gov/publications/nistpubs/800-53-Rev3/sp800-53-rev3-final_updated-errata_05-01-2010.pdf

content on a website or portal. User Access Agreements also often contain a user code of conduct and statement of warranties (e.g. user warrants that they will only post content that they have rights to post, user warrants that they are solely responsible for their conduct and content). The specific content of a website's user access agreement will vary based on the nature of the specific services offered.

Recommended Policy Option: Careful consideration of user access issues and risks at the program, portal, and project level is recommended, along with development of tiered user access policies that match the level of restrictiveness to the sensitivity of various data, source code, or documentation. At the development portal level, Forge.mil[6] provides a good model for developing high-security user access policies, whereas ITdasboard.gov represents a well-designed approach to access in the case of low-security, public access data.

Data Quality Assurance

Definition: The data will support development of effective applications only if the data are of consistently high quality. The DMA application developers and other developers and system users must have confidence that the data exchange has policies and procedures in place to ensure data quality.

Data quality assurance is the process of profiling data to discover inconsistencies and other anomalies, and performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve data quality. These activities can be undertaken as part of data warehousing or as part of the Database administration of an existing piece of applications software

Recommended Policy Option: DCM / DMA data system and data set "owners" will find the Bureau of Transportation Statistics guidelines to be a recommended foundation for the development of data quality assurance protocols. State-of-the-practice examples provided in this whitepaper offer insights into different approaches to data quality assurance. The Center for AIDS Research Network of Integrated Clinical Systems,[7] and the National Data Buoy Center[8] stand out as having a particularly rigorous quality processes in place that make creative use of automated techniques for real-time data quality verification. It is likely that the DCM / DMA programs will need to go well beyond even this level of data quality assurance, to include formalized protocols for data review, error documentation, and error correction.

Intellectual Property

Definition: Licensing of intellectual property, and, in particular, the application of open source licensing models is among the most important issues the DCM and DMA programs will face. Open source applications raise unique Intellectual Property (IP) liability concerns. To meet Mobility Program goals, the Mobility Program must acquire and preserve the right to

[6] Forge.mil is an online portal for Agile open source application development in support of the Department of Defense (DoD).
[7] http://www.cnics.net/
[8] http://www.ndbc.noaa.gov/

provide developed applications under open source terms. Without these terms, the Mobility Program will infringe upon the intellectual property rights of the software developer, as will any downstream developers or users employing software acquired through the Mobility Program.

Recommended Policy Option: A thorough, well-documented and clearly communicated IP policy framework is necessary to provide all participants in the DCM / DMA application development efforts with a clear understanding of the rules of the game with respect to licensing, patents, and other aspects of intellectual property protection. This will be challenging, given the large number of participants and applications envisioned. Forge.mil[9] provides the recommended model for open source application development to support major public-sector initiatives.

Liability

Definition: In the context of the DCM / DMA programs, there are two types of liability that are not addressed through IP licensing:

1. Liability stemming from software or data quality problems (e.g., inaccurate data, applications failing to work as intended), or misuse of the data.
2. The potential liability that exists if security is breached and exposes PII or individual locational information (such as GPS data).

Recommended Policy Option: With regard to risk with open data, data errors, and unintentional system problems, the US DOT's connected vehicle legal policy team is exploring the concept of a shared risk environment. With regard to intentional misuse or exposure of data, this same team is analyzing the extent and applicability of existing tort law. The recommendation is to apply the results of this analysis to the DCM and DMA programs.

Governance

Definition: This report discusses governance from three perspectives.

1. **Data governance** is a set of processes that address quality, management, policies, standards, metadata organization, and other issues associated with data.[10] A governance structure frames roles and responsibilities in relation to authority (i.e., scope, sanctions, and enforcement), rules of conduct, standards, and metadata. The governance model offers a structure to define which people and entities can take what actions, with what information, under what circumstances, using what methods.[11] It also establishes the means by which those "governed" are able to influence the overall scope and decisions of the governing body, as well as mechanisms for appeal and/or adjudication of contestable actions.

[9] www.forge.mil
[10] Sarsfield, Steve (2009). "The Data Governance Imperative," **IT Governance.**
[11] Data Governance Framework. Data Governance Institute. At http://www.datagovernance.com/dgi_framework.pdf

2. **Project governance** is a framework for decision-making and management of a project. The governance structure of a project determines the roles and responsibilities of the participants, with a particular emphasis on how decisions are made. In the world of open source software development, project governance establishes the rules by which collaborators may contribute to a project, how contributions will be evaluated and accepted/rejected, and how disputes will be resolved.
3. **Portal governance** is a web portal or links page is a website that functions as a point of access to information on the World Wide Web. It presents information from diverse sources in a unified way. In the context of open source application development, a portal contains the tools through which the contributors, users, testers, and project leaders interact (e.g. source code repository, wiki, forums, bug tracker). Our research found no published material on how open source application development portals are specifically governed.

Recommended Policy Option: Data.gov[12] is recommended as a model for effective data governance, and Forge.mil[13] for project governance.

Open Data Maintenance

Definition: Maintaining the huge volume of data available via open data environments is critically important. Without ongoing attention, existing data sets can quickly become outdated and inaccurate – especially transportation data, which are frequently and rapidly changing, and which are increasingly flowing into data environments from myriad mobile sources.

Recommended Policy Option: Ongoing review and updating of the open data supporting the DCM /DMA programs and applications will be a significant, and vital, undertaking. Therefore, data maintenance policies and procedures need to be implemented in advance, and must establish the protocols in three key areas: data review, data monitoring and assessment, and data updating.

[12] http://www.data.gov/
[13] http://forge.mil/

Introduction

Intent of Report

This report presents the results of a review of current practices in open data and open source in order to develop working definitions and construct useful lessons that may be applied to the U.S. Department of Transportation's (USDOT) Data Capture and Management (DCM) and Dynamic Mobility Applications (DMA) programs. These programs are providing foundational research for the Intelligent Transportation Systems (ITS) Program's connected vehicle environment. Their focus is to address technical and policy issues associated with technological innovations in the capture and management of real-time data (in particular, data generated by vehicles in motion) and the innovative uses of those data for transformative mobility and system management applications.

The goal in applying open data and open source practices to DMA and DCM is to expedite the development, testing, commercialization, and deployment of innovative mobility applications, fully leveraging both new technologies and federal investment to transform transportation system management, maximize the productivity of the system, and enhance the accessibility of individuals within the system. Program objectives also include the active acquisition and systematic provision of integrated, multi-source data to enhance current operational practices and transform future surface transportation systems management.

"Open" refers to the philosophy (or approach) that data, software, and other products should be free and accessible to anyone without restrictions or controls (although with clear attribution of intellectual property and clear permissions, typically described in terms of a license). "Open" also refers to a software development approach that allows the software product and source-materials to be available to other developers through virtual communities, thereby enabling developers to collaborate on product development or collectively enhance the software." Finally, the philosophy of "openness" is consistent with the President Obama's goal of making federal transportation policy "transparent and accountable to the American public, performance-based, focused on achieving strategic outcomes, and maximizing the value of public investments."[14]

Both of these approaches are embodied within the vision of what the USDOT hopes to achieve with its research in data capture and dynamic applications.[15] Both of these approaches are envisioned to play a role in executing the technical research for the DCM and DMA programs, and will continue to be considered in application to the extent that they offer practicable options in research, implementation, and operations.

This report offers definitions of practices for the DCM and DMA programs' consideration as well as examples of and insights into ways in which important technical, policy, and institutional issues can be addressed by using open approaches. Additional policy reports will draw from these practices to provide options and targeted recommendations for developing and implementing an

[14] US DOT Strategic Plan, at: http://www.dot.gov/stratplan/dot_strategic_plan_10-15.pdf

[15] The vision documents are located at: http://www.its.dot.gov/data_capture/datacapture_management_Federalrole7.htm and http://www.its.dot.gov/data_capture/datacapture_management_vision1.htm and http://www.its.dot.gov/dma/dma_vision2.htm

Open Source Applications Development Portal (OSADP) for the DMA program, developing a Research Data Exchange (RDE) and operational practices for the DCM program, and developing new and dynamic mobility applications within the OSADP.

Content of Report

The practices profiled in this report address the following issues, which were discussed in a previous paper, Identification of Critical Policy Issues for the Data Capture and Management (DCM) and Dynamic Mobility Application (DMA) Programs:[16]

- **Metadata**: These "data about data" make it easier and more efficient to manage large data sets, particularly those that integrate diverse data elements.
- **Securit**y: With respect to the ongoing development of DCM and DMA, a key challenge is how to maintain high levels of security while preserving the fundamental flexibility and "openness" of open data and open data environments.
- **Privacy**: The issue of privacy in the context of DMA and DCM applications refers to the ways to protect any data that contains personally identifiable information (PII) or that can be used to track the location of an individual. It also refers to how privacy principles are implemented to result in transparency about the use of data and more effective management practices.
- **User Access Policies and Controls:** A component of security, access policies and controls refer to mechanisms by which systems grant or revoke the right to access some data, or perform some action. There is a fundamental tension between controlling access and fostering open data environments. Consequently, developing appropriate access practices may be particularly challenging.
- **Data Quality Assurance:** The data will support development of effective applications only if the data are of consistently high quality. The DMA application developers and system users must have confidence that the data exchange has in place policies and procedures to ensure data quality.
- **Intellectual Property:** This broad term encompasses four distinct areas of law: patents, copyrights, trademarks, and trade secrets. All of these have direct bearing on the development of open data and data environments.
- **Liability:** For the DCM and DMA programs, open data and open-source applications raise primarily two types of liability concerns:
    - With respect to DCM, liability potentially exists if security around personally identifiable information (PII) or individual locational information (such as GPS data) is breached.

[16] Produced by the USDOT's John A. Volpe National Transportation Systems Center for the Federal Highway Administration (FHWA) and the ITS Joint Program Office, October 2011.

  - Liability concerns with respect to DMA center on intellectual property infringement and liability stemming from data quality problems e.g., applications failing to work as intended), or misuse of the data.

- **Governance**: Three approaches to governance concern the DCM and DMA programs—data governance, project governance, and portal governance. Data governance is a set of processes that address quality, management, policies, standards, metadata organization, and other issues associated with data. The governance of open data is of increasing interest within the Federal government because of the Open Government Initiative, which pushes government agencies to make high-value data more freely available to the public online. Project governance is the establishment of a framework that assigns roles and responsibilities to a set of stakeholders for decisions, conflict resolution, performance metrics, and other actions with regard to specific projects (particularly applies to user development communities around a project). Portal governance is similar except that the purview is the portal and its operations as opposed to specific projects. Both project and portal governance can range on a spectrum from full control (known as "benign dictatorship") to collective governance that is based on levels of participation and commitment (known as a "meritocracy").
- **Data Maintenance:** Maintaining the huge volume of data available via open data environments is critically important. Without ongoing attention, existing data sets can quickly become outdated and inaccurate. This problem is particularly acute for transportation data, which often are rapidly changing, and which increasingly are flowing into data environments from myriad mobile sources.

Chapters 1 through 9 provide a discussion of each of these issues, and then highlight one or more current examples of how each is being addressed. The examples do not constitute an exhaustive survey of the literature, but rather a sample, focusing on those examples that may be most instructive for addressing the challenges and for establishing policy for the DCM and DMA programs.

Chapter 10 provides considerations for applying or formulating open data policies. This discussion draws on recently published Office of Management and Budget (OMB) and Department of Transportation (DOT) guidance documents, along with publications by other entities.

The conclusion provides an analysis of how the state-of-the-practice examples might apply to the DCM and DMA programs.

Relationship to other Connected Vehicle Mobility Policy Reports

This report is one in a series of six policy reports that describe and analyze the policy issues associated with connected vehicle mobility. The series includes:

- Two foundational reports that identify the critical issues and describe the best practices and lessons learned from government, industry, and academia:
    - ***Identification of Critical Policy Issues for the Mobility Program***, FHWA-JPO-12-035
    - ***State-of-the-Practice and Lessons Learned on Implementing Open Data and Open Source Policies*** (this report), FHWA-JPO-12-030

- Four reports that analyze the specific policy issues in context of the goals of the DMA and DCM programs:
    - ***Policy Analysis and Recommendations for the Open Source Applications Development Portal (OSADP)***, FHWA-JPO-12-031
    - ***Policy Analysis and Recommendations for Development of the Dynamic Mobility Applications***, FHWA-JPO-12-033
    - ***Policy Analysis and Recommendations for the DCM Research Data Exchange***, FHWA-JPO-12-036
    - ***Privacy and Security Analysis and Recommendations for the DCM and DMA Programs***, FHWA-JPO-12-032

Chapter 1 Metadata

Recommended Policy Option

Adoption of the ASTM International Standard Practice for Metadata to Support Archived Data Management Systems (E 2468-05), as has been done for the ITS/JPO Research Data Exchange, is logical to support both research on and operationalization of metadata to support DMA / DCM applications. Furthermore, use of a modified Dublin Core as the metadata schema, as the National Transportation Library (NTL) does, is recommended. Dublin Core is a well refined, widely used, and effective approach to developing metadata.

Both of these elements have widespread industry acceptance; this will help ensure a smooth transition to private sector development of DCM/DMA applications.

Collaboration with the TRB Metadata Working Group to ensure consistency of standards is recommended.

Definition

The National Information Standards Organization (NISO) defines metadata as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource."[17] Often metadata are called "data about data." These data may be textual, numeric, spatial, verbal, or visual. For example, metadata have been used in the traditional library cataloging system where a library catalog card displays a book's title, author, subject matter, brief synopsis, and alpha-numeric identification. Such data help classify, aggregate, identify, and locate a particular book.

Using metadata to organize the wealth of information that could potentially be archived from ITS systems helps users understand which data apply to them and whether those data are appropriate for their needs.

There are three main types of metadata:

- Descriptive metadata: describe the data set for purposes of discovery and identification. They can include elements such as title, author, keywords, date, type, format, etc.
- Structural metadata: describe the structure of the data set indicating how the data are put together – for example, how pages are ordered to form chapters.
- Administrative metadata: describe information to help manage a data set, such as when and how it was created, file type, file size, and other technical information. There are two subsets of Administrative metadata:

[17] NISO. Understanding Metadata, http://www.niso.org/publications/press/UnderstandingMetadata.pdf

- Rights management metadata: describes the intellectual property rights providing copyright ownership, user privileges, restrictions, etc.
- Preservation metadata: contains information needed to archive and preserve the data set.

As defined by NISO, "interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality." The use of metadata promotes efficiency and interoperability by using defined metadata standards and crosswalks between standards. According to a recent report published by the Transportation Research Board (TRB) Metadata Working Group:

> . . . metadata can allow data producers to maintain control of how data are used. At the same time, it is a method for users of data to track where data can be found, how it might be accessed, what elements it contains, spatial and temporal timeframes of the data sets and what forms it is in, and thus, whether it is compatible with data user objectives or with other data sets. Metadata are quite valuable, saving data users both time and money. [18]

Metadata: State-of-the-Practice Examples and Policy Options

The practice of establishing metadata can be difficult to define as it is highly dependent upon the context in which it is used. A survey of available information indicates that there are many different metadata classification practices, standards, tools, and implementation examples but, importantly, few examples relating specifically to transportation metadata. Over the course of the next year, as the DCM and DMA programs develop and implement their own metadata strategies in support of the RDE and the OSADP, the following select examples may be candidates for application and/or modification for the program's purposes.[19]

Metadata with Transportation Applicability

North American Profile (NAP) of ISO 19115 "Geographic Information - Metadata"

North American Profile of ISO 19115 is a standard of the International Organization for Standardization (ISO) and a component of the ISO 19115 standards series for Geospatial metadata. ISO 19115 defines how to describe geographical information and associated services, including contents, spatial-temporal purchases, data quality, access and rights to use. The

[18] Transportation Metadata: Role of Data and Information Technology Section, Transportation Research Board – Metadata Working Group, 3/3/2006 located at:
http://www.nymtc.org/data_services/Data%20coordination%20files/final%20report%20Jun%201%202005.pdf

[19] Many other metadata standards are not listed as their practices and standards applied more specifically to library sciences, education, archiving, e-commerce, and the arts. Additionally, some of their practices appear similar to the examples included in this report.

standard defines more than 400 metadata elements, and 20 core elements including contents, spatial-temporal purchases, data quality, access, and rights to use.[20]

Content Standard for Digital Geospatial Metadata (CSDGM)

The CSDGM is the current US Federal Metadata standard. According to the Federal Geographic Data Committee (FGDC), "…all Federal agencies are ordered to use this standard to document geospatial data created as of January 1995. The standard is often referred to as the 'FGDC Metadata Standard' and has been implemented beyond the federal level with State and local governments adopting the metadata standard as well."[21]

The Content Standard for Digital Geospatial Metadata is the current standard, but ISO 19115 was formally adopted by the American National Standards Institute (ANSI) in June 2009. Once work on this standard is finished the FGDC will process the NAP as a federal standard and promote the implementation to the geospatial community. If this standard is adopted, nonfederal organizations will be obligated, as with the CSDGM, to create NAP compliant metadata if they apply Federal funds to development of geospatial data.

A recent example of transportation-related use of the Content Standard for Digital Geospatial Metadata comes from Washington State DOT, which used this standard to create their Washington State Transportation Framework (WA-TRANS). The data set in question (WA-TRANS) is a large geo-data set, which requires as complete documentation as possible to enable it to be stored and retrieved accurately.[22]

Dublin Core Metadata Element

This set of metadata elements provides a small and fundamental group of text elements through which most resources can be described and catalogued. Using only 15 base text fields, a Dublin Core metadata record can describe physical resources such as books, digital materials such as video, audio, image or text files, and composite media like web pages. Metadata records based on Dublin Core are intended to be used for cross-domain information resource description and have become standard in the fields of library science and computer science. Implementations of Dublin Core typically make use of XML and are Resource Description Framework based."[23]

The NTL Digital Repository contains digital objects and links to external websites and the catalog uses a modified Dublin Core[24] as the metadata schema. All elements in Dublin Core are optional and repeatable. The modified Dublin Core elements for the NTL Digital Repository are: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, and Edition. Also, Data.gov uses an adaptation of the Dublin Core metadata standard to display the metadata on Data.gov.

---

[20] http://en.wikipedia.org/wiki/ISO_19115. http://www.fgdc.gov/standards/projects/incits-l1-standards-projects/NAP-Metadata/napMetadataProfileV11_7-26-07.pdf/view

[21] http://www.fgdc.gov/metadata/geospatial-metadata-standards#valueofstandards

[22] http://www.wsdot.wa.gov/mapsdata/TransFramework/project_documents/WA-Trans_metadataStandards_Current.pdf

[23] http://dublincore.org/

[24] The name "Dublin" is due to its origin at a 1995 invitational workshop in Dublin, Ohio; "core" because its elements are broad and generic, usable for describing a wide range of resources. http://dublincore.org/

Other Metadata Standards

Data Documentation Initiative (DDI)

DDI is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.[25]

Statistical Data and Metadata Exchange (SDMX)

SDMX aims to foster standards for the exchange of statistical information and data. For example, Transmet, a website for metadata on Australian transportation data collections has developed the metadata statements in accordance with SDMX standards.[26]

ASTM International E2468-05 Standard Practice for Metadata to Support Archived Data Management Systems (ADMS)

This standard practice establishes the recommended metadata framework for archived data management systems and provides additional commentary and examples to assist ADMS developers and users.[27] It applies to data stored in archived data management systems. It includes metadata to describe the structure of the archive itself as well as the conditions under which the data were originally collected and processed. This standard is arranged into the following sections:

- Identification Information
- Data Quality Information
- Spatial Data Organization information
- Spatial Reference Information
- Entity and Attribute Information
- Distribution Information
- Reference Information

Lessons Learned

As is the case with most efforts to develop standards, achieving consensus on definition, implementation and use is perhaps the biggest challenge in the development of transportation metadata standards. According to the TRB Metadata Working Group, in addition to coordinating with other standards development activities, major hurdles include the following:

[25] http://www.ddialliance.org/what
[26] http://www.nss.gov.au/transportmetadata/standards.jsp
[27] http://www.standards.its.dot.gov/fact_sheet.asp?f=73

- Addressing the diversity of data types, sources, and users within the transportation community
- Reaching agreements on standards. Many of other metadata standards are purposely general in order to be flexible and adaptable. Consequently, there is a need for the development of "global" high-level metadata standards.
- Overcoming implementation barriers so that data producers routinely build in metadata and adhere to these standards consistently.

In support of the transportation community, the TRB Metadata Working Group is working to foster a consistent and coordinated approach to transportation metadata efforts. The Working Group's recent report emphasizes "…the need for the development of metadata standards and their potential benefit to transportation practices..." as such standards help all stakeholders fully understand all aspects of open data they use. The TRB Metadata Working Group report provides a strategic plan roadmap to establish a transportation metadata standard. According to the report, such a plan would:[28]

- Establish the need for the development of metadata standards and their potential benefits to transportation practice;
- Determine the general types of standards that are required and ascertain the policy implications of instituting metadata standards, especially barriers to acceptance;
- Identify other ongoing efforts that have an impact on the development of metadata standards for transportation data, and recommend how coordination may be achieved;
- Propose a schedule and recommend organizations (and other participants) that should develop and maintain metadata standards; and
- Identify methods to promote the use of completed metadata standards within the transportation community.

Conclusion

Metadata are critically important to guide both research and development of DCM/ DMA applications. As described in the DCM Program's vision document,[29] the data environments will have a requirement to include a high-level description of itself, what data types it contains, and general conditions under which data were captured. All these requirements are specific types of metadata.

This report has identified the creation of unified, high-level metadata standards as vital for the successful development of metadata. Adoption of the ASTM International E2468-05 standard for

[28] Ibid.
[29] Data Capture and Management Program Vision: Objectives, Core Concepts and Projected Outcomes located at: http://www.its.dot.gov/data_capture/datacapture_management_vision1.htm

metadata provides such a standard, although work remains to be done to produce the final set of metadata standards applicable to the DCM / DMA efforts that data providers and users from the public and private sectors can employ.

Next Steps

- Establish DCM / DMA metadata working group, comprising representatives from data provider and data user communities.
- Through the working group, identify key metadata standards issues and develop draft metadata standards for DCM / DMA data. Distribute draft standards to industry associations, government agencies and other stakeholders for review and comment.
- Collaborate with TRB metadata committee to ensure that US DOT's metadata standards are consistent with TRB's.
- Develop final DCM / DMA metadata standards. Modify ASTM International standard and Dublin Core as needed.
- Identify the similarities and differences in how the standard might be applied to the RDE and the OSADP. Determine if these differences might cause complications when transferring data between the two systems.
- Pilot test the standards to understand the level of burden they impose on data providers, and other potential barriers to adoption. Adjust standards where appropriate.
- Develop guidance documents for data suppliers, outlining the standards and their application.
- Agree on schedule for implementation of standards.

A final consideration for this chapter is that the TRB Metadata Working Group might provide a well-organized forum for tackling the issue of metadata standards; active collaboration with the Working Group is proposed.

Chapter 2 Security

Recommended Policy Option

Implement a robust policy framework and guidelines to ensure security of data sets, data environments, and the hardware and software associated with both. The framework and guidelines should ensure adherence with the Federal Information Security Management Act (FISMA), as well as data security guidelines from the National Institute of Standards and Technology (NIST) to effectively address DCM / DMA security concerns.

Definition

Security encompasses two aspects.

- Data security: Ensuring that data are kept safe from corruption and that access to them is suitably controlled
- Data Environments and System security: The use of software, hardware, and procedural methods to protect data systems from external threats[30].

Security measures that are built into applications, in conjunction with sound security protocols, form the foundation of a secure operating environment and help protect personal data. They minimize the likelihood that hackers will be able to steal, modify, or delete sensitive data.[31]

Open Data and Open Data Environments Security: State-of-the-Practice Examples

Federal Information Security Management Act (FISMA)

Title III of the E-Government Act of 2002, FISMA "requires each federal agency to develop, document, and implement an agency-wide program to provide information security for the information and information systems that support the operations and assets of the agency, including those provided or managed by another agency, contractor, or other source."[32] FISMA takes a programmatic approach to information security; it includes the following elements:

[30] http://searchsoftwarequality.techtarget.com/definition/application-security
[31] http://en.wikipedia.org/wiki/Data_security
[32] http://csrc.nist.gov/groups/SMA/fisma/overview.html

- Periodic risk assessments
- Policies and procedures based on the risk assessments
- Security awareness training
- Periodic testing and evaluation of security measures
- Process for remediating security deficiencies
- Procedures for detecting and reporting security breaches
- Continuity of operations plans

CONNECT

CONNECT is an open-source software and community that promotes IT interoperability in the U.S. healthcare system. Using security controls defined in the Nationwide Health Information Network services (NHIN), CONNECT enables secure electronic health data exchange among healthcare providers, insurers, government agencies and consumer services. It is the result of collaboration among Federal agencies and is coordinated through the Federal Health Architecture (FHA) Program.

Because CONNECT deals with sensitive personal information, security has been a key aspect in its development. Consequently, CONNECT has become a best practice for open data security.[33] A CONNECT Fact Sheet, available via the CONNECT web page, includes the following description of CONNECT's development and release of open source software:

> CONNECT is a Federal Health Architecture (FHA)[34] project that began in 2007 to share health-related data among federal agencies and their partners. CONNECT was built collaboratively by more than 20 federal agencies – each sharing the common goal of improving health information exchanges by making them more reliable and secure. Joint development eliminated duplicate, disjointed efforts and dramatically reduced costs for the federal government.
>
> After developing CONNECT in less than a year, federal agencies released CONNECT as an open source software for use throughout the industry. Today, the CONNECT open source community is comprised of more than 2,000 organizations – including federal agencies, states, healthcare providers, insurers, health IT vendors.[35]

[33] For example, experts at a USDOT-hosted Governance Roundtable in June 2011 frequently cited the CONNECT and Health IT governance framework.

[34] FHA is a coalition of Federal agencies that advance health IT interoperability within Federal agencies and outward with state, tribal, local and private sector organizations.

[35] CONNECT fact sheet at: http://healthit.hhs.gov/portal/server.pt?open=512&mode=2&objID=3340

The controls in use by CONNECT include the server-based Public Key Infrastructure (PKI) and the NHIN service registry which define and secure the NHIN backbone. The elements that meet the security and organizational requirements are:

- The **messaging platform and authorization framework** implement additional security and privacy controls to address the known threats for Web services implementations of service-oriented-architectures.
- The **audit log query service** is designed to meet the requirements for Health Insurance Portability and Accountability Act (HIPAA) disclosure accounting.
- The **consumer preferences profile** allows consumers to express their preferences for whether or not to share their information on the NHIN and for more granular control over access to their private information.
- The **CONNECT policy engine** enforces those preferences in the runtime environment to ensure that the access policies of the organization and the preferences of the consumer are honored in the decision to release health information in response to a request from the NHIN.

Federal agencies using CONNECT must adhere to FISMA (Federal Information Security Management Act of 2002) requirements in addition to meeting the HIPAA requirements. CONNECT has been engineered to meet these exacting security requirements and is undergoing the United States Department of Health and Human Services (HHS) Security Certification and Accreditation (C&A) process. Those implementing CONNECT are required to undergo a C&A in order to get an authority to operate in their environment; they will be able to leverage the security testing that CONNECT has undergone for the HHS C&A to speed them through their own process. Private sector organizations using CONNECT get the benefit of a solution that is built to meet the stringent requirements that the federal agencies must meet in their operational systems.[36]

Cloud Security Alliance

Cloud computing security practices are germane to the security challenges facing the open data and environments envisioned for the DCM and DMA programs, because cloud computing environments must enable the delivery of computing services via shared resources, software, and information. The Cloud Security Alliance (CSA) is "a not-for-profit organization with a mission to promote the use of best practices for providing security assurance within cloud computing, and to provide education on the uses of cloud computing to help secure all other forms of computing. The Cloud Security Alliance is led by a broad coalition of industry practitioners, corporations, associations and other key stakeholders."[37] One of the CSA's ongoing research areas is Security Guidance for Critical Areas of Focus in Cloud Computing, and CSA has several excellent resources on this topic, including an eponymous white paper that addresses security with respect to system architecture, governance, and operations.[38]

[36] http://www.connectopensource.org/about/governance
[37] https://cloudsecurityalliance.org
[38] https://cloudsecurityalliance.org/wp-content/uploads/2011/07/csaguide.v2.1.pdf

NIST Cloud Computing Security Guidelines

NIST report 800-144[39] addresses key security issues relevant to cloud computing. This document makes the important point that "Because cloud computing has grown out of an amalgamation of [existing] technologies . . . many of the privacy and security issues involved can be viewed as known problems cast in a new setting." Specifically, because applications and IT infrastructure are decentralized under the cloud computing paradigm, maintaining security is more challenging. The NIST report details nine areas requiring special attention in the cloud computing environment:

1. Governance
2. Compliance
3. Trust
4. Architecture
5. Identity and Access Management
6. Software isolation
7. Data Protection
8. Availability (continuity of operations)
9. Incident response

Data.gov

The Open Government Initiative's Data.gov Privacy Policy[40] addresses the website's protection of both the security and the privacy of visitors to the website. Under this policy, Data.gov gathers specific user data to measure the number of visitors to the various sections of the site and to identify system performance or problem areas. Users are not required to provide any information to search, retrieve, download, filter, and otherwise use the data available on Data.gov. Some optional uses exist that require a user account. To protect privacy and prevent PII breaches, raw data logs are scheduled for regular destruction in accordance with National Archives and Records Administration (NARA) guidelines.[41]

Regular Checks on Security Threats and Patches

Protecting systems requires periodic updates for software and checking for emerging security threats. Two websites, provided by the Federal Trade Commission (FTC), include the Open Web

[39] http://csrc.nist.gov/publications/drafts/800-144/Draft-SP-800-144_cloud-computing.pdf
[40] http://www.data.gov/privacypolicy
[41] For additional information about NARA guidelines on destruction of electronic records, see: http://www.archives.gov/records-mgmt/pdf/dfr-2000.pdf and http://www.archives.gov/frc/flyers/e-media-destruction-faqs.html

Application Security Project at www.owasp.org or SANS Institute's Top Cyber Security Risks[42] at www.sans.org/top-cyber-security-risks for information on threats as well as solutions.

Lessons Learned

With respect to the development of DMA and DCM, a key challenge is how to maintain high levels of security while preserving the fundamental flexibility and "openness" of open data and open data environments. State-of-the-practice approaches indicate that achieving this essential balance is possible. Important differences exist, however, between existing systems such as CONNECT, and the envisioned connected vehicle infrastructure. In particular, the envisioned infrastructure will eventually be ubiquitous, will be mobile, and will comprise myriad users and numerous applications. Security approaches will need to be tailored to this environment. Current approaches to mobile broadband security and cloud computing security may provide useful analogs that could be tailored to meet the unique requirements of an open environment.
As with all IT security programs, user compliance is a potential "weak link" for DMA / DCM security. Protocols that minimize the need for voluntary compliance will be essential. An analogy is a system that automatically scans all email attachments for viruses, versus one that relies on recipients not to open attachments of unknown origin.

Conclusion

Security of open data and open data environments presents special risks and challenges, all of which will need to be addressed in the context of DCM / DMA. Fortunately, the security risks are well understood, and regulatory, policy, and standards frameworks exist to guide the development of effective security approaches. Nevertheless, DCM / DMA applications and data environments are complex and dynamic; security frameworks must be applied to meet the risks that are specific to these new systems, technologies, and applications.

Next Steps

- Guided by the critical areas highlighted by NIST, document the security risks for open data and data environments in the DCM / DMA programs.
- Develop procedures to mitigate the security risks.
- Pilot test the proposed security procedures to gauge user acceptance. This should include understanding the burden the proposed procedures impose on data providers, computing services providers, and public sector participants. Adjust standards where appropriate.
- Develop guidance documents for all participating entities, outlining the security procedures and their application.
- Develop schedule for implementation of security procedures.

[42] SANS stands for SysAdmin, Audit, Network, Security.

Chapter 3 Privacy

Recommended Policy Option

The NIST 800-53 Appendix publication provides a well-organized "toolbox" of approaches for mitigating privacy risks, the use of which is recommended to address individuals' concern for collection of PII. The toolbox guides an organization through implementation of the privacy practices that are appropriate to the level and type of personally-identifiable information (PII) at issues.

Federal organizations and their programs must, at a minimum, employ Federal policies that require the following protections: *notice; protection against unauthorized disclosures; the right of individuals to review their records and to find out if these records have been disclosed; the right to request corrections or amendments; assurances that the information collected or maintained is accurate, relevant, timely, and complete; and accountability for violations of personal privacy*.

For the DCM's RDE and the DMA's OSADP, in addition to developing a FIPPs implementation plan, managers may also consider use of Privacy Enhancing Technologies (PET), specifically, user identity management functionalities that include password resetting and management of lost passwords; and data masking or automated data-de-identification as data enters the data environments (if data is not anonymous when collected – a decision being driven by how vehicle-to-vehicle safety is implemented). For the applications, the recommendation during development is to ensure that only the minimum amount of PII data is needed for functionality.

A separate report that addresses privacy from the DCM and DMA perspectives will draw from these options to consider the risks and recommendations for the DCM and DMA technologies and applications.

Definition

Privacy policies predominantly reflect policies for use PII and the policies associated with the ways in which organizations treat information that is provided by the user. For the DMA and DCM programs, a further consideration is the set of policies that guide how the data warehouses protect any data that contains personally identifiable information (PII) or that can be used to track the location of an individual. Within the transportation environment, two elements of privacy are particularly sensitive:

- **Confidentiality** of PII contained in data transmitted for transportation management or personal mobility purposes.
- **Locational privacy** – issue is whether someone's present location or past patterns of movement can be determined through electronic means.

There are at least three categories of practices for protecting and managing privacy that are in use in today's world:

- **Developing and implementing policies** that define requirements for privacy protection and risk mitigation. (Privacy by Policy)
- **Designing privacy into systems, technologies and applications** to protect users' data and to provide users more control over use of that data. (Privacy by Technology)

- **Implementing appropriate physical and technical security measures** to mitigate the risk of exposure or breach. (Privacy by Security)

These three categories, importantly, are not exclusive, but are often combined in practice. The following defines the options and presents examples in two of these categories—Privacy by Policy and by Design. Security options were presented in Chapter 2.

The following examples provide the background and definitions that the upcoming report will reference.

Privacy Policies: State-of-the-Practice Examples

Privacy policies exist for Federal, State, and Local governments, private sector organizations, and academia.

Federal Policies and Practices

At the Federal level, the 1974 Privacy Act provides the Federal policy as to how Federal agencies approach privacy; privacy practices are implemented based on the analysis of what data are required for collection and whether or not they must contain any identifying information, the implementation practices that describe how the data will be stored/archived and who will have access to them, the policies and systems that will secure the data against unauthorized use, and the risks associated with data exposure. Along with this analysis is the commitment to transparency for users. This approach is known as Fair Information Principle Practices (FIPPs),[43] a practice developed by NIST for Federal computer systems and the leading best practice within the Federal government.

NIST identifies five over-arching principles for protection of PII. Organizations should:

- Identify all PII in their operating environments. While it may seem self-evident, a fundamental requirement for protecting data privacy is having clear knowledge of the data that are collected or maintained.
- Categorize the PII confidentiality impact level for all categories of its PII.

  NIST cites such factors as the as the ability of individuals to be personally identified, quantity or the number of individuals who could be identified, sensitivity of the data, legal and regulatory obligations for privacy protection of particular data sets, and the extent of access to data by individuals and systems.
- Apply appropriate safeguards based on the PII confidentiality impact level. NIST recommends operational safeguards, privacy-specific safeguards, and security controls consistent with varying levels of confidentiality impact. Safeguards include policies and procedures; training for staff having access to PII; de-identifying PII by removing identifiers that can be traced to specific individuals; controlling access, particularly for mobile devices; encryption of

[43] http://www.itl.nist.gov/fipspubs/geninfo.htm

information or communication; and auditing or monitoring processes and events that affect confidentiality.

Additional FIPP policies call for "openness about developments, practices and policies with respect to personal data," individual "participation" or rights to obtain data pertaining to that individual, and organizational accountability for complying with policies protecting PII. Consent, which is integral to the principle of individual participation, requires individual authorization of the collection, use, maintenance, and sharing of PII prior to its collection. Timely, uncomplicated, and inexpensive access to an individual's PII records is also fundamental to the principle of participation. Similarly, the organization must provide a process for individual redress or correction of inaccuracies in PII identified as a result of the review process.

The NIST Special Publication Recommended Security Controls for Federal Information Systems and Organizations (NIST 800-53)[44] includes an appendix (Appendix J)[45] dealing specifically with privacy. This appendix catalogs a "structured set of controls for protecting privacy" covering issues such as individual consent (for collection of PII), organizational authority to collect data, minimization of data (i.e., collecting as little PII as possible), and data disposal. These controls are intended for use by organizational privacy officials when working with IT staff and project managers. As such, Appendix J is an effective "toolbox" for protecting cyber-privacy.

Another reference for the ITS Program's connected vehicle environment is the 2007 document that describes the Vehicle Infrastructure Integration (VII): Privacy Policies Framework. The VII Program was the technical research program that engineered technology prototypes for testing; these technologies form the basis for the connected vehicle environment. The framework presents definitions for what constitutes privacy within a connected transportation environment, offering nine principles on which to base the technical research and development, and identifying limitations and exclusions. This framework is very similar to the NIST 800-53 guidelines and offers a first step in identifying options for implementing privacy for the connected vehicle mobility technologies and applications. However, with changes in the mobile environment (particularly relevant to the DCM and DMA programs) and new policies emerging regarding privacy,[46] this document will undergo review and update in 2012 with an eye toward ensuring a solid privacy foundation for implementation of the new safety , mobility and environment applications, data environments, and data sets.

In reviewing examples of federal agencies that have implemented the NIST policies, the Department of Homeland Security (DHS) is widely-cited as a solid example. The Federal Aviation Administration's (FAA) Aviation Safety Reporting System (ASRS) offers an interesting example of how the Federal government protects privacy under highly sensitive circumstances that require private sector personnel reporting safety violations or problems.

[44] http://csrc.nist.gov/publications/nistpubs/800-53-Rev3/sp800-53-rev3-final_updated-errata_05-01-2010.pdf

[45] http://csrc.nist.gov/publications/drafts/800-53-Appdendix-J/IPDraft_800-53-privacy-appendix-J.pdf

[46] In particular, the emerging research into the definitions for trusted identities and the identity ecosystem being led by NIST. Information can be found at: http://www.nist.gov/nstic/identity-ecosystem.html. The definitions of these leading-edge practices are being developed with industry and are similar to work done by the European Union over the last five years.

Department of Homeland Security (DHS) [47]

DHS's updated its FIPPs in 2008, to keep pace with changes in technology, and, where warranted, to make them consistent with the FIPPs of other countries. In keeping with the Department's mission, its FIPPs regulate PII tightly; but yet call on the Department itself to provide transparency, security, notice, and choice to users. The DHS is accountable and responsible for training all employees and contractors who use PII, while the FTC assumes an oversight role. The FIPPs do not clearly address how companies or online retailers should work with privacy laws or how governance should work. Instead, the FTC FIPPs employ an explicit auditing process to check compliance. Although serving different purposes, the DHS FIPPs are considered to be comprehensive and highly regarded by the privacy community, and somewhat stronger than the FTC principles.[48]

Aviation Safety Reporting System

To be effective and well-used by the intended audience, privacy is a key driver in the system and organization design of the Aviation Safety Reporting System (ASRS). Pilots, air traffic controllers, flight attendants, mechanics, ground personnel, and others involved in aviation operations submit reports to the ASRS when they are involved in, or observe, an incident or situation in which aviation safety may have been compromised. All submissions are voluntary.
Reports sent to the ASRS are held in strict confidence. More than 880,000 reports have been submitted to date and no reporter's identity has ever been breached by the ASRS. ASRS de-identifies reports before entering them into the incident database. All personal and organizational names are removed. Dates, times, and related information which could be used to construe an identity are either generalized or eliminated.

The Federal Aviation Administration (FAA) has set policies that provide ASRS reporters with further guarantees about privacy and additional incentives to make reports. For instance, the FAA has committed itself to not use ASRS information against reporters in enforcement actions. It has also chosen to waive fines and penalties, subject to certain limitations, for unintentional violations of federal aviation statutes and regulations which are reported to ASRS. The FAA's initiation, and continued support of the ASRS program and its willingness to waive penalties in qualifying cases is a measure of the value it places on the safety information gathered, and the products made possible, through incident reporting to the ASRS.[49]

State and Local Government Policies and Practices

The majority of State privacy laws are derived from Federal law and practice. A number of States have instituted statutes that go above and beyond Federal practices. In some cases, it has been determined by the legal system that Federal law preempts these statutes.

To address the private sector, individual States have adopted variations on the data privacy laws that are specific to the protection of customer data[50] and attempt to balance legal requirements

[47] http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf
[48] Robert Gellman, "Fair Information Practices: A Basic History" (unpublished essay, July 15, 2011), at: http://bobgellman.com/rg-docs/rg-FIPShistory.pdf
[49] http://asrs.arc.nasa.gov/overview/confidentiality.html
[50] For instance, California law 1386 or Massachusetts law MGL 93H.

with commercial interests. In recognizing differences in commercial interests, State law tends to address privacy under specific circumstances—online privacy, spam, spyware, opt-in/opt-out provisions, labeling requirements and use of radio frequency identification (RFID) technologies, financial information, insurance regulations, and others. In addition to these various statutes, many States have developed statutes that regulate apply restrictions to business; for example, regulations on merchants to prohibit requiring any personal information in associated with a credit card purchase or specific regulations regarding destruction of data. Some States require any electronic transmission to encrypt personal data. And further, many States have expanded the definition of PII to include, among other data, driver's license numbers.

Importantly, many States have enacted laws on security breaches that require consumer notification when there is a security breach involving PII. Not all States, however, define breach in the same manner. Some States require notification upon exposure; other States require notification only after an investigation has resulted in a finding that the exposure results in "likely harm".[51]

From a review of a number of sources, California stands out being the most active in its development of privacy legislation, with a number of States following California's example in one or more areas of privacy law. However, the laws are specific and most do not apply to the connected vehicle environment. Key lessons learned from review of State privacy laws for connected vehicle are:

- Most State practices are similar to or are pre-empted by Federal Privacy law.
- Many States have explicit policies for requiring on-line businesses to post a privacy policy where it is easily seen and accessed.
- Many States define a driver's license number as personal information but we did not find evidence of Vehicle Identification Numbers as part of the States' definition of PII. However, a number of States note that any details, when coupled with an individual's name or other identifying information are then considered PII.
- Over 19 states have laws that regulate disposal of business records that contain personal information.
- A frequent exemption is for encrypted information. Since encryption practices can range widely, some States have sought to provide guidance through a common definition that notes that encryption requires transformation of data into unreadable form.
- One State, Massachusetts, obliges companies to encrypt the personal information of MA residents on systems, laptops, and portable devices.[52]
- There are differing practices for breach of data notifications across the States. This applies when a business owner outsources data collection and maintenance to another party. States have varying definitions of who is responsible both for maintaining the security of the data and notification of consumers in the event of a breach. Some States

[51] Synopsized from Proskauer on Privacy: State Privacy Laws at: http://www.pli.edu/product_files/booksamples/11513_sample5.pdf

[52] 201 CMR 17.00 at www.mass.gov/Eoca/docs/idtheft/201CMR1700reg.pdf

place the full responsibility on the business owner and others look to the party who maintains the data. [53]

- Last, businesses that operate in multiple States must know and adhere to varying State laws.

Because the connected vehicle technologies and applications are expected to use encryption within a dedicated system, most of the State privacy laws are unlikely to apply. However, laws on breach of data notification and data disposal statutes may be applicable to regional operators and/or private businesses that collect, aggregate, refine, own a data environment, or provide data in the connected vehicle system.

Importantly, it is incumbent upon the agencies and businesses to develop their own privacy practices in line with Federal, State, and local laws.

Private Sector Policies and Practices

The Federal Trade Commission's (FTC) Bureau of Consumer Protection (BCP) enforces laws that protect consumers against unfair or deceptive practices. Specifically, the FTC oversees consumer privacy, children's privacy, credit reporting, data security, and financial institutions. The FTC also develops and provides guidance on advertising and marketing practices, credit and finance practices, and privacy and security practices.

For many companies, collecting sensitive consumer and employee information is an essential part of doing business. The FTC has the authority to ensure that if a company collects this type of information, that they follow their legal responsibility to take steps to properly secure or dispose of that data.

The FTC privacy and security guidelines for business are derived from the NIST 800-53 guidelines. The FTC recognizes that there is no "one-size-fits-all" approach[54]; however, businesses recognize that consumers' value privacy in addition to convenience, many private organizations have gone beyond what they are legally required to do and adopted additional practices that are similar to the FIPPs.

In the early 2000s, the FTC experienced pressure from advocacy groups and critics to change its approach to FIPPs, which had been criticized as weakly written and weakly enforced. In response, the Commission began to explore privacy issues and legislations through roundtables, studies, and by talking to stakeholders. From that research, the FTC developed a new set of FIPs principles. In drafting its new principles, the FTC responded to public outcry over well-publicized instances in which companies disclosed users' personal information.

The FTC released a preliminary Proposed Framework for Businesses and Policymakers in December 2010.[55] The new principles would apply to all commercial entities that use consumer

[53] State security breach notification laws http://privacylaw.proskauer.com/2010/04/articles/data-breaches/its-not-too-late-to-come-to-the-party-mississippi-joins-45-other-states-by-enacting-security-breach-notification-law

[54] See the FTC guidance, Protecting Personal Information: A Guide for Business at: http://business.ftc.gov/sites/default/files/pdf/bus69-Protecting-Personal-Information-guide-business_0.pdf

[55] "Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework for Businesses and Policymakers," Federal Trade Commission, last modified December 2010, http://www.ftc.gov/os/2010/12/101201privacyreport.pdf. 41-42.

data. The principles encourage companies to incorporate consumer privacy into all aspects of company culture and product development and service lifecycles. In following the principle of Policy by Design, companies incorporate substantive privacy protections into practices applying to data security, collection limits, retention, and accuracy. Similarly, the FTC requested that companies make their data practices more transparent by shortening, clarifying, and standardizing privacy notices. Companies must also obtain affirmative express consent to use consumer data in a method that differs from that originally expressed.

On the consumer side, the FTC wants consumers to be able to access data about themselves in order to be able to correct it and/or limit the information. Additionally, FTC is encouraging companies to simplify the choice mechanism that they present to consumers; that simplification means only notifying them of data usage policies that fall under a list of "commonly accepted practices."

Significantly, the Proposed Framework for Businesses and Policymakers Model outlines no punishments or enforcement methods for companies that do not follow these rules. Predictably, privacy advocates have criticized the FTC's new privacy principles for being too weak. On the other hand, the FTC has earned praise for some aspects of its FIPPs, such as the privacy principles lasting for the entire lifecycle of a product or service, and the Do Not Track (DNT) mechanism to protect consumers from online behavioral advertising.

Academic Research Policies and Practices

Academic institutions that conduct research have unique challenges with regard to privacy—

- Highly sensitive personal information is typically the foundation of research and can include health information, behavioral and psychological information, economic data, and others. Exposure can lead to job loss, insurance loss, social stigmatization, or damage to economic or social status.
- Large databases are developed to look longitudinally across time at individuals as well as to compare and contrast segments of individual data against others.

A common practice is to mask data so that an individual is not easily identified by a record. However, characteristics associated with an individual are critical elements in research. The more characteristics that are recorded, the more a researcher is able to control for variables that might influence results. However, each additional characteristic adds to the ease of associating a record with an individual.

A number of policies tend to govern academic research; many of them are specific to the type of research being conducted. Most of these regulations are targeted at researchers that deal with human subjects (for instance, National Institutes of Health's Certificates of Confidentiality, the Food and Drug Administration's Informed Consent requirements in 21 CFR, or what is known as the "Common Rule", CFR Title 45, Part 46, which defines human subjects and provisions for privacy).

Two lessons learned are identified by reviewing academic research practices regarding privacy:

- If the research has Federal funding associated with it, Federal regulations apply and the institution typically follows FIPPs to provide appropriate notifications regarding its collection and use of data. A sampling of academic examples noted that most appear to ensure that as long as an individual is associated with research, the individual is kept informed about data use. Also, most Federal regulation allows the institution to decide what level of security is appropriate.

- Most of the research practices reviewed tend to employ de-identified data sets. In the past, this was done manually through coding a record and storing the data about the individual (name, address, etc.) in a separate location with the code. As the next section will describe, such practices are becoming increasingly automated.

Privacy Technologies: State-of-the-Practice Examples

PETs provide a set of tools for "de-identification" of data to protect PII. PET allows online users to protect the privacy of their PII through the use of computer applications and mechanisms implemented in conjunction with online services or applications. PET functions as a system of measures that minimize personal data, preventing unnecessary or unwanted processing of personal data, without the loss of the functionality of the information system.[56] Goals of PET are to provide users with the ability to control their personal data, minimize the extent of personal data maintained by organizations, protect their identities, effectuate informed consent, track the transfer of their data, and exercise their legal rights of data inspection, correction, and deletion.

Examples of existing privacy enhancing technologies include:

- Communication "anonymizers" – hide real online identity of individuals and replace it with a non-traceable identity;
- Shared fake online accounts with false identifying information that users then use to publish their user-ID and password on the Internet.
- Access to personal data – service provider infrastructure allows users to inspect, correct, or delete their personal data stored by a service provider.

Electronic Identity Management (IdM) – renders electronic communications anonymous. IdM applies to user credentials, the means by which users might log on to an online system and – since the advent of phishing attacks – the management of individual identities by service providers. The term "National Identity Management" has been used in relation to online government services.

Removal of PII from data sets prior to release presents a challenge to entities administering open data environments. Manual data scrubbing is highly labor intensive and, in any case, is unlikely to be able to keep pace with the volume of data in active environments such as those envisioned for DCM / DMA. Fortunately, automated PII redaction software solutions exist; and public agencies are beginning to use them to provide identity theft protection for public records.[57]

PETs can be grouped into four categories[58]:

[56] Van Blarkom, Borking and Olk 2003.

[57] See, for example: http://www.csisoft.com/company/displaynews.php?item=46\

[58] This typology and decision process is summarized from Chapters 4 and 7 of Koorn, R. et al. (KPMG), Privacy-Enhancing Technologies: White Paper for Decision-Makers. 2004. Produced by KPMG Information Risk Management for the Dutch Ministry of the Interior and Kingdom Relations. http://www.dutchdpa.nl/downloads_overig/PET_whitebook.pdf

- **General PETs**: Commonly practiced techniques for maintaining and managing personal data, including encryption, user access control, and data masking;
- **Separation PETs**: Separating the personally identifiable data from other data and linking the two (or more) databases through an identity protector; also known as developing "pseudonymity" or employing "communication anonymizers";
- **Privacy Management Systems**: Privacy management software tightly integrated with databases containing personally identifiable data; and
- **Anonymization**: Destruction or avoidance of personally identifiable data, including automated data scrubbing or cleansing; also known as "de-identification"

A KPMG report on Information Risk Management offers a graphic that summarizes the different levels of effectiveness among the different categories:

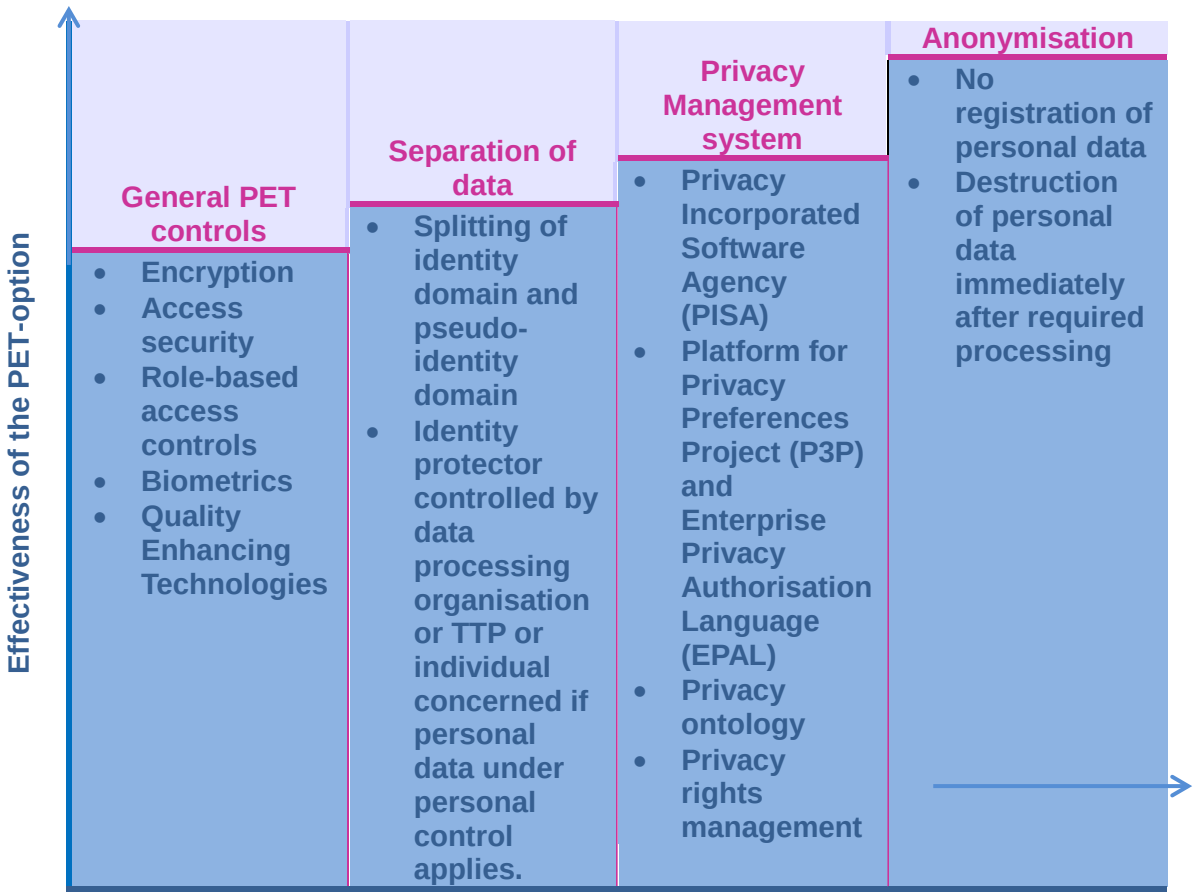

Figure 3.1. PET staircase: Effectiveness of the different PET options[59]

In overarching terms, general PET controls are less effective than a separation approach, which is in turn less effective than a privacy management system. Anonymization is the most effective. The selection of a PET approach is guided by the extent to which the application in question requires PII to function. Where PII must be used and maintained to allow full functionality, a less effective PET approach must be used. Where PII is not required, anonymization is a viable option.

Potential future PET that currently are the subject of research or development include: User identity management functionality includes user information self-service; password resetting; management of lost passwords, workflow, provisioning and de-provisioning of identities from resources.

- Wallets of multiple virtual identifies – allow the efficient and easy creation, management, and use of virtual identifies;
- Anonymous credentials – reveal only as much information as the holder of the credential is willing to disclose;
- Negotiation and enforcement of data handling conditions – limit the type of data and the conditions that apply to handling personal data such as whether it may be sent to third parties and the conditions under which the data will be deleted.

The National Renewable Energy Laboratory and the FAA's Confidential Close Call Reporting System are prominent examples of the use of data privacy technologies and protocols.

Renewable Energy Laboratory (NREL)

NREL has developed the Transportation Secure Data Center (TSDC) to provide an option to transportation planning agencies[60] to collect and analyze trends based on data that has identify information associated with it, and to house detailed transportation data from a variety of sources for continued and expanded research. The TSDC securely archives data sets and supports research efforts to build accurate and reliable real-world models. The TSDC will make past and future data available to broader groups, such as metropolitan planning organizations (MPOs) or municipalities, as well as to new data users at automobile manufacturers, national laboratories, the DHS and the U.S. Department of Energy (DOE).

TSDC scrubs raw data for use by a wider group, removing any confidential information. The resulting cleansed data, which includes high-level summary statistics and second-by-second speed profiles (with latitude/longitude information removed), is freely available for download. The TSDC's two levels of access make composite data available with simple on-line registration, and

[59] From page 33 of Koorn, R. et al. (KPMG), Privacy-Enhancing Technologies: White Paper for Decision-Makers. 2004. Produced by KPMG Information Risk Management for the Dutch Ministry of the Interior and Kingdom Relations. http://www.dutchdpa.nl/downloads_overig/PET_whitebook.pdf

[60] Agencies at the Federal, State, and municipal levels.

allow researchers to use detailed spatial data after completing a more rigorous clearance process.[61]

The Confidential Close-Call Reporting System

The Confidential Close-Call Reporting System (C^3RS) is a Federal Railroad Administration (FRA)-funded, voluntary, confidential demonstration project designed to improve safety practices within the railroad industry. It is designed to gather information about potentially unsafe conditions, or close call events, that pose the risk of more serious consequences. As a demonstration project, the C^3RS is exploring how to adapt a confidential reporting system to the needs of the U.S. railroad industry and to evaluate its effectiveness in improving safety programs for railroad carriers and their employees to report close calls without receiving disciplinary action. Within C^3RS, confidentiality is ensured in the removal or de-identification of personal and carrier information from a close call report. In other words, the identity of the reporting employee or anyone mentioned in the report cannot be determined. This creates an environment in which more information is likely to be disclosed.[62]

Lessons Learned

Privacy and open data are not antithetical. Protecting privacy in an open data environment does, however, require careful attention. Entities supplying open data and/or administering data environments must thoughtfully develop their own principles regarding the need for identifying data and the uses of such data or "scrub" data (using automated PII redaction systems if possible) prior to release, to remove PII. Additionally, agencies must ensure that strict policies and procedures that protect the data and guide who can access it and for what purposes. The above described models present a range of policy and procedural options to guide the DCM / DMA development effort—the VII Privacy Framework, FIPPs, secure and protected data environments, and secure reporting systems with limitations on data use. In addition, Chapter 4 presented options for User Access Policies and Controls which illustrated how access to open data environments can be controlled, and, when warranted, limited to further enhance privacy.

One common approach to protecting the privacy of personal information online used in private sector commercial applications is the "opt out," in which users can use tools or follow procedures to limit and control the tracking of their online activities. A recent Carnegie Mellon University's study, "Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising," found, however, that online opt-out tools were difficult for users to understand and configure.[63] Users approve of "Do Not Track" features provided by major web browsers, but tend to harbor doubts concerning their effectiveness, due to concerns that advertisers will not honor privacy restrictions. Primary lessons relevant to DMA and DCM are that user-initiated privacy controls must be clearly comprehensible and user-friendly. A preferable approach may be to provide "opt in" options. Also, it is crucial that organizations communicate privacy protections clearly to users and earn user trust by adopting data privacy as a fundamental cultural value.

[61] http://www.nrel.gov/vehiclesandfuels/secure_transportation_data.html
[62] http://www.closecallsrail.org/Default.aspx
[63] *Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising,* CyLab Carnegie Mellon University, October 31, 2011.

Additionally, as the DCM and DMA programs start their efforts to prototype systems, technologies, and applications, the use of PETs should be considered as a means of demonstrating effectiveness but also to identify costs and institutional challenges. The proper approach to managing privacy through PETs, however, is best identified only after conducting an *application-specific or technology-specific analysis* to identify data needs and determine if PETs limit the information to much and thus limit functionality or performance.

Conclusion

Protection of privacy in open data environments is challenging, as is the case with data security. However, well-established policies, guidance, and tools exist to address this critical need.

Next Steps

- Identify the key privacy risks inherent in the open data and open data environments anticipated for the DCM / DMA technologies and applications. Structure this inquiry phase according to the privacy risk categories outlined in NIST 800-53 “Appendix J” (See paper on Privacy Analysis).
- Using the tools provided in Appendix J, develop approaches to mitigate privacy risks.
- Analyze where PETs most effectively apply to the DCM / DMA technologies and applications.
- Pilot test the privacy approaches to understand the level of burden they impose on participating entities. Adjust where appropriate.
- Develop guidance documents outlining the privacy protection approaches and their application

Chapter 4 User Access Polices and Controls

Recommended Policy Option

Careful consideration of user access issues and risks at the program, portal, and project level is recommended, along with development of appropriate user access policies that match the level of restrictiveness to the sensitivity of various data, source code, or applications. Access classes and policies must be developed in conjunction with the Open Source community to foster user collaboration and acceptance.

At the development portal level, Forge.mil provides a good model for developing high-security user access policies, whereas ITdasboard.gov represents a well-designed approach to access in the case of low-security, public access data.

Definition

User Access policies specify who can access, use, and contribute to a website. User Access is often defined in legal documents called "User Access Agreements" (also sometimes referred to as "User Agreements," "Terms of Service" or "Terms of Use"). These documents set the terms and conditions under which users are permitted to access and use content on a website or portal. User Access Agreements also often contain a user code of conduct and statement of warranties (e.g. user warrants that they will only post content that they have rights to post, user warrants that they are solely responsible for their conduct and content). The specific content of a website's user access agreement will vary based on the nature of the specific services offered.[64]

High-level policy makers, as well as managers at the program, development portal, and project levels, need to think carefully about user access polices and controls. This is because these different levels often face different types and severities of access-related risk, and, consequently, require different, although related, approaches to dealing with access control. Important access-related decisions include the following:[65]

- Development of policies governing differential access permissions to users based on their roles as data providers, application developers, project managers, etc.

[64] User Agreements are also ubiquitous on software – even free software – that is downloaded directly from websites, or delivered via a DVD.

[65] A recent whitepaper, "Policy and Institutional Issues Analysis for the Dynamic Mobility Applications (DMA) Open Source Application Development Portal (OSADP)", provides a comprehensive discussion about access controls and policies.

- Creation of access controls that are flexible, to provide more restrictive control over access to sensitive and/or proprietary data, source code and applications, and easier (less cumbersome) access to other data.

User Access Policies: State-of-the-Practice Examples

Forge.mil

Forge.mil is an online portal for Agile open source application development in support of the Department of Defense (DoD). Forge.mil restricts user access to U.S. military personnel, DoD civilian employees and DoD contractors. All users require a DoD Common Access Card or a PKI certificate issued by a DoD-approved External Certificate Authority with a government sponsor to access Forge.mil

Projects housed within the SoftwareForge module of Forge.mil are open to all users, and users are encouraged to view and contribute to the projects housed there. Additionally, Forge.mil offers the ProjectForge module which features the same tools as SoftwareForge but with user access controlled on a project-specific level by the project manager.[66]

ITDashboard.gov

The IT Dashboard provides the public with an easy way to access data on federal agency IT spending and procurement. The IT Dashboard provides a good example of multiple user access levels. The general public can access all of the data on the IT Dashboard without the need to register. Data are also made available for download in both CSV and XML formats. However, only agency-authorized users with valid MAX credentials have permission to submit data for inclusion in the IT Dashboard. Data submissions are executed through the "My Investments" page, (which only authorized users) can view once they log in.[67]

User Access Controls Definition

Access controls are the mechanisms by which user access policies are enforced. They grant or revoke rights to access data or perform other actions. Access controls include the following:

- *File permissions*, such as create, read, edit, or delete on a file server.
- *Program permissions*, such as the right to execute a program on an application server.
- *Data rights*, such as the right to retrieve or update information in a database.[68]

[66] http://www.forge.mil/Faqs.html,http://www.forge.mil/UserAgreement.html
[67] http://www.itdashboard.gov/faq-agencies
[68] http://hitachi-id.com/concepts/access_control.html

Open Data Access Controls: State-of-the-Practice Examples

IT Dashboard

Access controls incorporated by the Office of Management and Budget (OMB) in the IT Dashboard (http://www.itdashboard.gov) enable Federal agencies, industry, the general public, and other stakeholders to view open data providing details of Federal IT investments. In response to high demand from Federal agencies and the software development community, the source code of the IT dashboard was made available to the public. This first open-source release represents a starting point, enabling communities of interest to adapt and mature their own versions of the Dashboard to their own unique needs.

The Dashboard provides information on the effectiveness of government IT programs and to support decisions regarding the investment and management of resources. The Dashboard is now being used by the Administration and Congress to make budget and policy decisions.[69]

TRANSIMS

The Transportation Analysis and Simulation System (TRANSIMS)[70], illustrates a selectively open-access environment. TRANSIMS is an integrated set of tools that were identified to conduct regional transportation system analyses. In this system, an open source community has been developed into an independent and self-governing collaboration of TRANSIM users, researchers and developers. A web-based infrastructure provides access to TRANSIMS core assets (software, data sets and documentation) and supports community interaction. Members collaborate by sharing code, enhanced documentation and submitting proven data sets back into a public clearinghouse. Access controls adopted by the allowed members to conduct regional transportation system analyses.

Open source portals typically manage user access to protect content and track usage. User access controls enable the system owner to acquire information about anyone/anything wanting to use the system, decide who may or may not enter, and limit the range of portal elements the user can see. For the OSADP, this means that user access controls can adjust access to each application as needed.

Data.gov

Relative to the concept of open portal, Data.gov is a key initiative of Open Government, and the portal for public access to data from 172 Federal agencies and sub- agencies via cloud-based, open data services platform. It offers raw data, interactive data sets, applications, and a developer community. Data sets that contain sensitive information (personally identifiable information, national security), that are limited by technology (e.g., not machine readable), or that

[69] http://www.itdashboard.gov/
[70] http://www.transims-opensource.net

do not belong to the Executive Branch of the Federal Government are not available on Data.gov.[71]

Lessons Learned

User access restrictions are compatible with open data environments. Indeed, user access to open data environments is often controlled to preserve privacy and provide security. This can include limiting access to certain persons (e.g., those with proper clearance), and imposing specific terms and conditions on access. Two factors are important for achieving this balance. First, the open source community (including the public agencies participating) must develop user access policies collaboratively, so that they are understood and accepted by all participants. These policies can then be used to develop user access controls. Second, certain data sets (e.g., those that cannot be scrubbed of PII) must be excluded from the data environment as necessary.

Conclusion

User access controls must reflect the variable nature of users, source code, applications, and data at the program, portal, and project levels. Consequently, development of a comprehensive user access policy framework covering all levels is a critical first step in addressing access risks. Such a framework will guide development of appropriate user access controls that are restrictive where necessary, and more permissive where appropriate.

Specific user access issues directly related to the DCM and DMA programs will be addressed in separate reports that are forthcoming.

Next Steps

- Catalogue risks related to user access to data, source code, portals, and applications.
- Develop a user access policy framework to address the various risks identified.
- Develop draft access controls in accordance with the policy framework.
- Vet the draft access controls with users at all levels; revise as appropriate
- Finalize access controls and develop implementation timeline.

[71] http://www.data.gov

Chapter 5 Data Quality Assurance

Recommended Policy Option

DCM / DMA data system and data set owners/operators should consider the Bureau of Transportation Statistics guidelines as the foundation for the development of data quality assurance protocols. In addition, the Center for AIDS Research Network of Integrated Clinical Systems, and the National Data Buoy Center stand out as having particularly rigorous quality processes in place that make creative use of automated techniques for real-time data quality verification. It is likely that the DCM / DMA programs will need to go well beyond even this level of data quality assurance (given the potential for real-time, safety-critical nature of the data), to include formalized protocols for data review, error documentation, and error correction.

Use the DCM Program's Evaluation Framework to quantify the benefits of the data sets and data environments developed through the DCM / DMA program.

Definition

The data will support development of effective applications only if the data are of consistently high quality. The DMA application developers and other developers and system users must have confidence that the data exchange has in place policies and procedures to ensure data quality.

Data quality assurance is the process of profiling data to discover inconsistencies and other anomalies, and performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve data quality. These activities can be undertaken as part of Data warehousing or as part of the Database administration of an existing piece of applications software.[72]

Data Quality Assurance: State-of-the-Practice Examples

Bureau of Transportation Statistics (BTS)

In its *Guide to Good Statistical Practices in the Transportation Field*[73] BTS recommends a number of elements that, taken together, constitute a comprehensive, programmatic approach to ensuring data quality. These elements include:

- **Periodic data quality assessments,** undertaken by data system owners. These assessments should include consultation from data users, including secondary data users,

[72] http://en.wikipedia.org/wiki/Data_quality_assurance
[73] http://www.bts.gov/publications/guide_to_good_statistical_practice_in_the_transportation_field/html/chapter_06.html

to suggest areas to be assessed, and to provide feedback on the usefulness of the data products.

- **Periodic data quality evaluation studies.** Data quality studies use various experimental techniques to verify data validity and accuracy, to check for coverage bias and measurement error, to establish error tolerance, to assess user satisfaction, and to evaluate other aspects of data quality that cannot be determined by examining the end-product data.
- **Quality Control Processes.** These processes are applied to each data collection activity to double-check accuracy and catch errors. Activities that require particular attention to quality control include data entry, data coding, and data editing.
- **Error Correction.** Plans should be in place to address errors that are discovered in data sets after they have been released. These plans should include procedures for replacing faulty data, and for documenting the errors so that data users are aware of them.

Data.gov and ITdashboard.gov

One approach to data quality assurance, for public-domain data, is to depend upon the providing entities to verify data quality. Data.gov, for example, does not assume responsibility for ensuring the quality of the publically available data available through it. Instead, it requires each participating agency to confirm that the data it supplies meet the agency's Information Quality Guidelines. Similarly, itdashboard.gov, which "enables federal agencies, industry, the general public and other stakeholders to view details of federal information technology investments," does not independently verify the quality of data accessed through it.

The "hands off" approach to data quality that data.gov and itsdashboard.gov have adopted will not be appropriate for the vast majority of the data related to the DCM / DMA programs. With the possible exception of purely descriptive statistics, all data will need to be rigorously checked to ensure their accuracy and validity, and corrected as necessary.

Transportation Secure Data Center

Geo-spatial data collected from individual vehicles are valuable for transportation planning and for the development of a variety of transportation-related applications, including DMA. Launched in 2010, the TSDC at the NREL provides a repository for these types of data, and addresses the attendant privacy concerns by scrubbing the data to remove PII before releasing them for public use. The TSDC also screens raw data for missing values before releasing them.
The preceding approaches present little risk when used with data that are static once collected (even if periodically updated), and tend to be historical. Indeed, as the DCM program notes:

> Most current data environments are archival in nature. Data captured and managed in the current data environments tend to be collected over time, assessed for quality and potentially aggregated at some intermediate point, and

> then at a later date (days, months or even years later) made available to researchers or other interested users.[74]

To support the envisioned connected vehicle applications, data environments will need to evolve to contain dynamic, multi-modal, real-time data, actively captured from myriad sources, including mobile sources. For such data, quality assurance must be performed "on the fly." Automated processing routines that evaluate incoming data can be used for this, as the next two examples illustrate.

National Data Buoy Center

The National Oceanographic and Atmospheric Administration (NOAA) National Data Buoy Center (NDBC) employs an automated Data Quality Control Checks and Procedures system (NDBC QC Program). The NDBC QC Program ensures that incoming sensor data are within NDBC total system accuracy. The system compares incoming sensor data (e.g., atmospheric pressure, wind direction, water temperature) from a given sensor against duplicate sensors on the same buoy, sensors on an adjacent buoy, or established standard values.

Center for AIDS Research Network of Integrated Clinical Systems

Access to real-time data is beneficial for clinical research about a variety of diseases; in these applications, ensuring data quality is essential. To meet this need, the medical research community is introducing automatic data quality assurance systems. For example, the Center for AIDS Research (CFAR) Network of Integrated Clinical Systems (CNICS) collects clinical data from the large and diverse population of HIV-infected persons.[75] CNICS captures a broad range of information associated with the rapidly changing course of HIV disease management through collection of "real-time" data at the point-of-care. These data are rapidly available to researchers through a "peer reviewed open access platform."

To ensure data quality, CNICS uses two modes of quality verification: synchronous and asynchronous. Uploaded data are checked for adherence to the existing CNICS metadata and coding standards. Synchronous validation occurs prior to the loading of data into the repository and verifies that all data elements are reported using a valid format and value. Asynchronous validation involves applications used after data are loaded into the CNICS repository to monitor data quality centrally. For example, "patient data are evaluated for potentially invalid date ranges; deceased patients should not have clinical events occurring after their date of death, start and stop dates for courses of therapy with antiretroviral medications and episodes of clinical conditions must produce positive durations."[76]

The DCM and DMA programs will have data quality assurance requirements; the NDBC or CNICS may provide useful models for how to meet those requirements, at least in terms of certain features. For example, in the same way that NDBC uses data from adjacent buoys to

[74] Data Capture and Management Program: Transforming the Federal Role, at: http://www.its.dot.gov/research_docs/pdf/25Data%20Capture%20Federal%20Role.pdf
[75] http://www.uab.edu/cnics/
[76] Ibid.

verify quality, in DCM / DMA applications, data from adjacent vehicles could be compared to identify vehicles that are out of synch and possibly transmitting erroneous information. Incoming data could be compared with established "normal" data ranges, and asynchronous validation could check information in databases for errors.

Lessons Learned

Open data and open data environments – particularly those dealing with dynamic, real-time data from numerous sources – present unique data quality assurance challenges. Data quality assurance protocols that function effectively for static, historical, and/or "closed" data sets will not suffice. Automated systems for checking data quality in real-time (i.e., synchronously) as well as and prior to release (i.e., asynchronously) will be necessary. In addition, open data environments will in many cases need to rely on the participating entities to independently verify the quality of the data they collect before submitting them (or before linking their databases to a virtual data environment). Consequently, all entities providing data to DCM /DMA open data environments will need to agree in advance to a set of data quality standards.

A recent whitepaper prepared for the US DOT ITS Joint Program Office presents an evaluation framework and performance measures to assess datasets and data environments developed through the DCM program, and to quantify their benefits.[77] The whitepaper includes valuable guidance on a number of key technical and institutional issues, including compatibility of data formats (e.g., spot speeds and link speeds); the use of open data standards; verification of data quality and validity; and the importance of archiving data for the purposes of reporting and evaluation.

Conclusion

In the examples above, all the organizations have well-developed data quality assurance processes in place, although they vary widely in their approaches and level of rigor. However, none of them are at the level that will almost certainly be required for the DCM / DMA programs.

Data quality assurance cannot be done on an ad hoc basis; it must be planned in advance, and must entail a programmatic approach that includes processes for ongoing data quality control, data quality assessment, and error correction. Data system operators and data "owners" need to take primary responsibility for data quality assurance, and should consult with data users for insights into data quality issues.

[77] United States Department of Transportation, Research and Innovative Technologies Administration, "Real-Time Data Capture and Management Evaluation and Performance Measures -Evaluation Framework..." September 6, 2011, Report No. FHWA-JPO-11-136, at: https://one.dot.gov/rita/ProgOffs/ITS/docworkspace/Mobility%20Initiative%20DL/DCM%20Program/Track%2002A%20Research%20and%20Development/Data%20Business%20Plan/Data%20Business%20Plan%20APPROVED%20EDITION-Evaluation%20Framework%20Report%20Draft%20Final1Sept2011.pdf

Data quality assurance protocols will need to be established very early in the DCM / DMA programs, before data collection efforts get underway. Further, the protocols will need to evolve as new data sources and/or data collection methods are introduced.

Next Steps

- Very early in the DCM / DMA programs, identify all current data sources and data collection methods at the program, portal and project levels. Catalogue data quality risks for each data source.
- Use the techniques and process developed in the Evaluation Framework whitepaper to quantify the benefits of the data sets developed through the program.
- Using the BTS guidelines as a foundation, develop draft data quality assurance protocols.
- Vet the draft protocols with data system/data set owners to make sure they are sufficiently rigorous yet not unduly burdensome. This vetting process would entail conducting "test runs" of each protocol and then collecting comments and suggestions from the participants.
- Develop a timeline for implementation of data quality assurance protocols that ensures they are in place prior to data collection efforts being launched.

Chapter 6 Intellectual Property (IP)

Recommended Policy Option

A thorough, well-documented and clearly communicated IP policy framework will be necessary to provide all participants in the DCM / DMA application development efforts with a clear understanding of the rules of the game with respect to licensing, patents, and other aspects of intellectual property protection. This will be challenging, given the large number of participants and applications envisioned. Nevertheless, a rigorous and defensible IP Policy Framework must be developed very early in the program.

IP issues should be dealt with at the application levels of the DMA program to allow third-party developers to make downstream enhancement.

Forge.mil provides the recommended model for open source application development to support major public-sector initiatives.

Definition

Licensing of IP and, in particular, the application of open source licensing models is among the most important issues the DCM and DMA programs will face. Open source applications raise unique IP liability concerns. As the American Bar Association (ABA) points out:

> The typical open source project is a grass-roots effort that contains contributions from many people. This method of development can be worrisome from an intellectual property standpoint because it creates multiple opportunities for contributors to introduce infringing code and makes it almost impossible to audit the entire code base.[78]

To meet Mobility Program goals, the Mobility Program must acquire and preserve the right to provide developed applications under open source terms. Without these terms, the Mobility Program will infringe upon the intellectual property rights of the software developer, as will any downstream developers or users employing software acquired through the Mobility Program. A licensing model is needed that protects the intellectual property of all parties, while;

- Giving the government the required flexibility to use, maintain, and modify software in a collaborative environment, between multiple organizations. This means the government must have unlimited rights to all software (including source code);

[78] American Bar Association, at: http://www.americanbar.org/groups/intellectual_property_law/resources/an_overview_of_open_source_software_licenses.html

- Allowing open source third-party developers to make downstream enhancements to the applications; and
- Remaining consistent with business models (so that the developers will remain engaged, and produce useful applications).

United States intellectual property law views computer applications as creative works, and automatically assigns the ownership of the intellectual property to the software's creator in the form of copyright. Reproduction, distribution, modification, public demonstration and public display of software that is "substantially similar" to the original software are illegal without the creator's permission. A license is the formal grant of rights by the creator to engage in conduct that otherwise would be a violation of the licensor's intellectual property rights.[79]

The extent to which databases and data sets may be covered by copyright varies with the specific product. "Facts" are not copyrightable; the data must be the product of some additional treatment or manipulation to be arguably protected.[80] Software is covered by copyright[81] and may also be patented.[82] In open software development, copyrights are typically licensed via an open source license. Databases are also copyrightable under the concept of compilation copyright.[83]

Open Source Licensing Practices

Free and open software (FOSS) licensing emerged in direct response to the restrictions on access to source code imposed by conventional licensing, and on the user's consequent inability to fix bugs, tailor or improve the software to meet individual needs, or pass these improvements on to other users without incurring additional costs. There are over a thousand FOSS licenses in use at this time. FOSS licensing is recognized in US and international law as an alternative to conventional licensing.

A license defines the rights and obligations that a licensor grants to a licensee. Open Source licenses grant licensees the right to use, copy, modify and redistribute source code (or content). These licenses may also impose obligations (e.g., modifications to the code that are distributed must be made available in source code form; an author attribution must be placed in a program/ documentation using that Open Source, etc.).

When an author contributes code to an Open Source project (e.g., Apache.org) they do so under an explicit license (e.g., the Apache Contributor License Agreement) or an implicit license (e.g., the Open Source license under which the project is already licensing code). A second option is to require each contributor to submit a Contributor License Agreement (CLA) which specifically grants the project permission to use the submitted content.[84] Some Open Source projects do not

[79] http://www.utahbar.org/sites/midyear/html/introduction_to_software_licen.html
[80] **FEIST** Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991)
[81] 17 USC 101.
[82] 35 USC Part II.
[83] Bitlaw, *Database Legal Protection*. http://www.bitlaw.com/copyright/database.html
[84] Fogel, Karl. Producing Open Source Software: How to Run a Successful Free Software Project. Ch. 9. http://producingoss.com/

take contributed code under a license, but actually require (joint) assignment of the author's copyright in order to accept code contributions into the project (e.g., OpenOffice.org and its Joint Copyright Assignment agreement).[85]

The most common types of licenses used in open software development are: the GNU General Public License (GPL), the Massachusetts Institute of Technology/X Window License (MIT/X), and the Berkley Software Distribution License (BSD). All three of these common license types allow for free copying and distribution; however there are some important differences relating to incorporation of open software into proprietary software. [86]

- GPL requires that the software be made freely available at no cost and that all future works containing some aspect of the software be licensed under GPL. This requirement means that GPL-licensed software cannot be sold or incorporated into proprietary software. This is likely the most widely-used software license.
- MIT/X enables free copying and distribution for any purpose. MIT/X – licensed software may be incorporated into for-profit proprietary software and is compatible with GPL-licensed software.
- The revised BSD license is essentially the same as the MIT/X license, but for historical reasons, it contains a clause requiring prior written permission to use the names of the project contributors to endorse or promote derivative products that use the BSD-licensed software.

The key challenge for the mobility program is that current federal procurement practices reflect conventional copyright law, not open source. Consequently, selecting appropriate licenses and contributor agreements to meet DCM / DMA program goals is complex, yet essential.

In the application development process for the DCM / DMA programs, the need to manage intellectual property risks will come into play at three points: when the development of the application is being arranged; when the application has been accepted and is being offered to users; and when users are contributing enhancements to the application back into the repository.[87]

The key to successfully defining the right intellectual property terms is the distinction between the licenses that the DMA Program will offer to users (the "outbound" licenses) and the licenses it will be receiving from developers and contributors (the "inbound" licenses or contributor agreements) who will provide "seed" code, develop the applications, or contribute enhancements.

At the most basic level, developing an effective approach to intellectual property issues in the context of the DCM / DMA programs is a matter of governance. A rigorous and defensible IP policy framework must be developed, vetted, accepted and implemented early. In this way, all parties (applications developers, source code providers and end users) will know the rules of the game, and understand how their interests are being protected (and also those instances where

[85] http://en.wikipedia.org/wiki/Open-source_software#Licensing

[86] Fogel, Karl. Producing Open Source Software: How to Run a Successful Free Software Project. Ch. 9. http://producingoss.com/

[87] For a comprehensive discussion of IP issues in the context of the DCM / DMS programs, see "Policy and Institutional Issues Analysis for the Dynamic Mobility Applications (DMA) Open Source Application Development Portal (OSADP)."

they are not protected, and why). A thorough, well-documented and clearly communicated IP policy framework will make all parties comfortable about participating in the application development process.

Intellectual Property: State-of-the-Practice Examples

Forge.mil

Forge.mil is an online portal for open source application development in support of the Department of Defense (through its' SoftwareForge and ProjectForge modules). Forge.mil allows a large set of approved users to contribute to development projects, thus each project has potential exposure to IP violations. The Forge.mil user agreement addresses IP with the following conditions:

- Users must own or have sufficient rights to post and distribute their content.
- No user may post information that is classified.
- No user may post information that violates the privacy rights, publicity rights, copyrights, or other IP of any person.
- Users must ensure that they have complied with any third-party licenses and they agree to pay any royalties, fees or other monies owed to any person as a result of posting content.
- If the user's employer has rights to the content or other IP created by the user, the user must receive permission to make the content available or secure a waiver of all rights to the content from the employer.
- Software shared on SoftwareForge, and developed by government vendors may only be posted if the U.S. Government has an Unlimited Rights or Government Purpose Rights license.[88] Software developed using ProjectForge is entirely governed by the project manager. It is up to the manager to ensure that all IP requirements are met and to later determine if the content can or should be shared with the wider Forge.mil community and posted on SoftwareForge.

Because Forge.mil is a portal for the development of numerous applications, specific licenses are determined at the project-level. Forge.mil supports both Open Source Initiate (OSI)-approved licenses (e.g. GPL, MIT/X. revised BSD) and DoD Community Source Software licenses (DoD-specific licenses which restricts rights to view, use, modify, and distribute to the DoD, or DoD contractors).[89]

[88] For a discussion of these types of licenses, see www.wifcon.com/anal/GPR_TD.doc
[89] http://forge.mil

CONNECT[90]

As described previously in this paper, CONNECT is a software product that enables health IT systems to interface with health information exchanges. Funded collaboratively by a group of Federal agencies for the Federal Health Architecture, the project has now transitioned to an open source project.

CONNECT is licensed under a BSD license known as the "Three-Clause BSD" license which is certified by the Open Source Initiative (OSI) and is GPL-compatible. This license does not restrict the incorporation of CONNECT into proprietary software, and the intent is that CONNECT will generate commercial activity that builds upon the basic CONNECT solution.

CONNECT was developed under contract for the Federal Health Architecture by Harris Corporation and its partners. Subsequent revisions and additions have been executed under open source license.[91]

Open Data Licensing Practices

Collections of data in a database or compilation are generally protected under U.S. copyright law.[92] Although the contents of the database or compilation may not be copy writable (e.g. public domain, or factual information), if the structure of the database or compilation is sufficiently novel or creative it may be subject to copyright (e.g. a list of "best practices" identifies the factual information in the list as of high value).[93] For the purposes of ensuring unambiguous, free and open access to compiled material, organizations are beginning to adopt the use of Open Data Licenses.[94]

All data created by the U.S. Federal Government is in the public domain. Thus, it is not currently a standard practice to license federal data. However, this is not necessarily the case for State and Local data, which may be subject to copyright restrictions unless copyright is specifically licensed or dedicated to the public domain.[95]

The Open Knowledge Foundation refers to open data as "open knowledge" and offers the following definition:

> *Open knowledge is any material – whether content, data or general information – which anyone is free to use, re-use and redistribute without restriction*.[96]

Through its Open Data Commons project, The Open Knowledge Foundation also established the first open data license in 2008: the Public Domain Dedication License (PDDL).[97] The purpose of licensing a database or data compilation with an open data license is to ensure the long-term

[90] http://www.connectopensource.org/about/what-is-connect
[91] http://www.connectopensource.org/about/what-is-connect
[92] 17 USC 101
[93] Bitlaw, Database Legal Protection. http://www.bitlaw.com/copyright/database.html
[94] Open Data Licensing, The Open Knowledge Foundation. http://wiki.okfn.org/OpenDataLicensing
[95] Stanford University Libraries. *Copyright and Fair Use Overview.* http://fairuse.stanford.edu/Copyright_and_Fair_Use_Overview/chapter8/8-a.html.
[96] Ibid.
[97] Open Data Commons, *About.* http://opendatacommons.org/about/

open availability of the data. Open data licensing provides a clear signal to potential users as to their ability to freely use, break-up, redistribute, recombine, or reuse the licensed data.[98]

The governance of open data is a growing issue of concern within the Federal government, particularly in the context of the Open Government Initiative, which pushes government agencies to make high-value data more freely available to the public online.[99] There are three basic types of open data licenses as defined by Open Data Commons[100]:

- **Public Domain Dedication:** Puts the data in the public domain, allowing anyone to use or distribute for any purpose without restriction, including the use of the data to create proprietary products. Dedicating a work to the public domain means the original owner actual transfers ownership to "the public." Thus, public domain dedication is not actually a license at all.
- **Share-Alike (plus attribution):** As in the GNU GPL open source software license, a share-alike license requires users to retain the original license and to attribute any public use of the licensed data according to terms contained within the license. Share-alike licenses typically require the preservation of the original license for any derivative works.
- **Attribution:** Allows others to use, modify, and distribute with only one condition: all versions must credit the original creator. This license type is the least restrictive and allows for commercialization of derivative works.

Additional Resources

Open Source Initiative

The Open Source Initiative (OSI), founded in 1998, is a non-profit corporation that is the recognized entity within the open source development community for approving software licenses as conforming to the Open Source Definition. The OSI website provides a categorized list of open source agreements, which may provide applicable licensing models for the DCM / DMA programs.[101]

Department of Defense Open Technology Development

The DoD's Open Technology Development (OTD) initiative, which began in 2006, provides a great deal of excellent guidance to help government personnel and contractors implement open source software (OSS) development. A recent report, Open Technology Development: Lessons Learned and Best Practices for Military Software, documents the issues involved in public sector

[98] Ibid.
[99] Open Government Directive. http://www.whitehouse.gov/open/documents/open-government-directive
[100] The Open Knowledge Foundation Wiki. http://wiki.okfn.org/Main_Page, Open Data Commons. http://opendatacommons.org/
[101] http://www.opensource.org/licenses/category

open source development, and includes a detailed discussion on selecting an appropriate OSS license.[102]

Lessons Learned

Copyright and licensing present challenging policy issues in the context of the DCM and DMA programs. The choice of OSS license will have significant implications for how applications will be used by government agencies and the private sector. Similarly, copyright protections may become an issue for users of databases associated with the open data environments, unless steps are taken to ensure that the data products issued are either dedicated to the public domain or otherwise licensed for free and open use.

Conclusion

In both OSS licensing and open data licensing, options exist which ensure that the licensed product will remain free and open forever (e.g. GPL, Share-alike). There also are options that enable products to be incorporated into proprietary products (e.g. MIT/X, basic attribution licenses). Choosing which license is the best policy fit for each product will be related to the intended users and philosophy of the stakeholders involved. However, as DCM and DMA products are likely to involve parties who would naturally retain ownership and copyright of content they create in some circumstances, setting a clear policy on OSS licensing and open data licensing or public domain dedication appears to be important.

The analysis associated with these policies and recommendations for implementation will be documented in two forthcoming reports—one on the policies associated with the OSADP and one on open data environments.

Next Steps

- Develop a taxonomy identifying the key IP issues and their interrelationships. For example, this taxonomy would document both inbound and outbound software licensing issues.
- Engage legal counsel from participating agencies to draft IP policies and related licenses, terms-of-use statements, and other related documentation.
- Vet the draft policies and documentation with user groups, particularly the software development communities that are anticipated to participate in the DCM /DMA programs.
- Working closely with legal counsel, modify the draft policies where appropriate to reflect feedback from user groups.
- Develop implementation timeline.

[102] Open Technology Development (OTD): Lessons Learned and Best Practices for Military Software. 2011-05-16 Sponsored by the Assistant Secretary of Defense (Networks and Information Integration) (NII) / DoD Chief Information Officer (CIO) and the Under Secretary of Defense for Acquisition, Technology, and Logistics (AT&L).

Chapter 7 Liability

Recommended Policy Option

In concert with the development of licenses to protect intellectual property, agencies involved in the DCM / DMA programs must put in place clear limits to liability. These should include broad risk indemnification for the providers of open data, and also specific protections written into license agreements to protect developers and source code providers as well as participating public agencies.

Licenses also need to be explicit about the circumstances under which they do *not* offer liability protection, so that all parties involved understand all potential risks.

Definition

As discussed in Chapter 6, the practice of licensing is an important protection as well as foundation for business models: the result is to assign liability to the licensee, which frees the products, services, and practices from the associated intentional misuse or inadvertent mistakes of users.

The remaining concerns are:

- Liability stemming from software or data quality problems (e.g., inaccurate data, applications failing to work as intended), or misuse of the data.
- The potential liability that exists if security is breached and exposes personally identifiable information (PII) or individual locational information (such as GPS data).

Data quality liability is, perhaps, the simplest category to address, at least with respect to indemnification. Public agencies providing data for public use frequently publish "disclaimers of warranty" and "limitations of liability" on the websites through which they provide data access. These statements serve two main purposes. First, they insulate and agency (and its employees) from liability by indicating that the data are provided "as is" without any warranty of accuracy, and that the agency bears no responsibility circumstances stemming from the use of the data. Second, such statements also limit the extent of the indemnification to the agency itself – that is, they make clear that the agency does not indemnify the users of the data.

With respect to liability related to software quality, open source licenses shift all risk for intellectual property infringement to the licensee. According to the ABA, this is because:

> *…contributors do not vouch for the cleanliness of the code they contribute to the project; in fact, the opposite is true -- the standard open source license is designed to be very protective of the contributor. The typical license form does not include any intellectual property representations, warranties or indemnities in*

> *favor of the licensee; it contains a broad disclaimer of all warranties that benefits the licensor/contributors*.[103]

This type of license structure stands in contrast to most commercial software licenses, which typically require the licensor to provide a level of assurance that the licensed software technology does not infringe intellectual property rights.

In pursuing an "open approach" to software and actually providing funding for applications development, a critical and potentially conflicting issue arises for the federal government that has liability implications—the consequence of unintentionally violating intellectual property rights while pursuing the opportunity for third-party entities to use the open source applications to commercialize derivative products, thereby creating a diverse set of applications to benefit the public. This issue must be addressed through appropriate IP licensing options that protect original work and provide mitigation against infringement and liability.

There is a great deal of overlap between policies to address IP concerns and policies to deal with liability. Both can be addressed through licensing agreements that define specific protections and limitations. Additionally, unambiguous terms of use policies that protect public agencies from liability stemming from the end use of data and/or applications those agencies provide are important.

Liability Strategies: State-of-the-Practice Examples

The following examples occur in the categories of limitations of liability, open data licensing, software quality and open source application licensing, and PII security breaches.

Open Data Limitation of Liability: King County, WA

King County, Washington operates an open data website as a public service. The website provides access to a variety of data sets, including transportation data. The County has a detailed open data "terms of use" statement on the website, including disclaimer of warranty and limitation of liability statements, which read, in part:

> All information on this website is provided "as is", "with all faults" and "as available" without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of merchantability, fitness for a particular purpose, accuracy, or non-infringement nor shall the distribution of this information constitute any warranty. King County assumes no responsibility for errors or omissions in the information or software or other documents which are referenced by or linked to this website. Under no circumstances, including, but not limited to, negligence, shall King County, its officials and employees, or any contributor to this website be liable for any direct or indirect damages, even if both parties are aware of the possibility of such damages, including without limitation loss of profits or for any other incidental, special, consequential or exemplary damages, however caused, whether based upon contract, negligence, strict liability in tort, warranty, or any other legal theory, arising out of or related to your use of,

[103] American Bar Association, at: http://www.americanbar.org/groups/intellectual_property_law/resources/an_overview_of_open_source_software_licenses.html

> or the inability to use, this website or its content. King County is not responsible for or liable for any damage, including damage caused by viruses, to your computer, computer system, or other property, during or on account of access or use of this website or any sites to which this website provides links. [104]

Open Data Limitation of Liability: Southeast Atlantic Coastal Ocean Observing System

The Southeast Atlantic Coastal Ocean Observing System (SEACOOS) was an open data repository for a wide variety of oceanographic data.[105] The website's Data Liability and Access Policies provide an example of the use of an important nuance that makes explicit the fact that the government's own indemnification does not extend to users of the open data it provides:

> *The United States Federal Government and SEA-COOS associated partners do not assume liability to the Recipient or third persons, nor will the Federal Government and SEA-COOS associated partners indemnify the Recipient for its liability due to any losses resulting in any way from the use of this data set.*[106]

Open Data Licensing: London, Canada

The City of London, Canada's open license states plainly that it provides the open data on its site without copyright or licensing restrictions, and that third-party users of the data likewise are given no proprietary interests in the data:

> *The City of London (City) now grants you a world-wide, royalty-free, non-exclusive license to use, modify, and distribute the data sets in all current and future media and formats for any lawful purpose. You now acknowledge that this license does not give you a copyright or other proprietary interest in the data sets. If you distribute or provide access to these data sets to any other person, whether in original or modified form, you agree to include a copy of, or this Uniform Resource Locator (URL) for, these Terms of Use and to ensure they agree to and are bound by them but without introducing any further restrictions of any kind.*[107]

Software Quality and Open Source Application Licenses: NASA

NASA's Open Source Agreement (NOSA) permits licensees of open source software originally offered by NASA to offer their own (i.e., third party) "warranty, support, indemnity and/or liability obligations..." However, any such protection is tightly constrained:

[104] King County, Washington, at: http://www.kingcounty.gov/About/dataTermsOfUse.aspx

[105] In 2008 SEACOOS became part of the Southeast Coastal Ocean Observing Regional Association. The SEACOOS website is now and archived site.

[106] http://seacoos.org/Data%20Access%20and%20Mapping/Document.Disclaimer

[107] http://www.london.ca/d.aspx?s=/Open_Data/Open_Data_Terms_Use.htm

> *A Recipient may choose to offer, and to charge a fee for, warranty, support, indemnity and/or liability obligations to one or more other Recipients of the Subject Software. A Recipient may do so, however, only on its own behalf and not on behalf of Government Agency or any other Recipient. Such a Recipient must make it absolutely clear that any such warranty, support, indemnity and/or liability obligation is offered by that Recipient alone. Further, such Recipient agrees to indemnify Government Agency and every other Recipient for any liability incurred by them as a result of warranty, support, indemnity and/or liability offered by such Recipient*.[108]

PII Security Breach

Potential liability stemming from theft or accidental release of PII is a critical issue. Federal agencies have formal procedures in place for notification of and response to PII security breaches, but such procedures do not necessarily insulate agencies from liability.[109] Federal privacy statutes lag behind the technical capacity of criminals to breach privacy. Every state has its own laws on privacy and confidentiality, so in the absence of an overarching federal law, liability would most likely be determined based on the domicile state of a defendant or, in the case of a corporation, where it is incorporated.

Appropriately, federal agencies have put a great deal of effort into preventing the release of PII in the first place. In addition to robust security protocols, best practices in this area include storing all PII data in encrypted form.

Lessons Learned

Open source application development liability concerns, stemming principally from potential IP violations, can be addressed through licensing. Numerous models are of open source licensing agreements are emerging that delineate the intellectual property boundaries (or lack thereof) of open source products. However, developing an open source licensing arrangement that addresses liability concerns while simultaneously supporting commercialization may prove to be a key challenge for the DMA program in particular.

Conclusion

With freedom and flexibility comes risk—dissemination of open data and open-source application development create liability concerns. Detailed and unequivocal "disclaimers of warranty" and "limitations of liability" can insulate agencies providing open data, but existing examples of such use of open data (e.g., accessing a website and downloading an open data set) appear to be miniscule compared with the active use of such data by dynamic mobility applications. For the envisioned DCM / DMA programs, liability protections will need to be codified in licenses for data and applications on both the "inbound' and "outbound' sides of the anticipated development portal.

---

[108] http://www.opensource.org/licenses/NASA-1.3

[109] e.g., http://www.hhs.gov/ocio/policy/20080001.003.html, http://www.doncio.navy.mil/ContentView.aspx?ID=852

Next Steps

- In parallel with the development of IP licensing as discussed in the previous section, define specific liability limits, and protections for all involved parties, including application developers, source code providers, and participating agencies. This process will be time consuming, because it will need to involve legal counsel from all participating agencies.
- Draft limitation of liability statements for all involved government agencies, and related indemnification statements for applicable open-source licenses. Distribute to potentially affected entities for review and comment.
- Finalize and publish.

Chapter 8 Governance Options

Recommended Policy Option

Comprehensive and well-defined governance frameworks for data, projects, and development portals are essential to specify the roles and responsibilities of participants and the processes by which decisions are made.

Data.gov is recommended as a model for effective data governance, and Forge.mil for project governance. The DMA's OSADP will need to define portal governance based on an analysis of the portal's risks and opportunities.

This chapter discusses governance from three perspectives: data governance, project governance, and portal governance.

Data Governance Definition

Data governance is a set of processes that address quality, management, policies, standards, metadata organization, and other issues associated with data.[110] A governance structure frames roles and responsibilities in relation to authority (i.e., scope, sanctions, and enforcement), rules of conduct, standards, and metadata. The governance model offers a structure to define which people and entities can take what actions, with what information, under what circumstances, and using what methods.[111] It also establishes the means by which those "governed" are able to influence the overall scope and decisions of the governing body, as well as mechanisms for appeal and/or adjudication of contestable actions.

The governance of open data is of increasing interest within the Federal government, particularly in the context of the Open Government Initiative, which pushes government agencies to make high-value data more freely available to the public online.[112]

Data Governance: State-of-the-Practice Examples

Data.gov

Data.gov is the flagship product of the Open Government Initiative. It is a central repository for thousands of data sets submitted by every federal agency for transparent release to the public.

[110] Sarsfield, Steve (2009). "The Data Governance Imperative," **IT Governance.**
[111] Data Governance Framework. Data Governance Institute. At http://www.datagovernance.com/dgi_framework.pdf
[112] Open Government Directive. http://www.whitehouse.gov/open/documents/open-government-directive

Data.gov offers the clearest set of requirements for data governance by describing how submitting agencies must agree to conform to specific roles and responsibilities. The following are taken directly from the Data.gov website:

- **Public Information.** All data sets accessed through Data.gov are confined to public information and must not contain National Security information as defined by statute and/or Executive Order, or other information/data that is protected by other statute, practice, or legal precedent. The supplying Department/Agency is required to maintain currency with public disclosure requirements.
- **Security**. All information accessed through Data.gov is in compliance with the required confidentiality, integrity, and availability controls mandated by Federal Information Processing Standard (FIPS) 199 as promulgated by the National Institute of Standards and Technology (NIST) and the associated NIST publications supporting the Certification and Accreditation (C&A) process. Submitting Agencies are required to follow NIST guidelines and OMB guidance (including C&A requirements).
- **Privacy.** All information accessed through Data.gov must be in compliance with current privacy requirements including OMB guidance. In particular, agencies are responsible for ensuring that the data sets accessed through Data.gov have any required Privacy Impact Assessments or System of Records Notices (SORN) easily available on their websites.
- **Data Quality and Retention.** All information accessed through Data.gov is subject to the Information Quality Act (P.L. 106-554). For all data accessed through Data.gov, each agency has confirmed that the data being provided through this site meets the agency's Information Quality Guidelines. As the authoritative source of the information, submitting Departments/Agencies retain version control of data sets accessed through Data.gov in compliance with record retention requirements outlined by the National Archives and Records Administration (NARA).
- **Secondary Use**. Data accessed through Data.gov do not, and should not, include controls over its end use. However, as the data owner or authoritative source for the data, the submitting Department or Agency must retain version control of data sets accessed. Once the data have been downloaded from the agency's site, the government cannot vouch for their quality and timeliness. Furthermore, the US Government cannot vouch for any analyses conducted with data retrieved from Data.gov.
- **Citing Data.** The agency's preferred citation for each data set is included in its metadata. Users should also cite the date that data were accessed or retrieved from Data.gov. Finally, users must clearly state that "Data.gov and the Federal Government cannot vouch for the data or analyses derived from these data after the data have been retrieved from Data.gov."
- **Public Participation.** In support of the Transparency and Open Government Initiative, recommendations from individuals, groups and organizations regarding the presentation of data, data types, and metadata will contribute to the evolution of Data.gov.
- **Applicability of this Data Policy**. Nothing in this Data Policy alters, or impedes the ability to carry out, the authorities of the Federal Departments and Agencies to perform their responsibilities under law and consistent with applicable legal authorities, appropriations, and presidential guidance, nor does this Data Policy limit the protection afforded any information by other provisions of law. This Data Policy

> is intended only to improve the internal management of information controlled by the Executive Branch of the Federal Government and it is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity, by a party against the United States, its Departments, Agencies, or other entities, its officers, employees, or agents.[113]

The Transportation Secure Data Center

The National Renewable Energy Laboratory and the U.S. Department of Transportation launched the TSDC as a way to distribute valuable new information on vehicle travel patterns obtained through the use of GPS location technology obtained by planning agencies through vehicle travel surveys. The addition of GPS location data increases the value of traditional travel survey data significantly but raises serious concerns over how to prevent this information from being used to identify individual vehicles and their owners.

In order to widely distribute these data sets, the TSDC has developed a two-tiered access system. For each data set, TSDC creates "scrubbed" versions of the data sets with latitude and longitude information and other potentially PII removed. These data sets are openly available to all users after completion of a short registration form. Access to the detailed location data is only available to users who complete a rigorous screening process. Researchers who receive access to this detailed data may only access it within a secure online data environment and are not allowed to download it. They may however create aggregated results.[114]

Virtual USA

The Department of Homeland Security created Virtual USA to encourage information sharing and collaboration across jurisdictions within the homeland security and emergency management community. Links to information are voluntarily submitted by participating agencies for inclusion in the portal.

Virtual USA follows a highly decentralized data storage and governance model where the contributing agencies retain ownership, storage, security, updating responsibilities and access control responsibilities for the data they submit. Each submitting agency retains full control over the data they submit, with remote users only able to access read-only versions.

Virtual USA encourages the use of common, open, non-proprietary data formats that facilitate the easy sharing of data. However, Virtual USA is officially technology agnostic and does not require that data be made available using any particular software or in any particular data format. This enables all contributors to share their data without requiring users to adopt a particular solution.[115]

[113] http://www.data.gov/datapolicy
[114] http://www.nrel.gov/vehiclesandfuels/news/2011/1427.html
[115] http://www.firstresponder.gov/Documents/vUSA_FAQs.pdf

Project Governance Definition

Project governance is a framework for decision-making and management of a project. The governance structure of a project determines the roles and responsibilities of the participants, with a particular emphasis on how decisions are made. In the world of open source software development, project governance establishes the rules by which collaborators may contribute to a project, how contributions will be evaluated and accepted/rejected, and how disputes will be resolved. Open source project governance tends to encourage consensus decision-making through constructive debate. However, when consensus does not emerge in a timely fashion, projects tend to follow one of two models:

Centralized Control Model

One person is in charge of all final decisions. This person may choose to delegate some authority to others, but retains final approval and veto authority. This approach is most common in small projects where one team member has a much greater understanding of the project than others.

Group Decision-Making Model

All final decisions are made by the group. Decisions can be made through a variety of mechanisms including: simple majority vote, consensus, and lazy consensus (where not voting is counted as a consenting vote). A common mechanism for voting is the Apache Software Foundation scoring mechanism where "yes" votes receive a "+1" and "no" votes receive a "-1." Some projects choose to allow any group member to veto (consensus requirement), others set requirements for the total score that must be achieved for a vote to pass using a simple majority (e.g. "+1" passes) or some modified majority (e.g. "+3" passes). Another common practice is to allow all votes to pass unless someone vetoes the proposal (lazy consensus). There is always a risk of veto abuse in group decision-making. Therefore, many projects require that vetoes be justified and encourage voting as the method of last resort, placing a greater emphasis on consensus. In the Group Decision-Making Model, the voting group can be the pool of all project contributors, or some smaller subset of key contributors who have the power to vote contributors in or out of the voting group.

Forking

Another key aspect of open source project governance is the requirement that projects be "forkable." Forking is when a project contributor is so unhappy with the decisions made by the governance structure that he breaks off a parallel development effort with a different governance structure. Although these "forks" typically fail, the threat of a fork is a strong incentive to project leaders to be responsive to contributors, lest a strong fork develop into the dominant project, rendering the previous project irrelevant.[116]

[116] *OTD Lessons Learned*. Department of Defense. http://cio-nii.defense.gov/sites/oss/OTD-lessons-learned-military-signed.pdf; *Why does a project need a governance model?* OSS Watch Wiki http://wiki.oss-watch.ac.uk/GovernanceModel.

Project Governance: State of Practice Examples

Forge.mil

Forge.mil is an online portal for Agile open source application development in support of the Department of Defense (DoD). Agile software development focuses on producing frequent, small improvements to working software as opposed to large, comprehensive overhauls or creating extensive product documentation. The Agile philosophy is that by focusing on near-term deadlines and goals, project teams can better adapt to changing customer requirements while avoiding wasted effort working on things that will eventually be abandoned or discarded. The Agile philosophy also believes that self-organized, self-governing teams produce better plans, requirements, and product architectures than top-down project management structures.[117]

Forge.mil provides two environments for project development designed for different levels of project control. SoftwareForge is open to all approved users of Forge.mil (approved U.S. military, DoD civilians and DoD contractors only) and mirrors the popular SourceForge.net open source application development portal. ProjectForge is similar to SoftwareForge but allows the project manager to determine which users have access to view and contribute to the project.

Individual projects on SoftwareForge or ProjectForge determine their own project governance rules.[118]

CONNECT

CONNECT is governed by a centralized Managing Work Group which sets the business agenda for the web portal. This agenda is implemented by the Change Control Board (CCB) which is a representative board consisting of members appointed by each contributing federal agency. The CCB provides a centralized authority for proposing, reviewing, and incorporating changes. The primary functions of the CCB are to:

- Authorize the establishment of baselines
- Authorize additions of user stories to baselines
- Represent the interest of all groups who may be affected by changes to the baselines
- Evaluate and approve, disapprove or defer proposed system changes
- Set timeline for enhancements and changes to the baseline
- Ensure implementation of approved changes

[117] *"Principles Behind the Agile Manifesto,"* http://agilemanifesto.org/principles.html

[118] http://forge.mil/Faqs.html. Defense Information Systems Agency (DISA) Presentations on Forge.mil:
Forge.mil on Ramp to the DoD Cloud http://www.disa.mil/conferences/2011/briefings/Forge_On_Ramp.ppt
Forge 101: An Introduction to Forge.mil http://www.disa.mil/conferences/2011/briefings/forge_101.ppt

The FHA, which created and oversees the CONNECT solution plans to migrate CONNECT to a public/private governance model. This change enables private sector stakeholders to have a role in the project governance of CONNECT.[119]

Portal Governance Definition

A web portal or links page is a website that functions as a point of access to information on the World Wide Web. It presents information from diverse sources in a unified way. In the context of open source application development, a portal contains the tools through which the contributors, users, testers, and project leaders interact (e.g. source code repository, wiki, forums, bug tracker). Our research found no published material on how open source application development portals are governed, specifically. Therefore, this section focuses on models for web portal governance generally.

Portal governance is the structure that determines how the various different teams will interact together to ensure that the portal meets the needs of the customer and the business needs of the governing organization. A governance model determines how website developers, administrators, interface designers, content creators, business marketing, portal users, and IT support will interact to ensure the efficient and successful operation of the portal.

Portals can be structured along a continuum from fully-centralized to fully-decentralized, with most opting for a compromise, sometimes called “federated.” Descriptions of these three governance models follow:

Centralized

The centralized portal governance model follows a typical top-down organizational structure, where one person or small group controls all final decisions, sets rules, and enforces processes. This was once the dominant model for businesses of many types, although it has fallen out of favor in large organizations due to the resources required for sustaining it, and the negative effect that a single individual can have on the organization.

Decentralized

The decentralized portal governance model has no central command structure. All rules and decisions are made collectively by self-defined groups with common interests. This model offers freedom, but provides little consistency, guidance, or support.

[119] http://www.connectopensource.org/about/governance

Federated

The federated governance model retains a strong central entity, but with numerous loosely connected entities beneath it. In this model the central authority controls only those roles and process that benefit all stakeholder groups (e.g. portal policies and procedures).The smaller units are then provided the freedom to determine their own needs, structure and design. Most portals follow use some form of federated governance structure.[120]

For examples of typical portal roles and processes associated with web portal governance, see *Winning Strategies for Portal Governance*.

Portal Governance: State-of-the-Practice Examples

Our research was unable to identify published sources that detail specific governance structures for portals, either federal or otherwise. While it is clear that significant effort and expertise go into designing and executing portal governance structures, few if any of these governance structures are freely available to the public. If further information is desired, further research efforts could focus on interviewing the project managers of Forge.mil and other open source application development portals to seek clarification on the specific governance models that are currently being used in practice.
The following are summaries of web portal best-practices literature we found relevant to this research.

Defining a Governance Model for Portals

The governance model for a portal should define the roles, processes and implementation mechanisms that will be required to manage the portal throughout its development, operation and updating.

Defining objectives is a key aspect of early portal governance decision-making (e.g. minimize legal risk, enable quick decision-making). Objectives should reflect the need for the portal and be driven by the client.

Development phases and phase-specific governance focuses should be defined.
Defining roles and responsibilities is a critical step in portal governance design. Roles should reflect the existing structure and inter-relationships between portal stakeholders. Responsibilities associated with each role may change depending on the project phase.

Example roles include:

- Steering Committee
- Portal Governance Board
- Core Team
- Extended Team(s)
- Portal Services Team(s)

[120] Roth, Craig. *Website Governance: A How-to Guide.* http://www.craiglroth.com/Opinions%20In%20Depth%20-%20web%20governance.pdf Behl, Pardeep. *Winning Strategies for Portal Governance*. http://www.ibm.com/developerworks/websphere/library/techarticles/0904_behl/0904_behl.html

The composition of these teams will vary based on the needs of the portal and the services implemented.

Process definition is another step in the formation of a portal governance model. Processes must be identified, defined, and mapped to defined roles. Examples of processes include:

- Prioritization and Release strategy
- Site Brand Management and user Experience
- Communication
- Site Policies and Compliance
- Site Taxonomy
- Content Management

Comprehensive stakeholder involvement is important at all stages of governance definition.

- Implementation of the governance model requires clear communication with all stakeholders. Strategies for stakeholder engagement include: workshops, wikis, training capsules, and best practices dissemination.[121]

Conclusion

Development of governance frameworks at the data, project, and portal level for the DCM / DMA programs will be challenging, due to the wide variety of data involved, as well as the desired open-source development environment, and open data environments. Therefore, governance work should be among the first tasks initiated, so that it can start simply and evolve in parallel with the design of the programs.

Next Steps

- Identify and document the key risks for the DCM and DMA programs at the data, project, and portal levels.
- Draft governance policies to address the risks, using data.gov and forge.mil as models.
- Solicit comments on the policies from user groups, and revise as appropriate.
- Develop implementation timeline

[121] http://www.infosys.com/consulting/systems-integration/white-papers/documents/portals-governance-model.pdf

Chapter 9 Open Data Maintenance

Recommended Policy Option

Ongoing review and updating of the open data supporting the DCM /DMA programs and applications will be a significant and vital, undertaking. Therefore, data maintenance policies and procedures need to be implemented in advance, and must establish the protocols in three key areas: data review, data monitoring and assessment, and data updating.

Definition

Maintaining the huge volume of data available via open data environments is critically important. Without ongoing attention, existing data sets can quickly become outdated and inaccurate. The inherent variability of transportation data, along with the fact that these data increasingly are flowing into open data environments from myriad mobile sources, makes the need for data maintenance especially important. As one observer has commented:

> The aggregation of new data sources, such as smart phones, RFID and new wireless sensor technologies combined with emerging data-rich environments such as the connected vehicle will place a significant data management load on operating entities. New data resources and corresponding "big data" curation and management needs will require public agencies to implement "data managers", either in-house, or through supporting third party contracts, in order to establish proper data frameworks.[122]

Open Data Maintenance: State-of-the-Practice Examples

Open Data Kit (ODK)

Open-source tools are emerging that enable the collection and management of data via mobile devices (and attached sensors). One such set of tools is Open Data Kit (ODK) from the University of Washington, which "is a free and open-source set of tools which help organizations author, field, and manage mobile data collection solutions." [123] Among its many functions, the ODK allows the centralized collection (and therefore updating and management) of data from all participating devices. Numerous organizations worldwide are currently using ODK for a wide range of projects, including gathering emergency response and preparedness field data, conducting transit worker surveys, and collecting environmental data.

[122] www.terranautix.com
[123] http://opendatakit.org/

In addition to versatile tools that permit data resources to be maintained as needed, effective data maintenance processes are essential to ensure the ongoing timeliness, accuracy and utility of shared data. A recently published TRB paper documents the results of a thorough review of data maintenance processes at the New York Metropolitan Transportation Council (NYMTC), and provides detailed recommendations for a dynamic data maintenance procedure.[124] This timely paper thus provides a data maintenance best practices "blueprint." The recommended data maintenance procedure includes three steps – review, monitoring/assessment, and modification/improvement – that the researchers suggest be added to the typical data lifecycle of creation, distribution, access, and updating.[125] The data maintenance steps are described as follows:

1. **Review**. This step occurs after data have been submitted initially, and involves looking for any issues in terms of various dimensions of data quality such as coverage, appropriateness of format, availability and content of metadata, accuracy, integration, accessibility and so on. If the data product presents conformity in terms of these various aspects of data quality, it will be distributed to the end-users. If not, it shall undergo necessary modifications. This part of the procedure should be repeated until desired conformity to the desired level of data quality is achieved.
2. **Monitoring and Assessment**. This step occurs after data are distributed or made available to end-users, and involves using a variety of tools to determine if modifications are warranted. These tools include online user feedback forms; website traffic monitoring tools to track successful data retrievals, returning users, and data product popularity; stakeholder surveys; reviews of emerging best practices; and use of online collaborative tools to allow users to exchange information.
3. **Modification and Improvement**. This step is a "feedback mechanism" that allows the incorporation of the results from the monitoring and assessment component back into the initial data creation step.

Lessons Learned

Two fundamental challenges exist with respect to open data maintenance for DCM and DMA. The first challenge is technical: development of systems to allow automated updating of data sets whose content comes from numerous sources, many of them mobile. This challenge, while significant, can be addressed by drawing on examples of application of open-source tools such as ODK.

The second challenge is more difficult: implementation of a data maintenance process that enables ongoing data review, monitoring, and modification to occur. This challenge is significant precisely because it cannot be addressed by purely technical means; successful implementation will involve significant administration and collaboration. For example, to enable data review, review criteria will need to be developed through workshops and other collaborative means, all of

[124] Ozmen-Ertekin and Kaan, "Dynamic Data Maintenance for Quality Data, Quality Research". TRB Paper No. 11-3093.
[125] For brevity, this discussion omits some other recommendations in the TRB paper, such as conducting workshops with data providers to establish data submission procedures.

which will require significant effort. Similarly, data monitoring will include user surveys, best-practices reviews, and ongoing information exchange between users.
Maintaining open data for DCM and DMA will be an ongoing, shared enterprise. It will demand significant, sustained commitment and collaboration by agencies, developers, and outside data providers.

Conclusion

The tasks involved in open data maintenance are highly related to those for data quality assurance. Effective data maintenance requires the implementation of policies and protocols for data review and modification. Like data quality assurance, data maintenance must be approached programmatically and proactively.

Chapter 10 Data Management Policy Considerations

The DCM Program and, in particular, the RDE will be subject to federal data management policies for as long as the Federal government hosts the RDE. A preliminary review of existing DOT and OMB policies suggests, however, that current examples do not directly address the complex aggregation of data sets as envisioned in the RDE. Further investigation and possibly formal inquiry will be necessary in order to determine the degree of applicability of these policies to the RDE. Analysis[126,127] of OMB and DOT guidance[128,129,130,131,132] has surfaced two major issues which will require further consideration and resolution:

- Which data in the RDE will be subject to federal data quality guidelines? The RDE will provide access to data sets furnished by other entities, both public and private. Existing guidance does not directly address this ownership/access configuration, which goes beyond simple hyperlinks from a federal website to non-federal data.
- Who will ensure data quality, and to what degree? Much of the federal data management guidance is aimed at ensuring high quality data. The guidance requires intensive review and scrutiny of data sets before they are made available to the public.

If non-federal data in the RDE were determined to be subject to federal data quality policy, the time and cost burdens of federally-defined "pre-dissemination review" requirements could discourage participation in the RDE or otherwise interfere with the intended use of the site.

Because the full scope of the RDE is not yet understood, the DCM Program leadership has an opportunity to consider these two key questions and the options available to them. Key questions include the following:

[126] Review of DOT Order 1351.34 Departmental Data Release Policy (DRAFT). Noblis. April 8, 2011.

[127] Chapter 2 "Data Exchange Policy Issues" in Identification of Policy Issues for the DCM and DMA Programs (DRAFT). U.S. DOT/John A. Volpe National Transportation Systems Center. July 2011.

[128] OMB Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility and Integrity of Information Disseminated by Federal Agencies. January 2002 and Department of Defense Open Technology Development: Lessons Learned and Best Practices for Military Software. Department of Defense. May 2011.

[129] DOT Report for Implementing OMB's Information Dissemination Quality Guidelines. Bureau of Transportation Statistics. August 2002.

[130] DOT Order 1351.34 Departmental Data Release Policy. March 2011.

[131] NARA Bulletin 2010-05 *Guidance on Managing Records in Cloud Computing Environments*. National Archives and Records Administration. September 2010.

[132] Open Technology Development: Lessons Learned and Best Practices for Military Software. Department of Defense. May 2011.

- **Will Non-Federal Data Sets in the RDE Be "Federally Disseminated"?** The RDE will be "a *system of systems* linking multiple data management systems. The RDE is considered owned by the USDOT, but will be maintained and controlled outside of the USDOT, through a common web-based Data Portal"[133] (i.e., the portal will provide access to data housed and maintained by non-federal agencies, universities, and private entities). If any data coming into the RDE, or linked to via the RDE must be considered federal, those data would need to be scrubbed to meet federal standards. Current OMB guidance provides some direction on this topic, with respect to what data fall under the definition of "federally disseminated. [134] Where the RDE and associated third-party data sets would fall, however, is not entirely clear.
- **Who must ensure the quality of non-federal data sets in the RDE?** Any extra-DOT data sets which are ultimately judged to be "Federally disseminated" will be subject to procedures for ensuring that public-facing data are high quality and fully accessible. OMB and DOT guidance focus on the type of scrutiny required for federal data, to ensure overall data quality and of avoiding harm through release of incorrect data or PII.

The following requirements for Federally-disseminated data have the potential to discourage non-federal entities from publishing data sets to the RDE:

- DOT allows publication of aggregate data provided it is accompanied by micro-data. Micro-data sets may be too cumbersome to store and/or irrelevant to users of the RDE.
- Data "released through the Web" must be 508-accessible.
- Data must come from "reliable sources." This raises the question of whether potential users of the RDE must be screened for "reliability" before contributing a data set.
- Reproducibility standards must be applied to both original and supporting statistical data. It will be important to ensure that researchers are not discouraged from sharing statistical data if they cannot also provide raw data.
- When a data set is first released or substantially changed, the DCM program would need to conduct a "pre-dissemination review" in consultation with Counsel, the DOT, CIO, BTS, and the DOT Office of Intelligence, Security, and Emergency Response. This would incur delays and costs for which non-federal entities might not be prepared.
- The DCM program office would need to be able to provide additional data on the subject matter of any covered information it disseminated.

[133] *Concept of Operations: Data Capture and Management Research Data Exchange [DRAFT].* RITA. August 2011.
[134] OMB Part V Definitions.

- Data sets must be registered with the DOT Services/Data Architecture Group Metadata Registry, for which requirements are not currently available.

In the event that non-DOT data sets published to the RDE are not judged to be "Federally-disseminated," these requirements would become moot, and a potentially large threshold to participation would melt away.

One reference to pursue is in the supplemental information published with the Final OMB Guidelines:

> In some cases, for example, the data disseminated by an agency are not collected by that agency; rather, the information the agency must provide in a timely manner is compiled from a variety of sources that are constantly updated and revised and may be confidential. In such cases, while agencies' implementation of the guidelines may differ, **the essence of the guidelines will apply**. That is, these agencies must make their methods transparent by providing documentation, ensure quality by reviewing the underlying methods used in developing the data and consulting (as appropriate) with experts and users, and keep users informed about corrections and revisions."[135]

Another alternative may be to develop or encourage a "pre-accreditation" process or norm for data sets, building upon practices already routine in the data-gathering community. The DoD has been transitioning its software development from historical lengthy government review processes to very short test and build cycles. This is made possible by using components that are pre-certified or accredited.

[135] Supplementary Information provided with Final Guidelines.

Conclusion

The connected vehicles concept offers significant potential safety and mobility benefits. Before connected vehicles programs can be implemented, however, numerous important policy issues must be resolved, particularly with respect to DCM and DMA applications and associated data environments. The government's commitment to open-source and open data environments, while promising to accelerate applications development, creates additional complexities in terms of intellectual property, licensing, data quality, and other factors. Our research suggests that many state-of-the-practice approaches can serve as models for policies and protocols to address these and other issues. At the same time, however, it is clear that existing practices will need to be modified and, in some cases, expanded to meet the needs of the DCM and DMA programs, which are breaking new ground in many respects.

Certain issues, notably security, privacy, and user access, can be addressed in part through technical solutions (e.g., communication "anonymizers" to help ensure privacy). These approaches are comparatively simple; the non-technical approaches – the development of policies and protocols – will be quite challenging in the context of the DCM / DMA programs, given the diversity of data suppliers and users involved. Nevertheless, policies and protocols (e.g., for data governance) will be essential; they will provide the foundations upon which all other elements of the programs will be built. Work on the creation of policies and protocols will be time consuming, and therefore must begin during the earliest stages of DCM / DMA program development. The process should include significant outreach efforts to identify and engage all stakeholders – representatives from data provider and data user communities, as well as privacy advocates and other interest groups. Achieving consensus may, in many cases, be arduous, but "shortcutting" the process will ultimately lead to failure. For the policies and protocols to be effective, all stakeholders must accept them.

Bibliography

Behl, Pardeep. "*Winning Strategies for Portal Governance"*. Web article published by IBM, June 29, 2009. Located at: http://www.ibm.com/developerworks/websphere/library/techarticles/0904_behl/0904_behl.html.

Bureau of Transportation Statistics, ***Guide to Good Statistical Practice in the Transportation Field***, BTS, 2003. Report located at: http://www.bts.gov/publications/guide_to_good_statistical_practice_in_the_transportation_field/html/chapter_06.html

Cooper, Scott P., Navid Soleymani & Clifford S. Davidson (Proskauer Rose LLP) and Tanya L. Forsheit (InfoSecCompliance LLC). "*State Privacy Laws*", Synopsized from Chapter 5, ***Proskauer on Privacy: A Guide to Privacy and Data Security Law in the Information Age***, October 2006. Treatise located at: http://www.pli.edu/product_files/booksamples/11513_sample5.pdf

Federal Trade Commission, ***Protecting Personal Information: A Guide for Business***. FTC, 2011. Guidance located at: http://business.ftc.gov/documents/bus69-protecting-personal-information-guide-business

Federal Trade Commission, "*Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework for Businesses and Policymakers*," Guidelines produced by FTC, December 2010. Located at: http://www.ftc.gov/os/2010/12/101201privacyreport.pdf.

Fogel, Karl. ***Producing Open Source Software: How to Run a Successful Free Software Project***. Web published under the Creative Commons Attribution-ShareAlike License. Located at: http://producingoss.com/

Gellman, Robert, "*Fair Information Practices: A Basic History*" (unpublished essay, July 15, 2011), at: http://bobgellman.com/rg-docs/rg-FIPShistory.pdf.

Hoffman, Andrew, "*It's Not Too Late to Come to the Party: Mississippi Joins 45 Other States by Enacting a Security Breach Notification Law*", Web article published by Proskauer Law, April 13, 2010. Located at: http://privacylaw.proskauer.com/2010/04/articles/data-breaches/its-not-too-late-to-come-to-the-party-mississippi-joins-45-other-states-by-enacting-a-security-breach-notification-law/

Koorn, R. et al. (KPMG), ***Privacy-Enhancing Technologies: White Paper for Decision-Makers***. 2004. Produced by KPMG Information Risk Management for the Dutch Ministry of the Interior and Kingdom Relations. http://www.dutchdpa.nl/downloads_overig/PET_whitebook.pdf.

Leon, Pedro G., Blase Ur, Rebecca Balebako, Lorrie Faith Cranor, Richard Shay, and Yang Wang, "*Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising",* White Paper from the CyLab Carnegie Mellon University, October 31, 2011. Located at: http://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab11017.pdf

Mathews, Kristen J., ***Proskauer on Privacy: A Guide to Privacy and Data Security Law in the Information Age.*** Proskauer Rose LLP, 2006-2012

NISO Press, ***Understanding Metadata***, National Information Standards Organization, 2004, ISBN: 1-880124-62-9, located at: http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

Ozmen-Ertekin, Dilruba (Hofstra University) and Kaan Ozbay (Rutgers University), "*Dynamic Data Maintenance for Quality Data, Quality Research*". TRB Paper No. 11-3093 and Ozmen-Ertekin, Dilruba (Hofstra University),Kaan Ozbay (Rutgers University) and Cynthia Chen (City College of New York), ***Final Report Improvements on NYMTC Data Products***, Final Report prepared for the Region 2: University Transportation Research Center, November 2011. Located at: http://www.utrc2.org/research/assets/147/Final-Rept-Data-NYMTC1.pdf

Ross, Ron and the Joint Task Force Transformation Initiative Interagency Working Group with representatives from the Civil, Defense, and Intelligence Communities , ***Recommended Security Controls for Federal Information Systems and Organizations***. National Institute of Standards and Technology, NIST Special Publication 800-53 Revision 3, August 2009 and updated errata May 2010. Located at: http://csrc.nist.gov/publications/nistpubs/800-53-Rev3/sp800-53-rev3-final_updated-errata_05-01-2010.pdf

Roth, Craig. ***Website Governance: A How-to Guide.*** Web published by the Burton Group (now part of Gartner Research).

Sarsfield, Steve (2009). ***The Data Governance Imperative***, IT Governance Publishing, April 2009**.**

Scott, John, David A. Wheeler, Mark Lucas, and J.C. Herz, ***Open Technology Development: Lessons Learned and Best Practices for Military Software, Report for the Department of Defense***, May 2011. Located at: http://mil-oss.org/resources/otd-lessons-learned-military-v1.pdf
Sachdev, Tuschar. "*Defining a Governance Model for Portals*". White Paper for InfoSys, 2011. Located at: http://www.infosys.com/consulting/systems-integration/white-papers/documents/portals-governance-model.pdf.

Sloan, Suzanne, Alan Chachich, Ingrid Bartinique, and Linda Sharpe. ***Policy and Institutional Issues Analysis for the Dynamic Mobility Applications (DMA) Open Source Application Development Portal (OSADP)***, Report for the U.S. Department of Transportation, July 2012. FHWA-JPO-12-031 provides a comprehensive discussion about access controls and policies.

Sloan, Suzanne, Ingrid Bartinique, Josh Hassol, Deirdre Herring, Amy Sheridan, and Dicky Waldron. ***Identification of Critical Policy Issues for the Mobility Program***, White paper for the US DOT, June 2012. Publication Number: FHWA-JPO-12-035.

Tereschuk, George B., "*Government Purpose Rights in Technical Data and Computer Software in DOD Acquisition".* White paper located at: www.wifcon.com/anal/GPR_TD.doc.
Thomas, Gwen, ***The DGI Data Governance Framework***. White paper published by the Data Governance Institute and located at: http://www.datagovernance.com/dgi_framework.pdf.

Tysver, Dan of Beck and Tysver, ***Database Legal Protection***. Web published at: http://www.bitlaw.com/copyright/database.html

U.S. Department of Transportation, ITS Joint Program Office, "*Real-Time Data Capture and Management Program: Transforming the Federal Role*". May 2010. White paper posted at http://www.its.dot.gov/data_capture/datacapture_management_federalrole7.htm.

U.S. Department of Transportation, ITS Joint Program Office, "*Real-Time Data Capture and Management Program Vision: Objectives, Core Concepts and Projected Outcomes*". April 2010. White paper posted at: http://www.its.dot.gov/data_capture/datacapture_management_vision1.htm.

U.S. Department of Transportation, ITS Joint Program Office, "*Dynamic Mobility Applications Program Vision: Objectives, Core Concepts and Projected Outcomes*". April 2010. White paper posted at: http://www.its.dot.gov/dma/dma_vision2.htm

van Blarkom, G.W., Borking, J.J., Olk, J.G.E., ***PET: Handbook of Privacy and Privacy-Enhancing Technologies (The Case of Intelligent Software Agents)***. Web published by College bescherming persoonsgegevens, 2003. ISBN 90-74087-33-7. Located at: http://www.andrewpatrick.ca/pisa/handbook/handbook.html.

Vandervalk, Anita and Dena Snyder (Cambridge Systematics, Inc.), ***Real-Time Data Capture and Management Evaluation and Performance Measures -Evaluation Framework***. U.S. Department of Transportation, Research and Innovative Technologies Administration, September 6, 2011, Report No. FHWA-JPO-11-136, at: https://one.dot.gov/rita/ProgOffs/ITS/docworkspace/Mobility%20Initiative%20DL/DCM%20Program/Track%2002A%20Research%20and%20Development/Data%20Business%20Plan/Data%20Business%20Plan%20APPROVED%20EDITION-Evaluation%20Framework%20Report%20Draft%20Final1Sept2011.pdf

Wire, Richard, ***Disposition of Federal Records: A Records Management Handbook***. National Archives and Records Administration (NARA), 2000. Web version located at: http://www.archives.gov/records-mgmt/pdf/dfr-2000.pdf.

U.S. Department of Transportation
ITS Joint Program Office-HOIT
1200 New Jersey Avenue, SE
Washington, DC 20590

Toll-Free "Help Line" 866-367-7487
www.its.dot.gov

FHWA-JPO-12-030

U.S. Department of Transportation
Federal Highway Administration
Research and Innovative Technology Administration