REPORT S2-L13A-RW-1

Designing the Archive for SHRP 2 Reliability and Reliability-Related Data

SHRP2 RELIABILITY RESEARCH

SHRP 2
STRATEGIC HIGHWAY RESEARCH PROGRAM
Accelerating solutions for highway safety, renewal, reliability, and capacity

TRANSPORTATION RESEARCH BOARD
OF THE NATIONAL ACADEMIES

TRANSPORTATION RESEARCH BOARD 2015 EXECUTIVE COMMITTEE*

OFFICERS

CHAIR: **Daniel Sperling,** *Professor of Civil Engineering and Environmental Science and Policy; Director, Institute of Transportation Studies, University of California, Davis*
VICE CHAIR: **James M. Crites,** *Executive Vice President of Operations, Dallas–Fort Worth International Airport, Texas*
EXECUTIVE DIRECTOR: **Neil J. Pedersen,** *Transportation Research Board*

MEMBERS

Victoria A. Arroyo, *Executive Director, Georgetown Climate Center; Assistant Dean, Centers and Institutes; and Professor and Director, Environmental Law Program, Georgetown University Law Center, Washington, D.C.*
Scott E. Bennett, *Director, Arkansas State Highway and Transportation Department, Little Rock*
Deborah H. Butler, *Executive Vice President, Planning, and CIO, Norfolk Southern Corporation, Norfolk, Virginia (Past Chair, 2013)*
Malcolm Dougherty, *Director, California Department of Transportation, Sacramento*
A. Stewart Fotheringham, *Professor, School of Geographical Sciences and Urban Planning, University of Arizona, Tempe*
John S. Halikowski, *Director, Arizona Department of Transportation, Phoenix*
Michael W. Hancock, *Secretary, Kentucky Transportation Cabinet, Frankfort*
Susan Hanson, *Distinguished University Professor Emerita, School of Geography, Clark University, Worcester, Massachusetts*
Steve Heminger, *Executive Director, Metropolitan Transportation Commission, Oakland, California*
Chris T. Hendrickson, *Professor, Carnegie Mellon University, Pittsburgh, Pennsylvania*
Jeffrey D. Holt, *Managing Director, Bank of Montreal Capital Markets, and Chairman, Utah Transportation Commission, Huntsville, Utah*
Geraldine Knatz, *Professor, Sol Price School of Public Policy, Viterbi School of Engineering, University of Southern California, Los Angeles*
Michael P. Lewis, *Director, Rhode Island Department of Transportation, Providence*
Joan McDonald, *Commissioner, New York State Department of Transportation, Albany*
Abbas Mohaddes, *President and CEO, Iteris, Inc., Santa Ana, California*
Donald A. Osterberg, *Senior Vice President, Safety and Security, Schneider National, Inc., Green Bay, Wisconsin*
Sandra Rosenbloom, *Professor, University of Texas, Austin (Past Chair, 2012)*
Henry G. (Gerry) Schwartz, Jr., *Chairman (retired), Jacobs/Sverdrup Civil, Inc., St. Louis, Missouri*
Kumares C. Sinha, *Olson Distinguished Professor of Civil Engineering, Purdue University, West Lafayette, Indiana*
Kirk T. Steudle, *Director, Michigan Department of Transportation, Lansing (Past Chair, 2014)*
Gary C. Thomas, *President and Executive Director, Dallas Area Rapid Transit, Dallas, Texas*
Paul Trombino III, *Director, Iowa Department of Transportation, Ames*
Phillip A. Washington, *General Manager, Denver Regional Council of Governments, Denver, Colorado*

EX OFFICIO MEMBERS

Thomas P. Bostick *(Lt. General, U.S. Army), Chief of Engineers and Commanding General, U.S. Army Corps of Engineers, Washington, D.C.*
Timothy P. Butters, *Acting Administrator, Pipeline and Hazardous Materials Safety Administration, U.S. Department of Transportation*
Alison Jane Conway, *Assistant Professor, Department of Civil Engineering, City College of New York, New York, and Chair, TRB Young Members Council*
T. F. Scott Darling III, *Acting Administrator and Chief Counsel, Federal Motor Carrier Safety Administration, U.S. Department of Transportation*
Sarah Feinberg, *Acting Administrator, Federal Railroad Administration, U.S. Department of Transportation*
David J. Friedman, *Acting Administrator, National Highway Traffic Safety Administration, U.S. Department of Transportation*
LeRoy Gishi, *Chief, Division of Transportation, Bureau of Indian Affairs, U.S. Department of the Interior, Washington, D.C.*
John T. Gray II, *Senior Vice President, Policy and Economics, Association of American Railroads, Washington, D.C.*
Michael P. Huerta, *Administrator, Federal Aviation Administration, U.S. Department of Transportation*
Paul N. Jaenichen, Sr., *Administrator, Maritime Administration, U.S. Department of Transportation*
Therese W. McMillan, *Acting Administrator, Federal Transit Administration, U.S. Department of Transportation*
Michael P. Melaniphy, *President and CEO, American Public Transportation Association, Washington, D.C.*
Gregory G. Nadeau, *Acting Administrator, Federal Highway Administration, U.S. Department of Transportation*
Peter M. Rogoff, *Acting Under Secretary for Transportation Policy, Office of the Secretary, U.S. Department of Transportation*
Mark R. Rosekind, *Administrator, National Highway Traffic Safety Administration, U.S. Department of Transportation*
Craig A. Rutland, *U.S. Air Force Pavement Engineer, Air Force Civil Engineer Center, Tyndall Air Force Base, Florida*
Barry R. Wallerstein, *Executive Officer, South Coast Air Quality Management District, Diamond Bar, California*
Gregory D. Winfree, *Assistant Secretary for Research and Technology, Office of the Secretary, U.S. Department of Transportation*
Frederick G. (Bud) Wright, *Executive Director, American Association of State Highway and Transportation Officials, Washington, D.C.*
Paul F. Zukunft, *Adm., U.S. Coast Guard, Commandant, U.S. Coast Guard, U.S. Department of Homeland Security*

*Membership as of February 2015.

The Second
STRATEGIC HIGHWAY RESEARCH PROGRAM

Designing the Archive for SHRP 2 Reliability and Reliability-Related Data

Robert Hranac
Iteris, Inc.
Berkeley, California

Jorge A. Barrios
Kittelson & Associates, Inc.
Oakland, California

TRANSPORTATION RESEARCH BOARD

WASHINGTON, D.C.
2015
www.TRB.org

Subject Areas

Highways
Data and Information Technology
Operations and Traffic Management
Safety and Human Factors

The Second Strategic Highway Research Program

America's highway system is critical to meeting the mobility and economic needs of local communities, regions, and the nation. Developments in research and technology—such as advanced materials, communications technology, new data collection technologies, and human factors science—offer a new opportunity to improve the safety and reliability of this important national resource. Breakthrough resolution of significant transportation problems, however, requires concentrated resources over a short time frame. Reflecting this need, the second Strategic Highway Research Program (SHRP 2) has an intense, large-scale focus, integrates multiple fields of research and technology, and is fundamentally different from the broad, mission-oriented, discipline-based research programs that have been the mainstay of the highway research industry for half a century.

The need for SHRP 2 was identified in *TRB Special Report 260: Strategic Highway Research: Saving Lives, Reducing Congestion, Improving Quality of Life,* published in 2001 and based on a study sponsored by Congress through the Transportation Equity Act for the 21st Century (TEA-21). SHRP 2, modeled after the first Strategic Highway Research Program, is a focused, time-constrained, management-driven program designed to complement existing highway research programs. SHRP 2 focuses on applied research in four areas: Safety, to prevent or reduce the severity of highway crashes by understanding driver behavior; Renewal, to address the aging infrastructure through rapid design and construction methods that cause minimal disruptions and produce lasting facilities; Reliability, to reduce congestion through incident reduction, management, response, and mitigation; and Capacity, to integrate mobility, economic, environmental, and community needs in the planning and designing of new transportation capacity.

SHRP 2 was authorized in August 2005 as part of the Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU). The program is managed by the Transportation Research Board (TRB) on behalf of the National Research Council (NRC). SHRP 2 is conducted under a memorandum of understanding among the American Association of State Highway and Transportation Officials (AASHTO), the Federal Highway Administration (FHWA), and the National Academy of Sciences, parent organization of TRB and NRC. The program provides for competitive, merit-based selection of research contractors; independent research project oversight; and dissemination of research results.

SHRP 2 Report S2-L13A-RW-1

ISBN: 978-0-309-27431-9

Copyright Information

Note: SHRP 2 report numbers convey the program, focus area, project number, and publication format. Report numbers ending in "w" are published as web documents only.

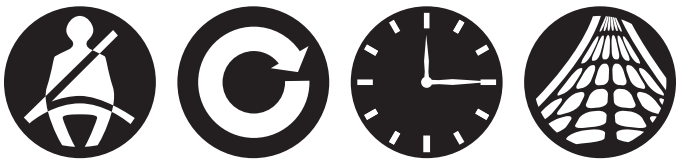
Notice

The project that is the subject of this report was a part of the second Strategic Highway Research Program, conducted by the Transportation Research Board with the approval of the Governing Board of the National Research Council.

The members of the technical committee selected to monitor this project and review this report were chosen for their special competencies and with regard for appropriate balance. The report was reviewed by the technical committee and accepted for publication according to procedures established and overseen by the Transportation Research Board and approved by the Governing Board of the National Research Council.

The opinions and conclusions expressed or implied in this report are those of the researchers who performed the research and are not necessarily those of the Transportation Research Board, the National Research Council, or the program sponsors.

The Transportation Research Board of the National Academies, the National Research Council, and the sponsors of the second Strategic Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of the report.



SHRP 2 Reports

Available by subscription and through the TRB online bookstore:

www.mytrb.org/store

Contact the TRB Business Office:
202-334-3213

More information about SHRP 2:

www.TRB.org/SHRP2

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. C. D. (Dan) Mote, Jr., is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Victor J. Dzau is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. C. D. (Dan) Mote, Jr., are chair and vice chair, respectively, of the National Research Council.

The **Transportation Research Board** is one of six major divisions of the National Research Council. The mission of the Transportation Research Board is to provide leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation. **www.TRB.org**

www.national-academies.org

SHRP 2 STAFF

Ann M. Brach, *Director*
Stephen J. Andrle, *Deputy Director*
Cynthia Allen, *Editor*
Kenneth Campbell, *Chief Program Officer, Safety*
Jared Cazel, *Editorial Assistant*
JoAnn Coleman, *Senior Program Assistant, Capacity and Reliability*
Eduardo Cusicanqui, *Financial Officer*
Richard Deering, *Special Consultant, Safety Data Phase 1 Planning*
Shantia Douglas, *Senior Financial Assistant*
Charles Fay, *Senior Program Officer, Safety*
Carol Ford, *Senior Program Assistant, Renewal and Safety*
James Hedlund, *Special Consultant, Safety Coordination*
Alyssa Hernandez, *Reports Coordinator*
Ralph Hessian, *Special Consultant, Capacity and Reliability*
Andy Horosko, *Special Consultant, Safety Field Data Collection*
William Hyman, *Senior Program Officer, Reliability*
Linda Mason, *Communications Officer*
David Plazak, *Senior Program Officer, Capacity and Reliability*
Rachel Taylor, *Senior Editorial Assistant*
Dean Trackman, *Managing Editor*
Connie Woldu, *Administrative Coordinator*

ACKNOWLEDGMENTS

This work was sponsored by the Federal Highway Administration in cooperation with the American Association of State Highway and Transportation Officials. It was conducted in the second Strategic Highway Research Program, which is administered by the Transportation Research Board of the National Academies. This project was managed by William Hyman, Senior Program Officer for SHRP 2 Reliability.

The research reported on herein was performed by Iteris, Inc., supported by Kittelson & Associates, Inc. Robert Hranac, Ali Mortazavi, and Karl Petty of Iteris, Inc., were the principal investigators. The authors acknowledge the contributions to this research and final report from Michael Darter, Sandra Lennie, Leon Raykin, Kavya Sambana, Michael Walls, Dustin Salentiny, Ken Yang, and Alex Kurzhanskiy of Iteris, Inc., and Jorge Andres Barrios, Alex Skabardonis, Senanu Ashiabor, Rick Dowling, and Daphne Dethier of Kittelson & Associates, Inc.

The L13A project team would like to acknowledge all the principal investigators of the SHRP 2 Reliability projects, the members of the SHRP 2 Reliability Technical Coordinating Committee, and the L13A Technical Expert Task Group.

FOREWORD

William Hyman, *SHRP 2 Senior Program Officer, Reliability*

The Reliability Technical Coordinating Committee recognized early in the formation of the Reliability research program that there would be a significant benefit to the research community, traffic engineers, planners, data managers, and others if the data from all the Reliability-related research projects could be preserved and easily accessed over the next 25 years or more. The inspiration for a repository was the Long Term Pavement Performance monitoring system from SHRP 1.

Because it was not clear how to proceed, the TCC determined that first there should be a feasibility study, and if that study found that it was both possible and desirable to develop the Reliability Archive, the Archive would be built. At the outset, the TCC also set aside resources to provide support to enable contractors to populate the Archive with data. The feasibility study strongly suggested that developing the Archive was both desirable and feasible. The study called for the use of open source software and determined that both the software to realize the functional capability of the Archive and the data should be stored in the cloud. The TCC recommended that the Archive be developed.

Although the priority from the outset was to make data sets used in the SHRP 2 Reliability research available for many decades, it quickly became apparent that the data would not be understandable or useful without sufficient contextual information. Thus, it was determined that all types of data should go into the Archive, both structured and unstructured. Structured data include comma-separated data, other flat files, and relational data. Unstructured data include data dictionaries, reports, presentations, video, spreadsheets, computer code, and other digital objects. In the parlance of the Archive, structured data are called "Data sets" and unstructured data, "Non–data sets."

The Archive's home page shows the five main use cases: upload, search, visualize, download, and discuss. Also shown on the home page are different statistics about the Archive and the latest four artifacts that have been uploaded so a user can drill down into the most recent additions. "Artifact" is the term used for each data set and non–data set in the Archive.

There are three ways to view the data once in the Archive: (1) a word search; (2) a search of the Archive that shows data sets by geographical location on a map of the country; and (3) a search by SHRP 2 focus area and project listing. The word search is a simple text search. A "search of the Archive" produces a literal spiral of data sets at each location; the user can click on each one and drill down to explore what the data are. A user can also select whether to look at a data set or non–data set and provide further filters for the selection. For example, a user can filter a data set according to whether it has speed, occupancy, and flow data. Finally, a user can search by project by first clicking on a focus area, then clicking a project listed under that focus area. An informative description accompanies each project to acquaint the user with the project objectives, key considerations in undertaking the research, and some of the important research products produced. Every project has metadata, and the metadata for data sets includes a data dictionary.

The Archive has substantial visualization capability for a user to preview and evaluate whether a data set is of interest. This feature is only for data sets, not non–data sets. This capability offers three different types of visualization. A user can visualize the first 300 records of a data set. On a map of highway facilities, reminiscent of a GPS navigation map, a user can see the precise location of the traffic detectors used to collect the data. Finally, a user can graph the relationships between different numerical fields in the data set. An example might be a scatter plot of the relationship between speed and flow.

The Archive was developed with the flexibility to be a dynamic and living system so that research that is new or related to the original Reliability research could be added to the system. There are significant administrative, maintenance, and operations costs to this capability. As of the end of the Archive contract, only SHRP 2 Reliability-related data had been entered into the system.

CONTENTS

Executive Summary

As part of Project L13A, the research team successfully developed, tested, and released a web-based, interactive archive system to store data and information from Reliability and Reliability-related projects from the second Strategic Highway Research Program (SHRP 2): http://www.shrp2archive.org.

The SHRP 2 Reliability Archive is designed to provide an open and accessible data hub. This hub creates a foundation that encourages additional transportation data research. Without the SHRP 2 Archive system, valuable data artifacts—including measured traffic data by various types of sensing equipment, structured analytic databases, spreadsheets, research reports, and other valuable transportation data—would be stored in scattered locations in forms that might not be publicly accessible. The SHRP 2 Archive resolves the scattered nature of data by creating a hub of data storage that is organized with built-in visualization tools for online analysis as well as downloadable structured data sets to support further research. By storing data in a consolidated and public manner, SHRP 2 has increased the likelihood that future researchers will discover data, thus increasing the value of the data and decreasing the cost of data acquisition for the collective transportation community. This structure also saves valuable resources for additional research by reducing the acquisition costs of time and funding that are associated with additional data collection.

As a part of the initial preparatory analysis, the L13A project team reviewed the findings of the SHRP 2 prototype L13 project report and past work on data archiving systems and technologies. This analysis resulted in the selection of WordPress as the core content management system (CMS), on which the Archive system was built. The selection of WordPress was based on its simplicity, flexibility, and extensibility.

The development of the Archive itself was an iterative and two-sided process involving the stakeholders or users on one side and the developers or the project team on the other side. An agile software approach was used to develop the Archive; using the methodology allowed for needs and requirements to evolve through collaboration between both sides. Through a series of phases, the L13A project team established a collaborative effort in which a subject matter expert (SME) group—consisting of the Technical Expert Task Group (TETG) members, future users, and experts—worked together to help the software design team identify practical user requirements and design a system that can be operational for more than 25 years.

The main steps for developing the Archive consisted of the following:

- Identify user needs, in collaboration with key stakeholders and potential users of the Archive;
- Design, test, and deploy a cost-effective, interactive, scalable, and robust archive system to store artifacts from SHRP 2 Reliability or Reliability-related projects;
- Develop processes and procedures to collect, upload, and use artifacts;

- Define a guideline to prepare artifacts for upload;
- Discuss challenges and issues regarding operations and maintenance of the Archive; and
- Provide technical information on the design and architecture of the Archive system.

The Archive is now operating on the Amazon cloud service (http://www.shrp2archive.org). The system has been developed based on open source web technologies. The system is expected to host more than 500 artifacts collected from more than 30 Reliability-related projects. The approximate size of the SHRP 2 Archive is anticipated to reach about 1 TB. A detailed Help section describing system features and providing a step-by-step guide on how to use the system is available online under the Archive Help pages.

As mentioned earlier, the SHRP 2 Archive is more than a data repository. It is a system that enables uploading of artifacts, searching with results arranged by list or map, and bulk and subset downloading of artifacts. Also, the Archive is a user-friendly toolset that facilitates visualizations of user-selected data and collaborations between multiple researchers.

A summary of the SHRP 2 Archive system's functionalities is provided below and shown in Figure ES.1:

- *Upload.* The SHRP 2 Reliability Archive's ingestion wizard allows Reliability project leaders to upload artifacts along with their related metadata and data dictionaries. The system categorizes the artifacts into two general groups: data sets and non–data sets (e.g., documents, computer codes, video, pictures). Artifacts submitted under each category need to meet certain requirements. Therefore, the producer of an artifact has to preprocess it before submitting it to the system.
- *Search and download.* The Archive provides faceted and text search tools to help users look for artifacts. The text search feature enables the users to conduct content searches within artifacts and metadata. In addition, the faceted search tool allows users to explore the Archive's repository. Users can filter the search results by selecting various related criteria. The system provides the search results on a map and in a list. Users can also search by project within a focus area.
- *Map visualization.* The system has the capability to map the geolocation information of the traffic detectors provided in a data set. Thus, the geofilter and map view capabilities can help users explore sensor locations.
- *Data visualization.* Archive users can explore and filter the content of a data set. They can make various two-dimensional (2-D) plots (e.g., lines, points, bar, and column) of any fields in the data set or a subset of a data set. The system also has the capability to plot different series on a common plot. Furthermore, users can save and print the plots for future use.

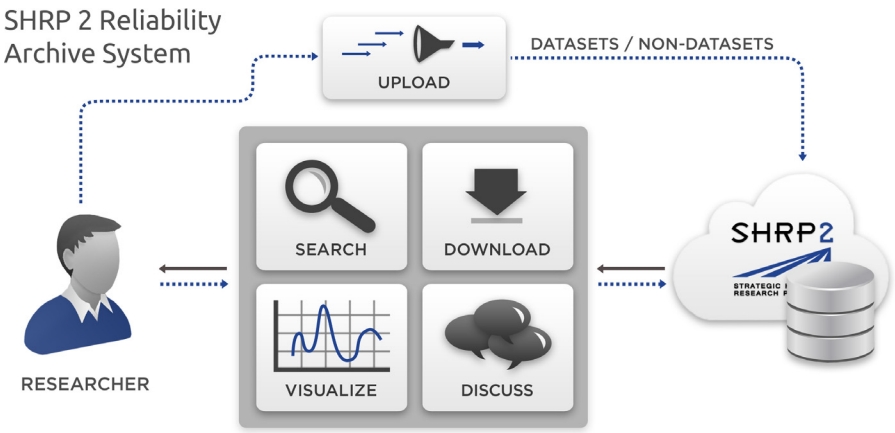


Figure ES.1. SHRP 2 Reliability Archive system.

- *Collaboration.* Registered Archive users can comment on project pages and artifact pages. They also can rate projects and artifacts. To comment on Archive pages and rate artifacts, users have to be registered.
- *Administration and user profile.* The Archive administrator is capable of granting access to users as well as adding, editing, and deleting site content. Through the administrator interface, the Archive administrator can also monitor the artifacts being ingested and moderate the comments being submitted by users.

The project team hopes the Archive developed out of this study will contribute to extending the state of the art in the use and development of data repositories in the transportation community.

CHAPTER 1

Background

1.1 Objective

The second Strategic Highway Research Program (SHRP 2) Reliability focus area has commissioned roughly 30 research projects costing approximately $27 million; together these projects were designed to lay the foundation for understanding and improving travel time reliability. In addition to these, seven other projects in the Capacity, Renewal, and Safety focus areas address travel time reliability. The SHRP 2 Reliability research and other Reliability-related projects have produced a large collection of raw data sets, analysis results, and documentation (called "artifacts"). Archiving these artifacts is pivotal to provide a foundation for future travel time reliability research efforts.

The objective of the L13A project is to ensure that this rich reliability data and information is available to the transportation research community via an archiving platform (called "the Archive") that is accessible through the Internet. The Archive will support the continued momentum of travel time reliability research efforts initiated by SHRP 2 for the coming decades.

1.2 Target Audience

The primary audience for the Archive is the following:

- University faculty, staff, and students in civil engineering, transportation planning, and logistics/supply chain management who conduct research on travel time reliability and closely related topics;
- Researchers from private consulting firms and other private enterprises involved in analyzing and modeling travel time reliability and closely related topics; and
- Analysts, traffic engineers, planners, and managers from road authorities interested in applying the findings to achieve improvements in their road networks.

While the primary audience is the transportation research community, the Archive is publicly accessible. Any member of the public or any organization may view and download the artifacts. Further, any member of the public or any organization may create a user account to participate in discussions about the research.

1.3 Benefits

The SHRP 2 Reliability Archive offers the following benefits:

- Complete access to all data, analysis results, documentation, and supporting information (nonsensitive information only) for SHRP 2 Reliability research, enabling users to understand how the original researchers came to their conclusions, replicate or validate their findings, and extend their research;
- Features for visualizing data sets in a grid, on a graph, or plotted on a map, enabling users to quickly form a mental model of a data set through visualization;
- Simple navigation and quick access to traffic data sets for the user's research project;
- Options for customizing the available traffic data sets to meet the user's needs and for downloading only the information that the user desires; and
- Quick registration process, enabling a user to create a user account and log in within minutes.

1.4 Summary of Archived Projects and Artifacts

The Archive includes a handful of projects from other SHRP 2 focus areas that also address travel time reliability. Nonsensitive data are provided so that they can be accessed and downloaded in an open manner. It should be noted that the Archive excludes data sets that contain personally identifiable information (PII).

The Archive includes more than 35 Reliability and Reliability-related projects. At the time of this writing, a total of 526 Reliability-related artifacts have been stored or identified for storage in the Archive. This set of artifacts includes 128 data sets and 398 non–data sets. (For more detailed information regarding the definition of data sets and non–data sets, see Section 5.1.) Table 1.1 and Table 1.2 provide a summary of archived/to-be-archived projects and artifacts.

1.5 Document Organization

Following this background chapter, the remainder of this final report is structured as follows:

- Chapter 2, Approach, describes the technical approach to developing the Archive;
- Chapter 3, Preparatory Analysis, presents the results of the project team's literature review and preparatory analysis;

Table 1.1. List of Archived and To-Be-Archived Projects

Number	Title
Reliability Focus Area	
L01	Integrating Business Processes to Improve Reliability
L02	Establishing Monitoring Programs for Mobility and Travel Time Reliability
L03	Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies
L04	Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools
L05	Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes
L06	Institutional Architectures to Advance Operational Strategies
L07	Evaluation of Cost-Effectiveness of Highway Design Features
L08	Incorporation of Travel Time Reliability into the *Highway Capacity Manual*
L09	Incorporation of Nonrecurrent Congestion Factors into the AASHTO Policy on Geometric Design
L10	Feasibility of Using In-Vehicle Video Data to Explore How to Modify Driver Behavior that Causes Nonrecurring Congestion
L11	Evaluating Alternative Operations Strategies to Improve Travel Time Reliability
L12	Improving Traffic Incident Scene Management
L13	Archive for Reliability and Related Data
L13A	Design and Implement a System for Archiving and Disseminating Data from SHRP 2 Reliabilities and Related Studies/Assistance to Contractors to Archive Their Data for Reliability Projects
L14	Traveler Information and Travel Time Reliability
L15	Innovative IDEA Projects
L16	Assistance to Contractors to Archive Their Data for Reliability and Related Projects (Combined with L13A)
L17	A Framework for Improving Travel Time Reliability
L32B	e-Learning for Training Traffic Incident Responders and Managers
L32C	Post-Course Assessment and Reporting Tool for Trainers and TIM Responders Using the SHRP 2 Interdisciplinary Traffic Incident Management Curriculum
L33	Validation of Urban Freeway Models
L34	e-Tool for Business Processes to Improve Travel Time Reliability
L35	Local Methods for Modeling, Economic Evaluation, Justification, and Use of the Value of Travel Time Reliability in Transportation Decision Making
L36	Regional Operations Forums for Advancing Systems Operations, Management, and Reliability
L38	Pilot Testing of SHRP 2 Reliability Data and Analytical Products
L55	Reliability Implementation Support

(continued on next page)

Table 1.1. List of Archived and To-Be-Archived Projects (continued)

Number	Title
Capacity Focus Area	
C04	Improving Our Understanding of How Highway Congestion and Pricing Affect Travel Demand
C10A	Partnership to Develop an Integrated, Advanced Travel Demand Model and a Fine-Grained, Time-Sensitive Network: Jacksonville-Area Application
C10B	Partnership to Develop an Integrated, Advanced Travel Demand Model with Time-Sensitive Networks: Sacramento-Area Application
C11	Development of Improved Economic Analysis Tools Based on Recommendations from Project C03 (2)
Renewal Focus Area	
R11	Strategic Approaches at the Corridor and Network Levels to Minimize Disruption from the Renewal Process
Safety Focus Area	
S04A	Roadway Information Database Development and Technical Coordination and Quality Assurance of the Mobile Data Collection Project (S04B)

- Chapter 4, System and User Needs and Requirements, discusses the general user needs and system requirements on which the system was built;
- Chapter 5, Artifact Upload, explains processes to prepare and upload Artifacts into the Archive;
- Chapter 6, User Guide—Working with the Archive, describes how the Archive system works;
- Chapter 7, System High-Level Architecture, provides technical information on the system design and architecture;
- Chapter 8, Test Plan, presents the plan that was developed for testing the software; and
- Chapter 9, Notes on Operations and Maintenance of the Archive, discusses operations and maintenance–related challenges and reviews the outreach plans for promoting the Archive among the transportation community.

This report also contains the following two appendices:

- Appendix A, Data Dictionary Template, is a template to which users may refer for providing information regarding data sets; and
- Appendix B, Federal System Security Guidelines, summarizes the federal requirements for system security.

Table 1.2. Summary of Artifacts

Project	Document	Spreadsheet	Data Set	Code	Other	Total
L01	5	0	0	0	0	5
L02	15	1	5	0	0	21
L03	10	110	31	4	0	155
L04	9	0	2	3	0	14
L05	9	4	0	0	0	13
L06	5	0	0	0	0	5
L07	6	1	5	1	0	13
L08	12	9	1	2	0	24
L10	4	0	1	1	0	6
L11	6	1	2	0	0	9
L12	5	0	0	0	0	5
L13	5	0	0	0	0	5
L13A	8	0	0	0	0	8
L14	17	0	3	0	0	20
L15A	3	0	0	0	0	3
L15B	3	0	0	0	0	3
L15C	3	0	0	0	0	3
L15D	3	0	0	0	0	3
L16	0	0	0	0	0	0
L17	14	0	0	0	0	14
L32B[a]	0	0	0	0	0	0
L32C[a]	0	0	0	0	0	0
L33	10	1	35	13	4	63
L34	4	0	0	0	0	4
L35A	6	2	7	0	0	15
L35B	6	14	5	0	0	25
L38A	4	0	5	0	0	9
L38B	4	0	5	0	0	9
L38C	4	0	5	0	0	9
L38D	4	0	5	0	0	9
L36	6	0	0	0	0	6
L55	3	0	0	0	1	4
C04	8	0	6	0	0	14
C05	6	0	4	0	0	10
C10A[a]	0	0	0	0	0	0
C10B[a]	0	0	0	0	0	0
C11	6	1	0	0	0	7
R11	7	0	0	1	0	8
S04A	4	0	1	0	0	5
Total	**224**	**144**	**128**	**25**	**5**	**526**

[a] Data are not available at this time.

CHAPTER 2

Approach

2.1 General Approach

The Archive system has to address both long-term and short-term user needs. A system that meets short-term customer requirements but cannot adapt to future and long-term needs is doomed to a short life span. However, it is almost impossible to grasp all short-term and long-term user as well as system needs at the planning stage. The user requirements need to be identified gradually—through keeping the users in the loop—while the system is being developed incrementally.

The development of system capabilities is an iterative, two-sided process. On one side, customers and users need to define the system's capabilities that will provide value. On the other side, the project team has to incrementally develop the system and constantly ask questions about whether the system functionalities being developed are worth the value they deliver.

To meet the characteristics critical to the success of the L13A project, the project team selected the agile software development approach to develop the Archive. The agile software development approach is an iterative, feature-based delivery process and is founded on continuous communication with users. This method is based on incremental development, in which needs and requirements evolve through collaboration between the developers and users/stakeholders. Agile development is a time-boxed approach that requires adaptive planning and provides evolutionary development and delivery.

The plan was to develop the Archive system and transfer the final product to SHRP 2 through six phases. Figure 2.1 shows the proposed approach. The proposed phases were as follows.

2.1.1 Phase 1—Requirement Definition

The focus of this phase was on understanding the system and user basic requirements. The collection of the requirements was conducted through two parallel tracks. On the first track, the project team performed literature studies on various archive systems and technologies to become familiar with the latest progress and findings in the data archiving and content management domain. On the second track, the team developed a stack of preliminary requirements in the form of user stories that were discussed and verified in a workshop consisting of a group of technical experts. The revised set of requirements was the basis for developing a prototype of the Archive.

2.1.2 Phase 2—Prototyping and Acceptance Testing

The team developed an archive prototype in Phase 2 based on the user stories defined in Phase 1. This phase was crucial since it was an opportunity to appraise various approaches and to solicit potential users' feedback on effectiveness and usefulness of archive features. The team closely collaborated with a group of subject matter experts, carefully put together by the project team, and the Technical Expert Task Group (TETG) to answer key questions regarding the operations and maintenance of the system. This collaborative effort was extremely helpful for the development team to make key decisions regarding the system's features and specifications. At the end of Phase 2, the completed prototype was demonstrated to stakeholders (Decision Gate 1) to gain their approval for starting Phase 3 (Archive Development).

2.1.3 Phase 3—Archive Development

Phase 3 focused on delivering an operating archive system and finalizing user interface and back-end coding. The system components were tested according to a test plan developed by the team to fix the bugs and errors that could not be identified during the prototyping phase. The team started loading the system with an initial set of SHRP 2 artifacts. At the end of this phase the project team held a user acceptance workshop to obtain TETG members' approval to release the system for operation (Decision Gate 2).

2.1.4 Phase 4—Outreach and Training

The objectives of this phase were to perform outreach and training activities after release of the Archive and to promote

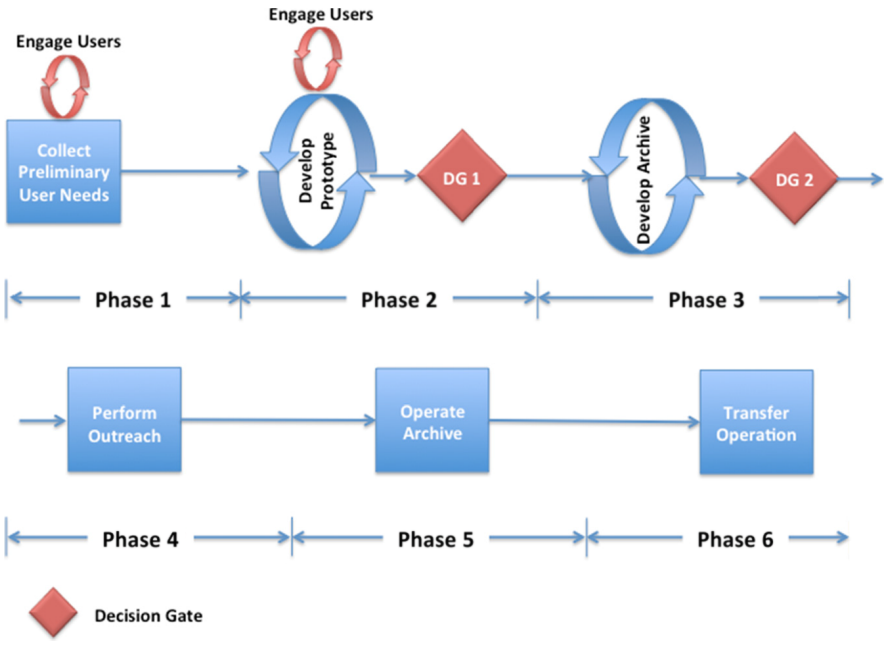


Figure 2.1. Project general approach.

the site through social media, academic conferences, or any transportation-related venue that discusses highway systems management and operations. Because of the significant workload associated with the artifact upload task, this phase was scaled down to save budget for the upload activities.

2.1.5 Phase 5—Operation and User Engagement

Phase 5 will deliver a fully functional data archive that is loaded with all SHRP 2 artifacts. This phase has not been started at the time of writing this report.

2.1.6 Phase 6—Transfer of Operations

During this phase the ownership of the Archive system will be transferred to SHRP 2. Extensive documentation on all the systems and frameworks will be provided and, ultimately, the project will be closed.

The Archive system was created through a collaborative effort between Iteris, Inc., and Kittelson & Associates, Inc. The Kittelson & Associates team assisted Iteris with the literature review, outreach, and data upload tasks.

2.2 Software Development Methodology

The L13A team used a hybrid method, which combined the prototyping and agile methodologies, to develop and deploy the Archive. The general development process was based on prototyping methodology in which the inception phase started with a prototype that was demonstrated to a group of users and subject matter experts to identify whether the design and features would address users' needs. At the software coding and system design level, the development was based on the agile approach.

2.2.1 Agile Approach

The team used agile schema to develop the software. The agile approach (see Figure 2.2) starts with creating a product backlog, a prioritized list of features that need to be developed for the product. The product backlog is a stack of stories that can be generated by the project team and other stakeholders. The project manager prioritizes the backlog and—with the help of the development team—breaks each story into smaller tasks. From that point on, each feature/story is coded iteratively—with the user/client in the loop—until it meets the desired outcomes.

The agile process is iterative and relies on continuous feedback from users/stakeholders during the course of development. To meet that requirement, the team met with a group of subject matter experts (SMEs) and users in various phases of the system development (i.e., requirement elicitation, prototype development, user acceptance) to make sure the final product met the needs of future users (see Figure 2.2). It should be noted that system and user requirements were not identified at the inception of the project, which made the communication task a crucial element to the success of the project. For more information on the user engagement strategy, see Section 2.3.

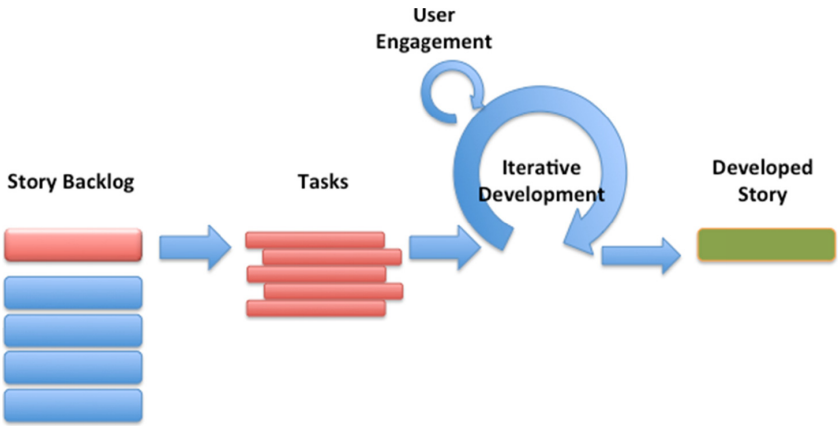


Figure 2.2. Agile approach.

2.3 User Engagement

As part of Phase 1, Phase 2, and Phase 3, the L13A project team established a collaborative effort in which the SME group—consisting of the TETG members, future users, and experts—worked together to help the software design team identify practical user requirements and design a system that can be operational for more than 25 years. This section sets out the general strategy for establishing the group and engaging members of the L13A SME group.

2.3.1 Objective of the L13A Subject Matter Experts Group

To develop an archive system for the SHRP 2 Reliability program, the L13A project team put together a core group of users and SMEs to get involved in identifying the system requirements and designing the Archive system. The goal was to make the Archive system more useful to the targeted audience. The project team assembled these stakeholders into a group, with a clear mission to help guide the development of the project.

Through effective stakeholder engagement, the team expected to

- Understand user and system needs;
- Use outside expertise and advice on system design and architecture;
- Obtain user feedback on the system requirements and interface; and
- Create awareness regarding the L13A project and the SHRP 2 Reliability Archive.

2.3.2 Mode of Engagement

The team leveraged two types of engagement tools to interact with the stakeholders: (1) facilitated workshop; and (2) stakeholder website.

2.3.2.1 Facilitated Workshop

The project team used the facilitated workshop approach to involve stakeholders. The Joint Application Design (JAD) method was used to elicit user and system requirements. JAD-like workshops provide various benefits for the users and developers. Some of the benefits are

- Reducing risk of scope creep;
- Accelerating delivery of product;
- Providing savings in time and effort; and
- Creating greater chance of consensus.

The group provided feedback and insight for the following deliverables (see Figure 2.3):

- User/system requirements and system design;
- Acceptance testing criteria;

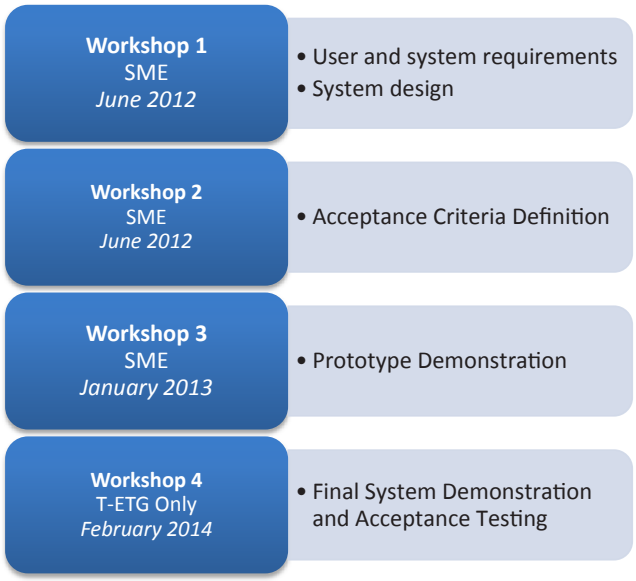


Figure 2.3. Stakeholder engagement: Dates and objectives.

- Archive system design review;
- Prototype; and
- Final system acceptance test.

2.3.2.2 Stakeholder Website

The stakeholder website is the SHRP 2 L13A project-restricted website that is available only to the SHRP 2 clients, the contractor team, and the TETG. The website can be accessed by those with proper permissions at http://sites.kittelson.com/SHRP2_ReliabilityDataArchive. Among other uses, the stakeholder website has been used to provide a link to a continuously updated spreadsheet with project progress. Figure 2.4 presents the home page of the stakeholder website for authorized users.

2.3.3 Group Structure

As mentioned before, the major objective of engaging the L13A SME group was to capture stakeholders' feedback on the Archive system design and features. The set of participants consisted of a representative sample of stakeholders and users that could help the project team develop a user-centric system successfully. The L13A SME group (see Figure 2.5) comprised

- The TETG group (see Table 2.1);
- An external advisory panel made up of
  - Senior researchers from major academic transportation centers,

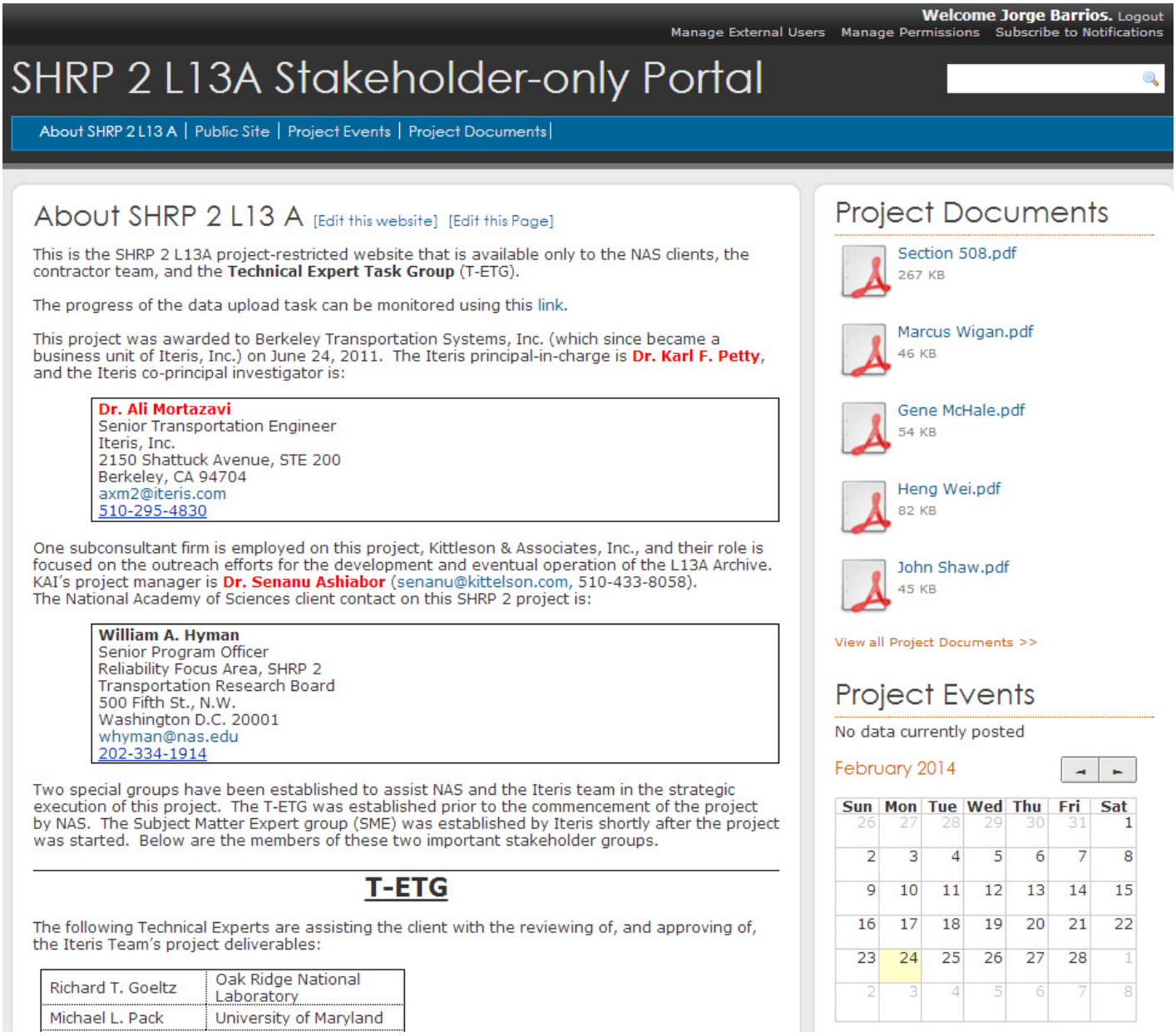


Figure 2.4. Stakeholder website.

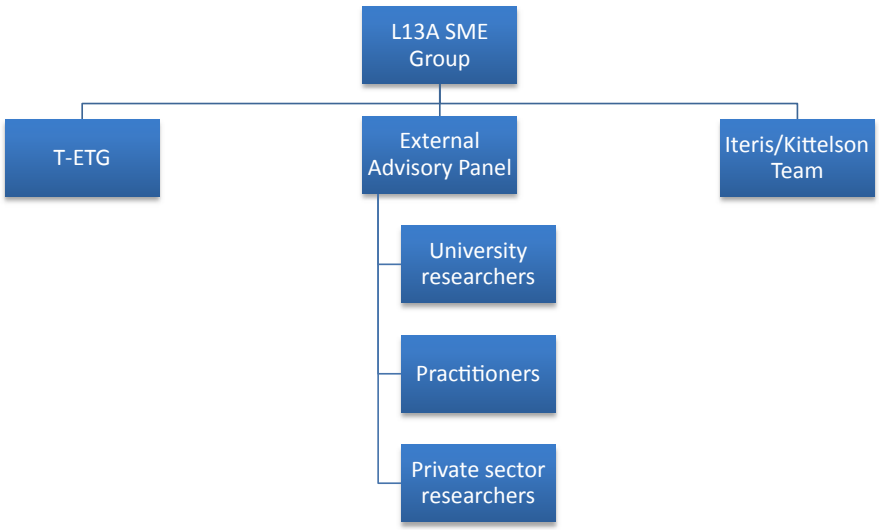


Figure 2.5. L13A subject matter expert group structure.

  - Practitioners from state and federal departments of transportation (DOTs), and
  - Private-sector researchers; and
- The Iteris and Kittelson project management team.

Members of the external advisory panel are listed in Table 2.2. Various characteristics were considered to assemble the group. Some of these characteristics are as follows:

- Being able to provide expertise in data archive systems, specifically those used for traffic data;
- Working with the project from the beginning, and staying engaged throughout the life of the system;
- Conducting research in topics related to SHRP 2 Reliability research fields;
- Being neutral; and
- Being potential users from different areas.

Table 2.1. Technical Expert Task Group Members

Name	Organization
Richard T. Goeltz	Oak Ridge National Laboratory
Michael L. Pack	Center for Advanced Transportation Technology (CATT) Lab, University of Maryland
William H. Schneider	The University of Akron
Dustin Sell	Microsoft
Theodore J. Trepanier	INRIX
Kristin A. Tufte	Portland State University
Marcus Ramsay Wigan	Oxford Systematics
Mike Bousliman	Montana Department of Transportation

Table 2.2. External Advisory Panel Members

Name	Organization
Brian Hoeft	Las Vegas Regional Transportation Commission of Southern Nevada's (RTC) Freeway & Arterial System of Transportation (FAST)
Dale Thompson	Federal Highway Administration (FHWA)
Daniela Bremmer	Washington State DOT
Gene McHale	FHWA
Heng Wei	University of Cincinnati
James Hall	University of Illinois
John Shaw	Wisconsin DOT
Ken Courage	University of Florida
Mark Hallenbeck	University of Washington
Mei Chen	University of Kentucky
Nazy Sobhi	FHWA
Paul Pisano	FHWA
Saman Moshafi	IndraSoft Inc.
Steven Beningo	FHWA
Steven Parker	University of Wisconsin
Walter During	FHWA

CHAPTER 3

Preparatory Analysis

The project team performed a thorough preparatory analysis task to get more familiar with the data archiving state-of-practice and understand the available content management technologies that can be used for L13A. This task included review of the L13 report (Section 3.1) as well as the past work on data archiving systems. This preparatory analysis was conducted to share the outcomes of the team's efforts on reviewing existing archived data user services (Section 3.2), online archive systems (Section 3.3), and commercially available archiving technologies (Section 3.4). The major objective of this analysis was to help the team with the following:

1. Come up with a preliminary system design; and
2. Identify an existing commercial off-the-shelf (COTS) content management system on which the Archive system could be built (Section 3.5.3).

3.1 Review of the L13 Report

3.1.1 Summary

The SHRP 2 L13 project, Requirements and Feasibility of a System for Archiving and Disseminating Data from SHRP 2 Reliability and Related Studies, was completed by Weris, Inc. between September 2008 and March 2010. The final report is available at http://onlinepubs.trb.org/onlinepubs/shrp2/SHRP2_S2-L13-RW-1.pdf (Tao et al. 2011).

The L13 (prototype) project report set out to identify the best way of meeting the three main goals of the Reliability Archive:

- Preserving the SHRP 2 digital assets for up to 50 years;
- Providing open access to transportation practitioners;
- Establishing a framework that can be used in other projects or for collaboration purposes.

Using those criteria, the SHRP 2 L13 research team focused on a version of an "active" archive system that could serve as a repository capable of managing files and metadata from different content sources. The aim was to preserve a diverse but related collection of digital artifacts and to make them accessible to practitioners and subsequent generations of researchers. The L13 research team proposed that the conceptual design pattern for the archival system follow that of a digital library or museum.

The research team assessed the technical, economic, and business aspects of the proposed archiving and dissemination system. This process was accomplished through interviews with the key stakeholders and a literature review of available and emerging technologies that might be applicable to the Archive. Based on this foundational work, the research team developed a vision for the Reliability Archive system that contained key high-level goals. The goals provided guiding principles for the development of a conceptual design and a detailed set of requirements for the Reliability Archive.

Starting from a conceptual design—based on their vision of a digital museum—the L13 authors created detailed system requirements and computed estimated life-cycle costs for three alternatives:

- An in-house File Transfer Protocol (FTP) web cluster;
- An in-house relational database;
- A commercial cloud-based system.

Of the three alternatives, the research team found that the commercial cloud-based system exhibited the lowest initial costs, the lowest recurring costs, the highest flexibility, and the best user accessibility. Given this finding, the team recommended a cloud storage system, which uses a pay-as-you-go, web-based access model.

The research team found that the in-house alternatives require significant up-front equipment purchase and installation that may be time-consuming and subject to bureaucratic delays.

3.1.2 Findings

3.1.2.1 SHRP 2 Management Perspective

One of the primary objectives of the Reliability Archive was to allow users to find and validate the research results from relevant SHRP 2 projects and to refine and build on research results in the future. Another primary objective of the Reliability Archive was to preserve research project data. In other words, there was agreement that the research conclusions need to be archived along with the data.

3.1.2.2 Project Contractor Perspective

The research team interviewed contractors of active Reliability projects and relevant capacity projects to help understand the data used and produced by these projects that would need to be archived.

3.1.2.3 Literature Research

As part of the L13 project, the research team conducted a literature review. A survey of the literature in the public domain revealed that the ability to archive digital resources—and the effectiveness of doing so—has grown considerably with the explosive growth of digital information.

The L13 report specifically discussed the Reference Model for an Open Archival Information System (OAIS) which has been adopted by the International Organization for Standardization. The OAIS model defines the major entities and functions of a digital repository. OAIS is a conceptual framework and does not prescribe any specific implementation on any level.

The OAIS paradigm has three general parts:

- Data ingestion—accepting digital objects into an archive with metadata in Metadata Encoding and Transmission Standard (METS) format;
- Data archive and management—storing, managing storage hierarchy, updating administrative and metadata, software and hardware maintenance; and
- Data access—locating, applying access controls, and generating responses.

3.1.2.4 Role and Importance of Metadata

The L13 report recognized METS as a suitable metadata standard for the Archive system. METS is an Extensible Markup Language (XML) schema that provides a mechanism for recording various relationships that occur between pieces of content and between the content and the metadata that make up a digital object. METS was specifically designed to act as an OAIS information package. Packaging the metadata with the digital object ensures that the object is self-documenting.

3.1.2.5 Conceptual Design for the Archival System

The research team's observations yielded the conclusion that the proposed archival system could not be thought of as a database, the structure of which is known up front. The team proposed that the conceptual design for the archival system follow that of a digital library.

The L13 report noted that the project teams would create initial Submission Information Packages (SIP) for conveyance to the archival system. Planning and preparation, for the eventual submission of SIPs to the Archive toward the end of each project, would need to commence early within each research project. This includes selecting the most preservation-friendly file formats and creating descriptive metadata. All aspects of copyright, privacy, and proprietary rights would need to be documented.

Once the necessary preaccessioning work has been performed, the six core archive functions of ingestion, data management, archival storage, access, administration, and preservation planning would be performed according to the OAIS model.

3.1.3 System Requirements Proposed by L13 Project

The research team noted that consumers are expected to be a worldwide community of transportation practitioners who would use the information directly, as well as researchers who would validate and build on the information base. In addition, the team expected consumers to interact with the archival system through a web-based portal.

According to the L13 report, all SHRP 2 Reliability projects would produce a range of document-centric files, such as reports and presentations, in various formats. Thus, there would be a need for a document management system. The use of COTS Enterprise Content Management (ECM) packages was discouraged in the L13 final report for various reasons.

The L13 report concluded that the Archive should be preserved for up to 50 years, even though the report only calculated the maintenance costs until 2035.

3.1.4 User Interfaces

The L13 report determined that the user interface (UI) would be based on four general user types:

- Managers of transportation agencies;
- Technical staff of transportation agencies;
- Nontransportation professionals; and
- Researchers and analysts.

Managers of transportation agencies are interested in business processes, strategies, institutional structures, and

performance measures. They need to quickly find the conclusions of each project, executive summaries, and presentations. The technical staff of transportation agencies are interested in various Reliability products, such as data sets, tools, and reports. They need to quickly find the end products of the projects, which may be organized and grouped by categories such as planning, design, and operations. Nontransportation professionals with some relationship to transportation, such as law enforcement, are interested in the end products related to operational strategies, incident management, and travel time reliability improvement. They need to quickly find project conclusions, results, and operational strategies. Researchers and analysts are interested in understanding transportation, conducting studies, and developing their own methods and technologies. Their focus is the interface to individual projects.

The L13 report proposed a UI consisting of the following pages:

- Home page;
- Navigation of Reliability research projects;
- Direct project lists;
- Reliability themes;
- Data set organization—including data set name, collection method, related projects, location, format, size, derived data, and research results;
- Grouping of research products—by the three general categories of planning, design, and operations; and
- Search—both simple and advanced, and navigation of project-level data and results.

3.1.5 Data Integrity and Quality

Data integrity and quality control were determined to be crucial for a successful archive. The L13 report identified three logical points of data quality control:

- Within individual Reliability projects;
- Through Reliability Project L13A (previously Reliability Project L16 for assistance in preparing data for submission to the Archive); and
- Through active enforcement of the preservation policy within the archival system.

When a Reliability project is ready to deliver its data to be archived, the project team would be expected to submit the data (and metadata) along with the project's quality control standards, methods, and assessment. The L13 report suggested Reliability Project L13A team would be responsible for reviewing the data quality assessment and would either confirm or modify the quality rating. The quality rating would be a metadata attribute that would be part of the metadata to be prepared and collected by individual projects, known as Preservation Description Information.

The research team identified two types of quality issues with project metadata. The first issue is that each project would most likely use and collect different metadata elements. The other issue is that some metadata information may be inaccurate or incomplete. Detailed metadata guidelines would need to be developed to define the mandatory metadata and the specifications for data quality. The team suggested that a quality control screen should be set up to assess the project metadata. Once the project metadata passes the data quality screen test, the metadata would be archived to the metadata repository in the L13A Archive.

3.1.6 Data Rights

The researchers found that generally there had to be few or no restrictions on the derived data from the Reliability projects. The raw data typically came from the contractor's existing data sets, a state DOT, or other transportation agency as well as the private sector. The report proposed that access to the data be protected with usage stipulations.

3.1.7 Institutional Framework and Governance

As with any large archive of information, the research team stated the necessity for a proven and reliable institutional framework to provide long-term stewardship of the Archive. The L13 research team documented some best practices of national systems and referred to the SHRP 2 implementation report (Committee for the Strategic Highway Research Program 2, 2009) for recommendations. One of the key recommendations of the SHRP 2 implementation report was to designate a principal implementation agent responsible for leading and supporting the SHRP 2 implementation. In addition, the recommendation was to have a similar role established for the Archive.

To support the principal implementation agent, the L13 research team recommended that a stakeholder advisory group provide strategic guidance and technical advice on the long-term stewardship and use of the Archive.

3.1.8 Technical Issues

The research team explored specific technical issues that were cited in the L13 Reliability Project request for proposal (RFP):

- Data normalization and denormalization;
- Online analytical processing (OLAP) and user-defined functions;

- Service-oriented architectures (SOA); and
- Virtualization.

3.1.8.1 Normalization and Denormalization

Normalization and denormalization are used to organize the data by efficient data storage and relationships (normalization), or optimization for quicker queries at the expense of duplicating data sets (denormalization). The research team determined that data normalization and denormalization do not have any application in the proposed Archive in terms of the postresearch part of the process of preparing data for preservation.

3.1.8.2 Online Analytical Processing and User-Defined Functions

The Archive's purpose is to serve the transportation community by preserving transportation project information and facilitating lookup, presentation, and downloading of such information. Therefore, it was the L13 research team's position that it is not within the scope of the archival system to perform analysis on the stored data, or to perform other open-ended or dynamic user-defined functions on the data.

3.1.8.3 Service-Oriented Architecture

Service-oriented architecture involves web-based services provided by a system that exposes their functionality. The report mentioned that SOA and web services could be used to deliver mashups and could also be expected to play other roles in the Reliability Archive.

3.1.8.4 Virtualization

Virtualization uses software to abstract a hardware environment. The virtualization software runs on a host operating system, allowing one or more guest operating systems to run on the same hardware platform. This application of virtualization was expected to play a role in the deployment of the Reliability Archive, particularly in terms of hosting application software involved in managing the repository or hosting software that provides user access to the repository. For storage, virtualization is used to abstract logical storage from physical storage. The research team found it likely that some form of storage virtualization would be used in the actual deployment of the proposed archival system.

3.1.9 Establishing Solution Alternatives

The research team mapped system requirements against potential solution building blocks and concluded that these requirements fell roughly into three blocks of functionality, connected via some kind of workflow as shown in Figure 3.1.

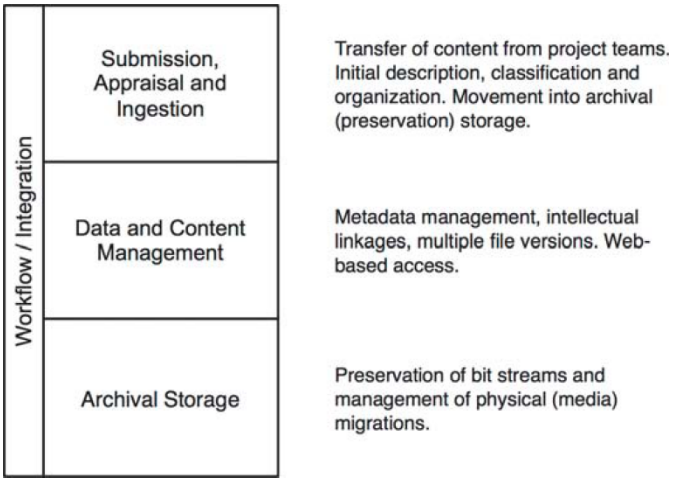


Figure 3.1. Functional blocks of proposed archival system.

The L13 research team identified and discussed two critical issues that would influence the selection of potential alternatives:

- The relative importance of certain system functionality over time; and
- The estimated total data volume to be preserved in the Archive.

3.1.10 Solution Components and Implementation Approaches

In coming up with solutions, the L13 research team considered a wide range of potential technology choices. The L13 team looked at commercial off-the-shelf (COTS) technology, standardized versus proprietary hardware, open source software (OSS), in-house developed software, hosting, and storage and software as a service (SaaS).

The L13 research team concluded that in-house software development should be considered only as a last resort and only for limited functionality for which the need is short-term. Based on the L13 report, community-supported OSS should also be considered only under similar circumstances because it generally requires developing significant in-house expertise to implement and support it. COTS software seemed to be the most attractive option for the application and infrastructure software portion of the system, eliminating the burden and issues that arise with self-support of either in-house developed software or community-supported OSS. The research team recommended that cloud storage be considered because the cost of acquiring and managing storage is likely the single largest cost of the system's lifetime.

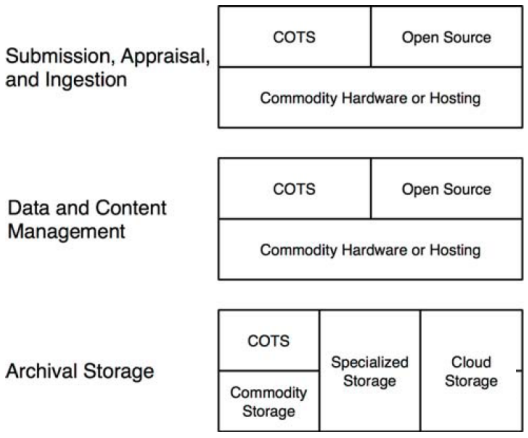


Figure 3.2. Solution framework.

The visioning and filtering process that the L13 research team went through led to the conceptual solution framework as shown in Figure 3.2.

Using this framework, the research team proposed a number of alternative system solutions, which are described next.

3.1.10.1 Alternative 1

This alternative is a bare minimum solution whose implementation is straight forward, but its capabilities are very limited: storing data in a file system. Its components are listed below and shown in Figure 3.3.

1. Research teams have password-protected access to a specific directory in which they build their project file tree.
2. Web cluster consists of FTP server for uploading data and Hypertext Transfer Protocol (HTTP) server for providing access to the data.
3. Archival storage is provided by self-hosted network-attached storage. Disk size per the network-attached storage is 16 TB.
4. Institution staff uses the Archivist Toolkit to catalog the files deposited into the storage.
5. User access to the Archive is provided through directory browsing in Windows Explorer fashion.

The L13 report concluded that this alternative was unattractive and could only be considered as the last resort.

3.1.10.2 Alternative 2

This alternative is based on digital object repository management software designed for libraries, museums, and archives. Known content management systems are listed here: http://en.wikipedia.org/wiki/List_of_content_management_systems.

The components of this alternative are listed next and shown in Figure 3.4.

1. Research teams submit the content into the repository via web interface that provides all the necessary forms and enforces access restrictions.
2. Review stage involves automatic, semiautomatic, and manual workflows resulting in editing, deleting, and approving the content before its ingestion into the repository.
3. The proposed Relational Database Management System (RDBMS) is Oracle. The idea is that the runtime database holding the content can be automatically built from the METS-formatted metadata. The web, application, and database cluster is a number of self-hosted commodity servers.
4. Digital objects themselves are stored in self-hosted archival class storage under write-once, read-only policy with

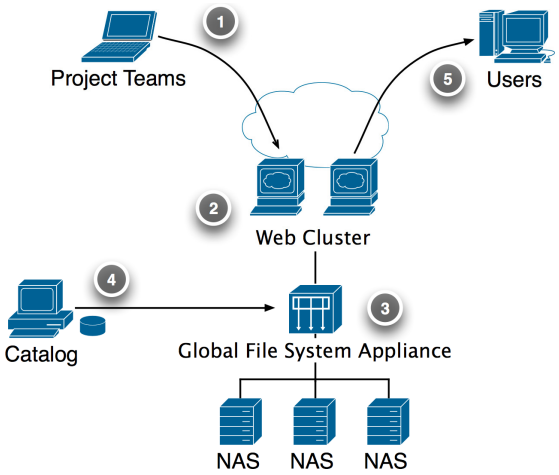


Figure 3.3. Alternative 1 concept.

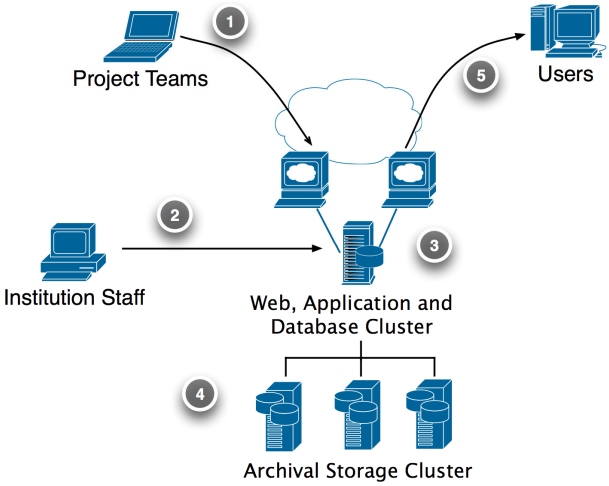


Figure 3.4. Alternative 2 concept.

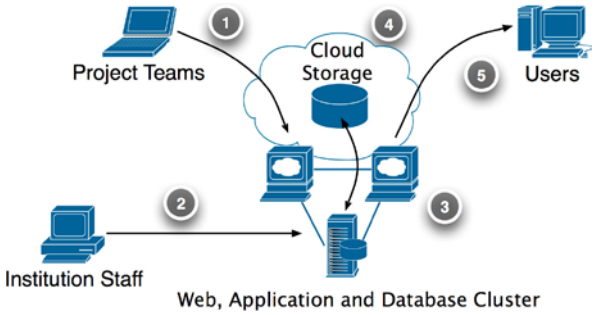


Figure 3.5. Alternative 3 concept.

object replication to ensure their security and integrity over time.

5. Researchers and practitioners access the repository through a web portal. Web publishing is automatic and driven by the repository metadata, the look and feel being customized by Extensible Stylesheet Language Transformations (XSLT) and Cascading Style Sheets (CSS). Users can navigate the repository through fixed and dynamic classification menus/paths and perform full-text and faceted searches.

3.1.10.3 Alternative 3

This alternative is almost the same as Alternative 2. The difference is that it is cloud-based. Items 1, 2, and 3 are the same as in Alternative 2. Digital objects are stored in the cloud. The UI is the same as in Alternative 2. This alternative was the solution promoted in the L13 final report. It was justified as a minimal cost alternative (equipment maintenance and system administration are outsourced). The alternative is shown in Figure 3.5.

3.1.11 Life-Cycle Costs Analysis

A section of the L13 report included the research team's estimates on the costs of each alternative archival system, while considering all the life-cycle costs that could be identified over a 25-year period. The life-cycle cost assumptions considered in the analysis included costs associated with initial acquisition, operations, and maintenance as well as periodic upgrades to accommodate technology advances and obsolescence.

The life-cycle costs of the three alternatives were summarized in the L13 report. Alternative 3 was the minimum cost alternative. The report estimated the cost of Alternative 3 at $5,530,132 over 25 years. The cost of implementation was estimated at $173,425 per year, and the duration was estimated to be 1½ years.

3.2 Archived Data User Services

The U.S. Department of Transportation included Archived Data User Services (ADUS) in the National Intelligent Transportation Systems (ITS) Architecture in 1999, envisioning "the unambiguous interchange and reuse of data and information throughout all functional areas" (FHWA 1998). ADUS requires that data from ITS systems be collected and archived for historical, secondary, and non–real-time uses, and that these data be made readily available to users.

This section reviews existing federal guidance on the development of ADUS systems and reviews transportation-related ADUS systems that have been developed in several states in the United States.

3.2.1 Introduction to Federal Highway Administration ADUS Guidelines

The FHWA funds and monitors many state ADUS programs. In the past 10 years, the FHWA has published a number of reports reviewing the progress of ADUS programs and summarizing the challenges of ADUS programs across the country (U.S. DOT 2003). The 2003 report identified the major functions of ADUS systems as

- Operational data control;
- Data import and verification;
- Automatic data historical archive, to store the data permanently;
- Data warehouse distribution, to provide data to the planning, safety, operations, and research communities; and
- ITS community interface.

A complete list of ADUS programming procedures and specifications has been compiled by Iteris, and is available online at http://itsarch.iteris.com/itsarch/html/user/usr71.htm.

3.2.2 FHWA ADUS Functions and Guidelines

Operational data control is extensively described in a report prepared by the Texas Transportation Institute (TTI) for the FHWA (Turner 2007). Data control—but most important, the resulting data quality—is an important aspect of ADUS systems, as users will likely disregard the validity of the entire system if they encounter erroneous data points. The TTI document provides data control guidelines to ensure data quality.

In the interest of promoting a unified approach to ADUS, the FHWA partnered with the ASTM International (formerly American Society for Testing and Materials) to devise national ADUS standards (ASTM 2011). The ASTM report focuses on the technical considerations of implementing an ADUS system, which is referred to as an Archived Data Management System (ADMS). ASTM developed 10 guiding principles, which it grouped on the basis of whether the focus is on (*a*) acquiring data, (*b*) managing the ADUS, or (*c*) retrieving data and serving information. Table 3.1 is an adaptation of these principles.

Table 3.1. Guiding Principles for ADMS Development

Acquiring Data	Managing the ADMS (ADUS)	Retrieving Data and Information
• Get archived data from other centers. • Integrate selected other transportation data, including roadside data collection.	• Manage the archive to account for data quality. • Provide security for the ADMS. • Specify and maintain metadata to support the ADMS. • Manage the interfaces of the archive data administrator. • Interact with other archives and monitor other standards.	• Process user requests for data. • Support analysis of the archived data. • Prepare data for government reporting systems.

Source: Adapted from ASTM (2011).

An April 1998 report to the FHWA's Office of Highway Policy Information is largely dedicated to ADUS's "institutional issues for implementation" (Margiotta 1998). Among the institutional issues, privacy concerns, liability, and training and outreach are the most relevant to the SHRP 2 L13 project. The Margiotta report describes ways to address these issues.

3.2.2.1 ADUS Transportation Research Board 2007 Workshop

An interesting review of the institutional issues described above was organized by the FHWA at the 2007 Transportation Research Board (TRB) annual meeting. Several presentations on ADUS implementation and the lessons learned from such implementations are described in Bertini (2007).

The workshop involved a discussion about issues with the use of ADUS systems and possible solutions. Table 3.2 provides a starting point for understanding the needs of transportation professionals by matching their needs to current

Table 3.2. Needs of ADUS Stakeholders

Stakeholder Group	Application	Method or Function	Collection and Use of	
			Current Data	ITS-Generated Data
Metropolitan planning organization (MPO) and state transportation planners	Congestion management systems	Congestion monitoring	Travel times collected by "floating cars": usually only a few runs (small samples) on selected routes. Speeds and travel times synthesized with analytic methods (e.g., *Highway Capacity Manual,* simulation) using limited traffic data (short counts). Effect of incidents missed completely with synthetic methods and minimally covered by floating cars.	Roadway surveillance data (e.g., loop detectors) provide continuous volume counts and speeds. Variability can be directly assessed. Probe vehicles provide same travel times as floating cars but greatly increase sample size and areawide coverage. The effect of incidents is embedded in surveillance data, and Incident Management Systems provide details on incident conditions.
	Long-range plan development	Travel demand forecasting (TDF) models	Short-duration traffic counts used for model validation. Origin–Destination (O-D) patterns from infrequent travel surveys used to calibrate trip distribution. Link speeds based on speed limits or functional class. Link capacities usually based on functional class.	Roadway surveillance data provide continuous volume counts, truck percentages, and speeds. Probe vehicles can be used to estimate O-D patterns without the need for a survey. The emerging TDF models [e.g., the Transportation Analysis and Simulation System (TRANSIMS)] will require detailed data on network (e.g., signal timing) that can be collected automatically via ITS. Other TDF formulations that account for variability in travel conditions can be calibrated against the continuous volume and speed data.

(continued on next page)

Table 3.2. Needs of ADUS Stakeholders (continued)

Stakeholder Group	Application	Method or Function	Collection and Use of	
			Current Data	ITS-Generated Data
MPO/state transportation planners (continued)	Corridor analysis	Traffic simulation models	Short-duration traffic counts and turning movements used as model inputs. Other input data to run the models collected through special efforts (signal timing). Very little performance data available for model calibration (e.g., incidents, speeds, delay).	Most input data can be collected automatically and models can be directly calibrated to actual conditions.
Traffic management operators	ITS technology	Program and technology evaluations	Extremely limited; special data collection efforts required.	Data from ITS provide the ability to evaluate the effectiveness of both ITS and non-ITS programs. For example, data from an incident management system can be used to determine changes in verification, response, and clearance times due to new technologies or institutional arrangements. Freeway surveillance data can be used to evaluate the effectiveness of ramp meters or high-occupancy vehicle restrictions.
		Predetermined control strategies	Short-duration traffic counts and floating car travel time runs. A limited set of predetermined control plans is usually developed, mostly due to the lack of data.	Continuous roadway surveillance data makes it possible to develop any number of predetermined control strategies.
		Predictive traffic flow algorithms	Extremely limited.	Analysis of historical data forms the basis of predictive algorithms: "What will traffic conditions be in the next 15 min?" (Bayesian approach).
Transit operators	Operations planning	Routing and scheduling	Manual travel demand and ridership surveys; special studies.	Electronic fare payment systems and automatic passenger counters allow continuous boardings to be collected. Computer-aided dispatch systems allow O-D patterns to be tracked. Automatic vehicle identification (AVI) on buses allows monitoring of schedule adherence and permits the accurate setting of schedules without field review.
Air quality analysts	Conformity determinations	Analysis with the MOBILE model	Areawide speed data taken from TDFs. Vehicle miles traveled (VMT) and vehicle classifications derived from short counts.	Roadway surveillance provides actual speeds, volumes, and truck mix by time of day. Modal emission models will require these data in even greater detail, and ITS is the only practical source.
MPO/state freight and intermodal planners	Port and intermodal facilities planning	Freight demand models	Data collected through rare special surveys or implied from national data (e.g., Commodity Flow Survey).	Electronic credentialing and AVI allow tracking of truck travel patterns, sometimes including cargo. Improved tracking of congestion through the use of roadway surveillance data leads to improved assessments of intermodal access.

(continued on next page)

Table 3.2. Needs of ADUS Stakeholders (continued)

Stakeholder Group	Application	Method or Function	Collection and Use of	
			Current Data	ITS-Generated Data
Safety planners and administrators	Safety management systems	Areawide safety monitoring; studies of highway and vehicle safety relationships	Exposure (typically VMT) derived from short-duration traffic and vehicle classification counts; traffic conditions under which crashes occurred must be inferred. Police investigations, the basis for most crash data sets, performed manually.	Roadway surveillance data provide continuous volume counts, truck percentages, and speeds, leading to improved exposure estimation and measurement of the actual traffic conditions for crash studies. ITS technologies also offer the possibility of automating field collection of crash data by police officers [e.g., Global Positioning System (GPS) for location].
Maintenance personnel	Pavement and bridge management	Historical and forecasted loadings	Volumes, vehicle classifications, and vehicle weights derived from short-duration counts (limited number of continuously operating sites).	Roadway surveillance data provide continuous volume counts, vehicle classifications, and vehicle weights, making more accurate loading data and growth forecasts available.
Commercial vehicle enforcement personnel	Enforcement of commercial vehicle regulations	Hazardous material inspections and emergency response	Extremely limited.	Electronic credentialing and AVI allow tracking of hazardous material flows, allowing better deployment of inspection and response personnel.
Emergency management services (local police, fire, and emergency medical)	Incident management	Emergency response	Extremely limited.	Electronic credentialing and AVI allow tracking of truck flows and high-incident locations, allowing better deployment of response personnel.
Transportation researchers	Model development	Travel behavior models	Mostly rely on infrequent and costly surveys: stated preference and some travel diary efforts (revealed preference).	Traveler response to system conditions can be measured through system detectors, probe vehicles, or monitoring in-vehicle and personal device use. Travel diaries can be embedded in these technologies as well.
		Traffic flow models	Detailed traffic data for model development must be collected through special efforts.	Roadway surveillance data provide continuous volume counts, densities, truck percentages, and speeds at very small time increments. GPS-instrumented vehicles can provide second-by-second performance characteristics for microscopic model development and validation.
Private-sector users	Truck routing and dispatching	Congestion monitoring	Current information on real-time or near real-time congestion is extremely limited.	Roadway surveillance data and probe vehicles can identify existing congestion and can be used to show historical patterns of congestion by time of day. Incident location and status can be directly relayed.
	Information service providers	Trip planning	Information on historical congestion patterns is extremely limited. This information could be used in developing pretrip route and mode choices, either alone or in combination with real-time data.	

Source: Adapted from Margiotta (1998).

practice and to equivalent solutions available from ADUS systems. The table was compiled by Margiotta and published in Margiotta (1998).

3.2.2.2 Summary of FHWA ADUS Guidelines

In summary, the FHWA has stressed the importance of addressing both the technical and institutional aspects of an ADUS system. The technical considerations have been widely studied and documented as a result of partnerships with TTI, ASTM, Iteris, and others. However, institutional concerns are not as well understood. For this reason, the FHWA has recently sponsored workshops, seminars, and research to exclusively deal with tailoring and promoting ADUS systems to transportation planners and engineers.

3.2.3 Existing ADUS Systems

This section presents a review of existing ADUS systems in the United States and other countries.

The purpose of the literature review was to guide the development of the Archive. Because of the prolonged development and data procurement period of the L13A Archive, the current versions of the ADUS systems below may be significantly different from their descriptions. Nevertheless, the literature review captures the features and concepts that were considered for the SHRP 2 L13A Archive.

3.2.3.1 PeMS, California

The California Department of Transportation (Caltrans) Performance Measurement System (PeMS) was established in the early 2000s with the help of University of California, Berkeley's Partners for Advanced Transportation Technology (PATH). The system was set up to process 30-sloop detector data from freeways across the entire California network. At the time PeMS was set up, it processed 2 GB of data per day (Choe et al. 2002).

The data are published in real time through a web interface and stored for historical analysis. Traffic volume, speed, and occupancy data for freeways are archived in PeMS. Travel time data of some freeways are collected through electronic toll-tag collectors. Data can be accessed by selecting the entire length of freeway or section of freeway. More recently the state has begun adding arterial roads to the PeMS system.

PeMS develops performance management information from fairly rudimentary and raw data (detector volumes and occupancies). Using the volumes and occupancies the PeMS system produces travel time estimates, time-space diagrams, count curves, and other graphic tools that can be used to understand and improve freeway operations.

The combination of both the input (volumes) and performance data (such as speed or VMT) enables the creation of contour and across-space plots that can aid in determining the location of bottlenecks. This can be done by comparing the occupancy and count curves of two nearby detectors. Whenever a bottleneck forms, occupancy spikes and starts a wave of increased occupancy that moves upstream to other detectors. PeMS contains algorithms to automatically identify, classify, and report bottlenecks to the graphical user interface (GUI), as shown in Figure 3.6.

Other potential uses of PeMS include level-of-service characterization, incident impacts, and anything that requires high-resolution speed data. Furthermore, PeMS has been used to calibrate simulation models and test new traffic flow theories by researchers throughout the state of California.

The strength of PeMS lies in its ability to combine multiple data sources into an easy-to-use interface that produces useful visualizations of the data. Some of the larger data sources are

- Loop detectors;
- Census detector stations;
- Weigh-in-motion stations;
- Toll-tags;
- Bluetooth sensors;
- Incident logs from the California Highway Patrol; and
- Transit schedules.

More detail on these sources can be found in Petty and Barkley (2011).

PeMS data are easily accessible. The only requirement in setting up a user account is indicating why one needs the data. Users only need to apply for an account once at http://pems.dot.ca.gov/.

3.2.3.2 PORTAL, Portland

The Portland State University (PSU) ITS laboratory is archiving Oregon Department of Transportation (ODOT) freeway inductive loop detector data in a systematic way. The data are streamed to the server located at PSU and then archived in a RDBMS. This system is known as the Portland Transportation Archive Listing (PORTAL). The system has been in operation since July 2004, streaming data from the ODOT Traffic Monitoring Operations Center to PSU (Bertini et al. 2005).

The PORTAL system focuses mainly on freeway data. One of the design goals of the system has been to adhere to the national ITS architecture. The PORTAL system includes a detailed metadata repository and maintains metaschema for all data entering the system, including information generated in the field at the controller and in the traffic management center.

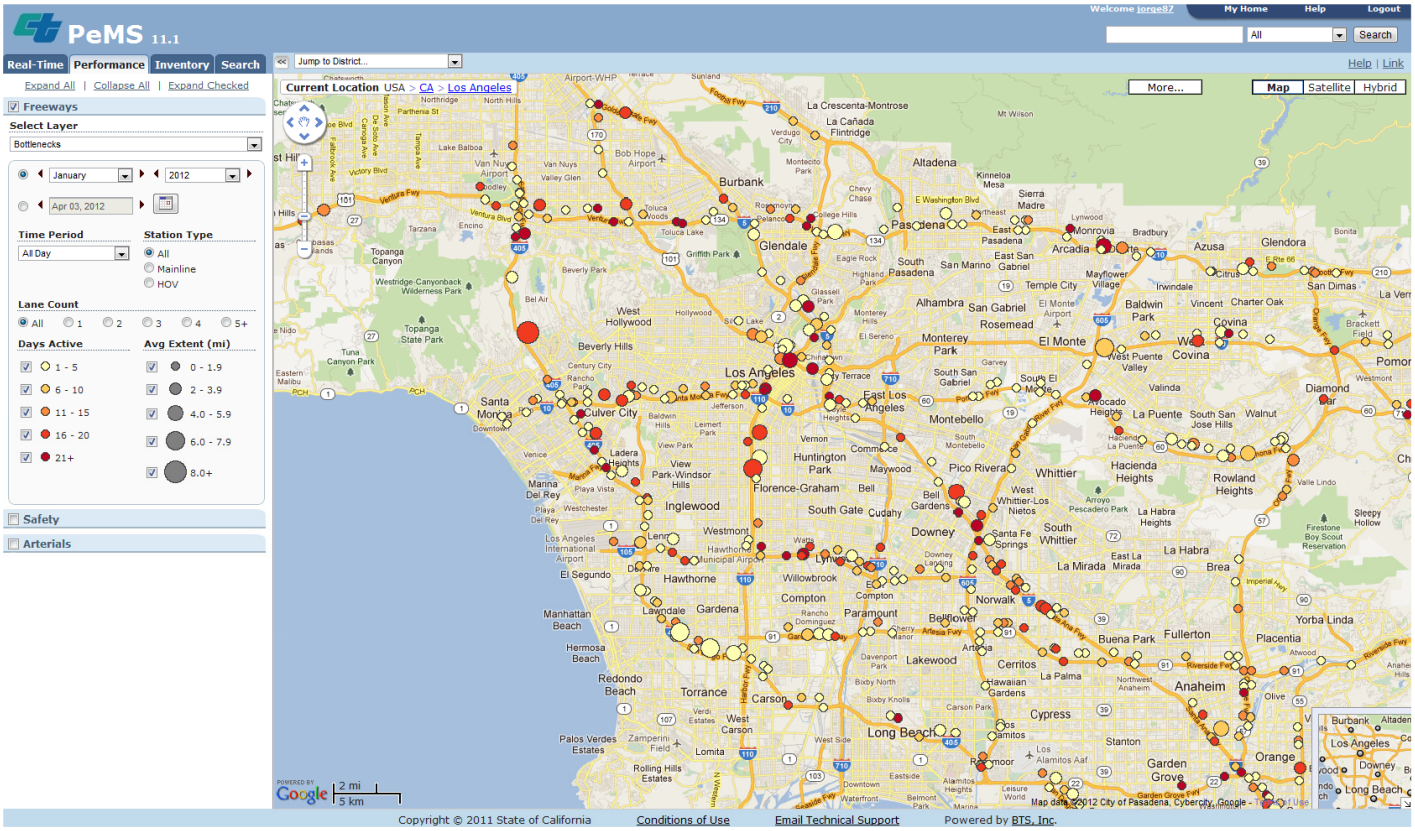


Figure 3.6. PeMS bottleneck identification.

The PORTAL system covers the Portland-Vancouver metropolitan region. The current system (as of the time when the literature review was conducted) archives a wide variety of transportation-related data including the freeway loop detector data from the Portland-Vancouver metropolitan region, weather data, incident data, transit data, and freight data. Information on available data can be obtained from the PORTAL website (http://demo.portal.its.pdx.edu/Portal/index.php/systems).

The system is very flexible and provides various user-configurable parameters. Among the options provided are the following:

- Systems. PORTAL provides a color-coded speed display of the Portland-Vancouver system. The user has the option to choose date and peak periods.
- Highways. This option displays volume and speed data for freeways. Users can choose any freeway within the system coverage area.
- Station. By choosing this option, the user can view different counting stations within the coverage area. Users can choose a specific detector station to obtain speed, travel time, number of lanes, and mile post information.
- Arterial. Volume and speed information can be obtained by selecting date and time ranges. The resolution of these data is available in 5-min, 15-min, 1-h, monthly, and yearly increments.
- Bluetooth. Travel time data are available at some selected locations. Users have the ability to select time and date for data, and start and end stations of the road segments.
- Transit. An interactive map displays different attributes in the PORTAL coverage area. These include transit service areas, transit stops, routes, and boarding frequency.
- Downloads. Speed, volume, and occupancy data can be downloaded from within the user interface by selecting start and end date. These data can be easily accessed using the PORTAL website.
- FHWA data. The data coverage includes freeway transit and arterial data for the I-205 corridor in Portland, Oregon. The selected corridor is approximately 10-mi long. The data set contains freeway loop detector data, weather data, incident data, arterial counts, signal phasing data, limited Bluetooth travel time data, and bus and light rail data.
- Data quality. Information on detector health is provided. These include offline detectors, communication errors, damaged detectors, and configuration errors.

3.2.3.3 CATT Lab, Maryland

The University of Maryland Center for Advanced Transportation Technology Laboratory (CATT Lab) builds, operates, and maintains the transportation data archive for the Washington metropolitan area and other states (University of Maryland 2012). The system is called the Regional Integrated Transportation Information System (RITIS). The data include volume, speed, incidents, weather, and system delays, which are collected by various state and local transportation agencies and transmitted to the CATT Lab's system. RITIS then parses, fuses, and loads the data into databases for analysis, redistribution, and display in near real time. CATT archives the majority of the data for use in other applications including real-time simulation, travel time estimation, traffic mapping and visualization applications, research, and planning.

The RITIS database can be accessed at https://www.ritis.org/. Users need an account to access certain data. A sample of one of the archived incident database application interfaces is shown in Figure 3.7.

3.2.3.4 Center for Transportation Studies, Virginia

The ADMS Virginia project is hosted at the Smart Travel Laboratory, a joint facility of the Virginia Department of Transportation and the University of Virginia. ADMS Virginia is a development effort to archive ITS data for transportation applications. The web-based system uses historical traffic, incident, and weather data to provide traffic data in a variety of formats to users of the system.

The website (http://adms.vdot.virginia.gov/ADMSVirginia) is integrated with Google Maps to produce graphical displays of color-coded travel patterns as shown in Figure 3.8.

To access the ADMS users need to have an account. The account can be requested online via e-mail at the project website.

3.2.3.5 AITVS, Virginia

Virginia Polytechnic Institute and State University's Spatial Data Management Lab has developed the Advanced Interactive

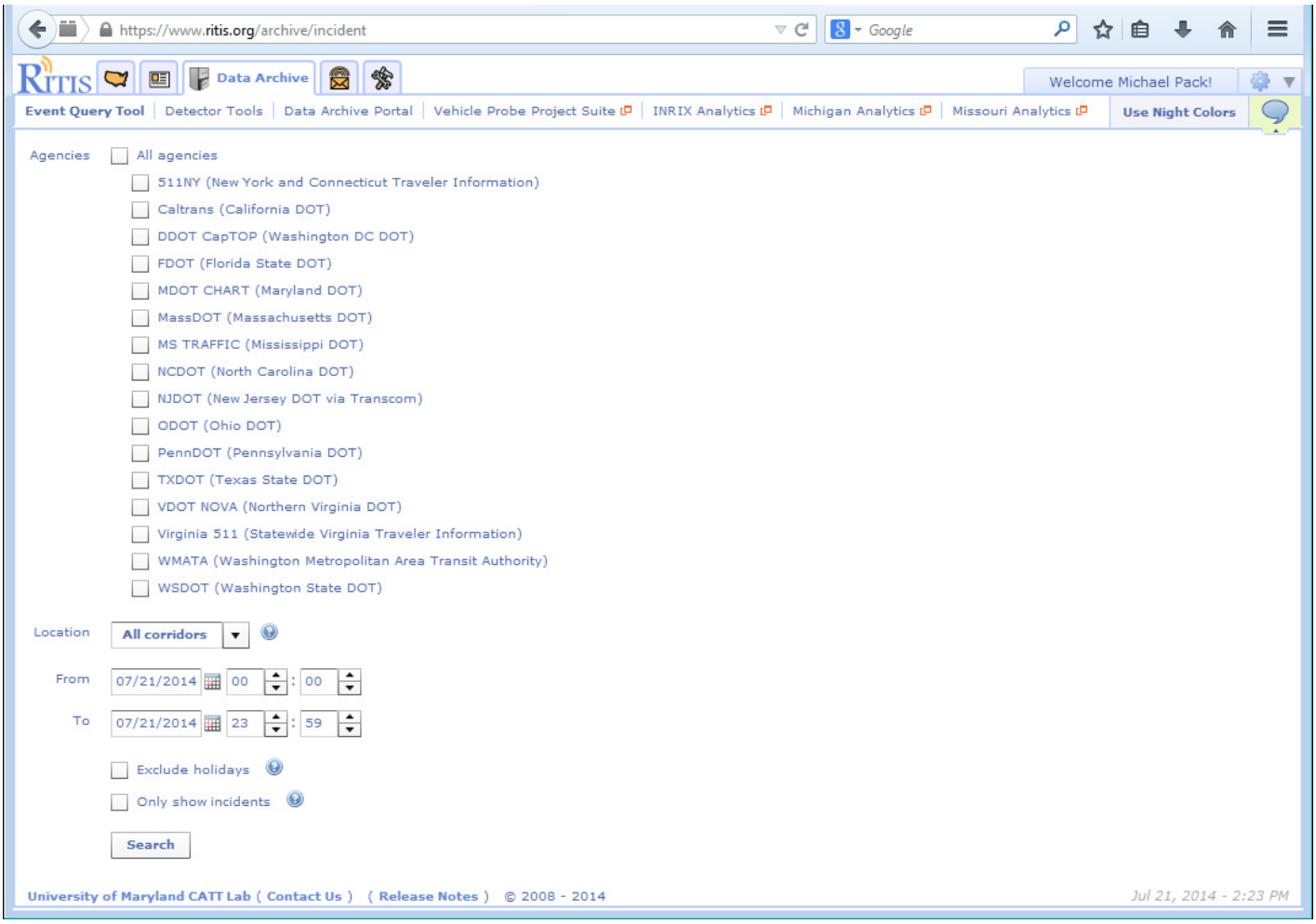


Figure 3.7. Screenshot showing data selection options in RITIS.

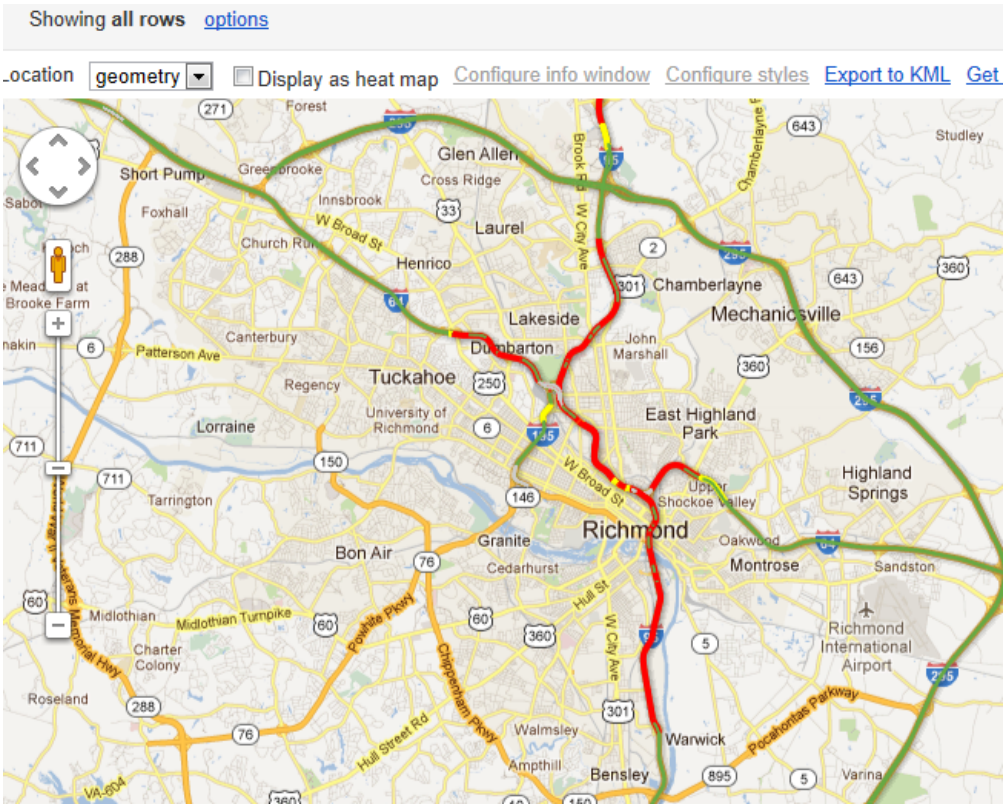


Figure 3.8. Screenshot from University of Virginia, Smart Travel Lab.

Traffic Visualization System (AITVS) that provides real-time highway monitoring capabilities via comprehensive visualization components. AITVS provides a rich set of multidimensional visual components for real-time and historical traffic data analyses (Lu et al. 2006).

The AITVS provides six distinct visualization components that comprehensively cover the various performance metrics of a road system. These visualization components are time plot, date plot, highway station plot, highway station versus time plot, highway stations versus day-of-the-week plot, and time versus day-of-the-week plot (Lu et al. 2006). The speed profile, volume, and occupancy plot, shown in Figure 3.9, can be obtained by selecting pairs of stations.

3.2.3.6 Houston TranStar, Texas

The Houston TranStar consortium is a partnership of four government agencies: Texas Department of Transportation, Harris County, the Metropolitan Transit Authority of Harris County, and the City of Houston. TranStar collects real-time data covering a total of 770 directional freeway miles. Traffic data collection in TranStar relies mostly on automatic vehicle identification (AVI) information. In addition, closed-circuit television (CCTV) cameras cover 335 freeway centerline miles. TranStar has been archiving 15-min aggregated AVI travel time and speed data since October 1993. In addition, the database has freeway incident data dating back to May 1996, emergency road closure data from August 2001, and construction lane closure data from May 2002.

Houston TranStar provides information for multiagency operations and management of the region's transportation system, motorists, and traffic management operators in Houston (Houston TranStar Consortium 2010). Real-time traffic information from the database is displayed in a map interface at the TranStar website (http://traffic.houstontranstar.org) as shown in Figure 3.10. Archived speed data from various freeway segments can be compared in different time horizons.

3.2.3.7 TDAD, Washington State

The Washington State ADUS project, named Traffic Data Acquisition and Distribution (TDAD), was set up to provide traffic data over a wide area over extended periods of time (Dailey et al. 2002). TDAD makes its historical data available online.

TDAD obtains its data from loop detectors across the state, which report volume and occupancy at 20-s intervals. TDAD depends on the state's ITS Backbone Project to obtain the

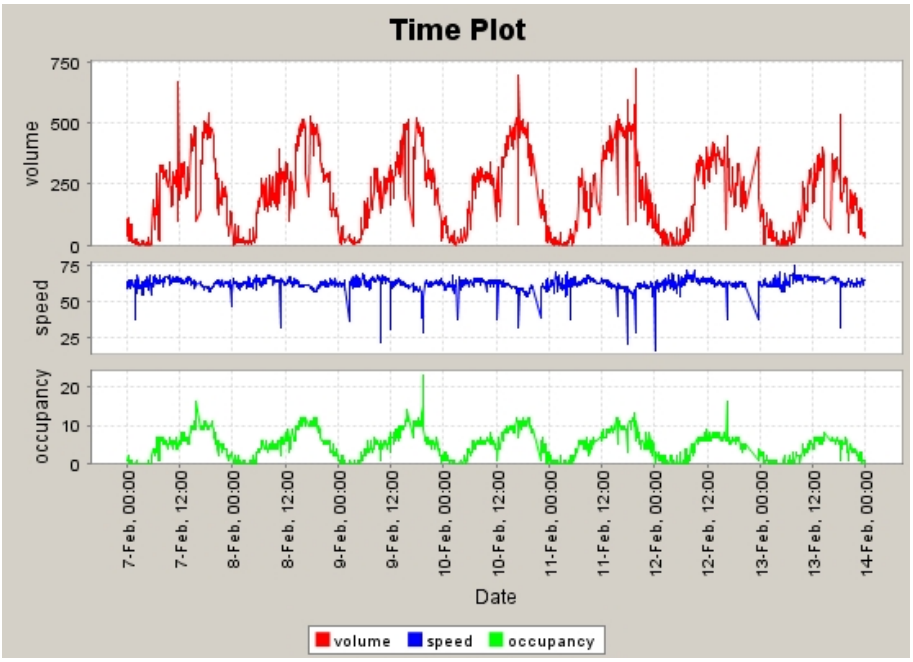


Figure 3.9. Sample plots of volume, speed, and occupancy from AITVS.

data and for operational support. The Backbone Project also serves transit and traveler information programs within Washington State DOT (WSDOT) (Dailey 2003).

To access TDAD data, individuals outside WSDOT must download a toolkit, the Self-Describing Data interface and software library. Several groups—including Iteris, Wavetronix, HERE (formerly NAVTEQ), and AT&T—have developed applications to continuously download, process, and reuse the WSDOT data. Unfortunately, according to the University of Washington's ITS website, the funding for the data feed has not been renewed; thus, the ADUS is unavailable at the moment. This is an example of what can happen if adequate funding is not set aside for operations and maintenance when an ADUS system is initially designed.

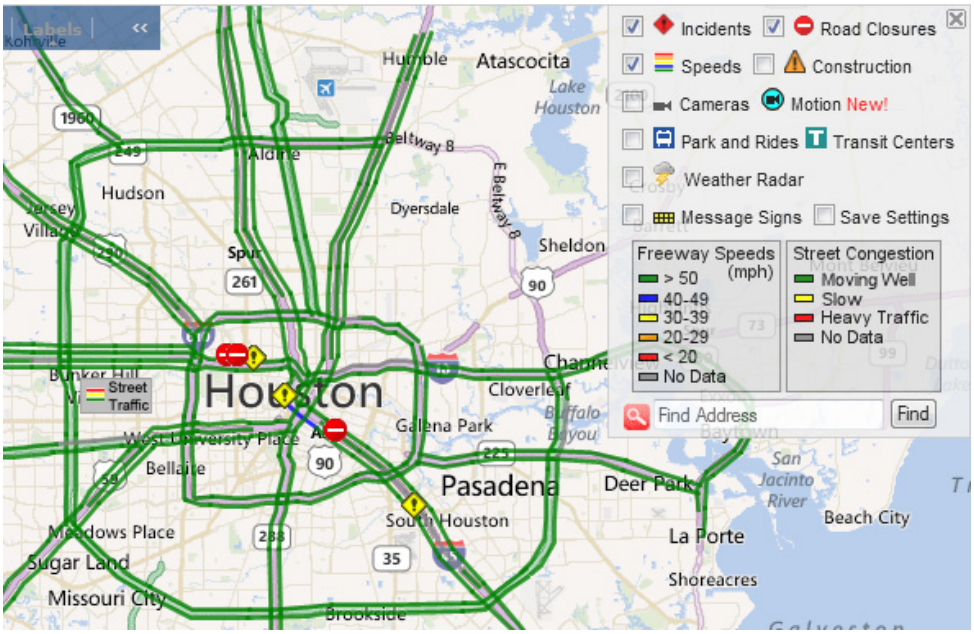


Figure 3.10. Houston TranStar traffic map.

3.2.3.8 Minnesota DOT RTMC

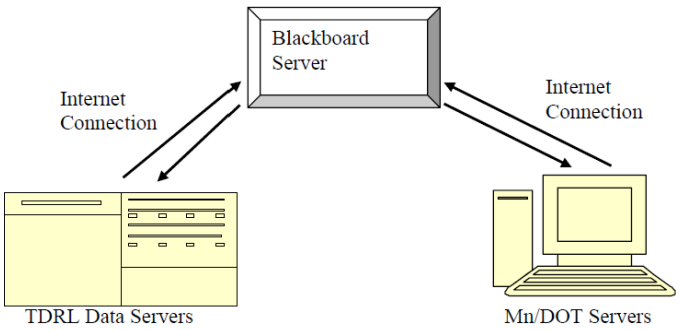
The Minnesota Department of Transportation (MnDOT) built the original transportation management center in 1972 to manage the freeway system in the Twin Cities metropolitan area. The primary purpose of the facility is to integrate MnDOT's Metro District Maintenance Dispatch and MnDOT's traffic operations with the Minnesota Department of Public Safety's State Patrol Dispatch in a unified communications center. The Regional Transportation Management Center (RTMC) now monitors 340 mi of metro-area freeway with 4,500 loop detectors and 450 CCTV cameras (Minnesota Department of Transportation 2012). The RTMC also covers 85 electronic message signs in the region. The RTMC can be accessed at http://www.dot.state.mn.us/rtmc.

MnDOT has developed interface software that transmits a minimum 30-s interval loop detector count and other traffic data from the site to the server located at the RTMC. The data are continuously archived, and more than 6 years are available for download. Lane-by-lane traffic data including volume, speed, occupancy, headway, and density are collected from the permanent loop detectors. The data are available to the public.

MnDOT designed the system to provide data through the Internet. An online relationship was established between the data production capability of the Data Center at the University of Minnesota Duluth's Transportation Research Data Lab (TDRL) and the servers at MnDOT. This concept is shown in Figure 3.11 (Kwon 2004). Data can be written to, or read from, the blackboard server by the TDRL Data Center or MnDOT servers.

3.2.3.9 STEWARD Database, Florida

The Florida statewide ITS architecture contains an archived data management subsystem known as the Statewide Transportation Engineering Warehouse for Archived Regional Data (STEWARD). STEWARD collects and stores statewide data, including daily summaries of traffic volumes, speeds, occupancies, and travel times obtained from SunGuide Transportation Management Centers (TMC) in Florida. The summaries are accumulated over periods of 5 min, 15 min, and 60 min. STEWARD can be accessed at http://cce-trc-cdwserv.ce.ufl.edu/steward/.



Note: The arrow lines indicate Internet data connections and the sequence of data flow.

Figure 3.11. System-level concept of data automation (MnDOT).

Several options are available for users to screen the data they want from STEWARD. Interactive maps for all detectors within District 1 to District 7 of the Florida DOT can be displayed in the STEWARD system. A sample of TMC coverage data selected for download is shown in Figure 3.12.

STEWARD has been designed to appeal to TMC managers, district ITS program managers, and traffic engineers. Some of the useful functions built into STEWARD to make it appealing to managers include the following (Courage and Lee 2008):

- Identify detector malfunctions;
- Provide calibration guidance for detectors;
- Perform quality assessment data reliability tests on data;
- Provide daily performance measures for system, and statewide performance measures;
- Facilitate periodic reporting requirements; and
- Provide data for research and special studies.

The existing STEWARD database contains traffic sensor subsystem data from all TMC stations over a 24-h period. STEWARD serves as a central data warehouse for SunGuide data. The STEWARD output can be used for a variety of purposes. Separate processes involved in the operation of STEWARD are shown in Figure 3.13 (Courage and Lee 2009).

3.2.3.10 The Regiolab-Delft, the Netherlands

The Regiolab Project is a collaborative project between public agencies, research institutes, and industry partners in the Netherlands. The project involves collecting real-time traffic monitoring data from all relevant roads in the region, archiving the data, and developing services and tools that make it easier for researchers to use the data for regional analysis. The public agencies involved in the project are the municipality of Delft, the Province Zuid-Holland, and Rijkswaterstaat. Delft University of Technology, TRAIL Research School, and Connekt institutes are the researchers; and the industry partners are Vialis and Siemens.

According to the project website (http://www.regiolab-delft.nl), the data being archived consist mainly of minute data from inductive loop detectors and variable message signs on the national highways in the province of South Holland. Traffic data are collected from detectors on approximately every 500-m interval on motorways. In addition to the loop detectors, local data from traffic control systems and

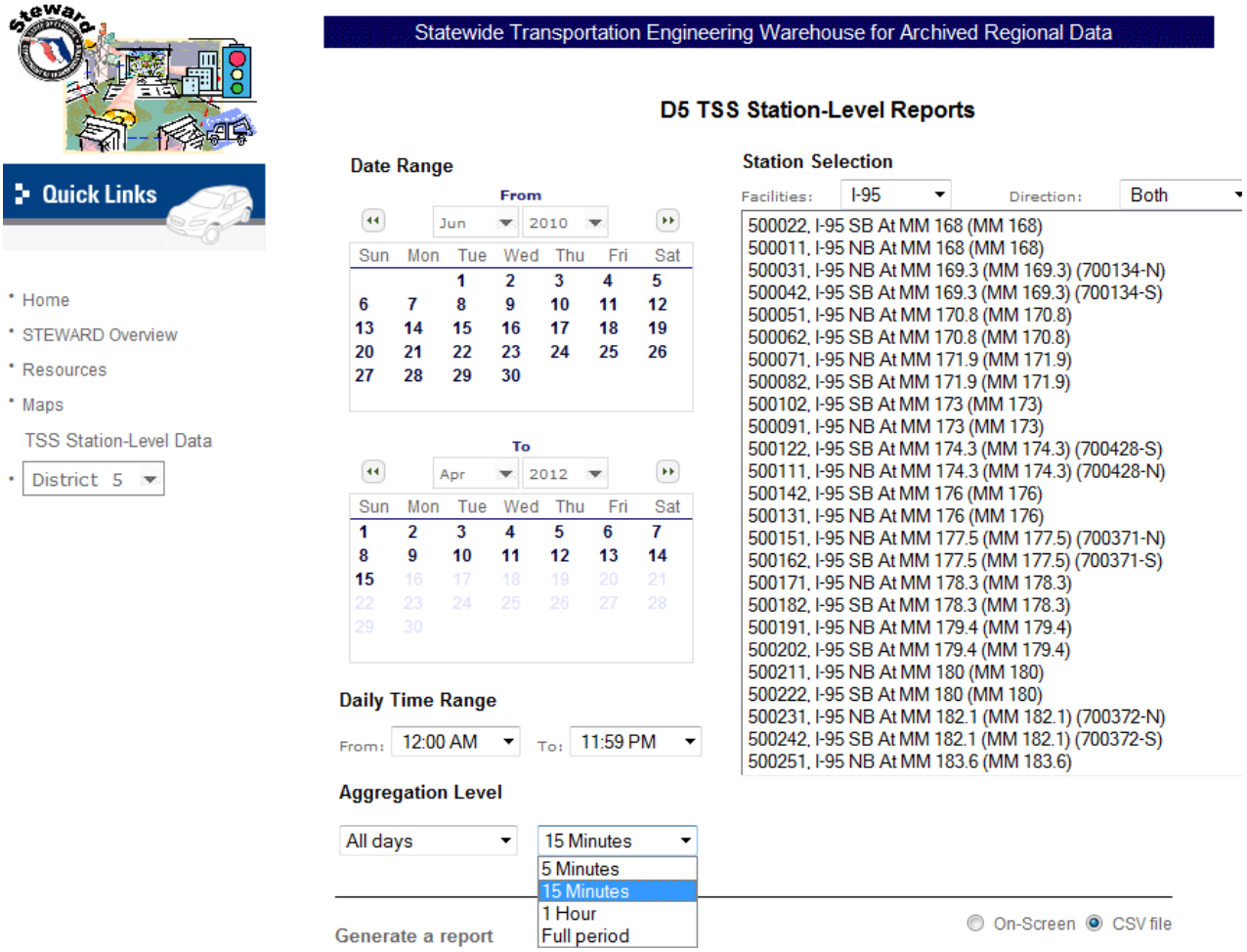


Figure 3.12. STEWARD Florida database.

cameras in the municipality of Delft are also being archived. Sample camera locations are shown in Figure 3.14.

The data archive is being stored and managed using the Drupal content management system.

The traffic data are available for download to registered researchers from the Regiolab website. The website provides a Matlab Toolbox (the program is written in Matlab software) and Structured Query Language (SQL) and other database software tools for extracting data from the archive.

The regional traffic data archive is capable of analyzing traffic flows during the day and can be used to estimate travel times and predict future conditions in the network. Sample charts and visualization tools available from the archive are shown in Figure 3.15.

3.2.3.11 Traffic Data Clearinghouse, Japan

The Kuwahara Laboratory at the University of Tokyo has teamed up with the Delft University of Technology to create a traffic data clearinghouse for researchers (Traffic Data Clearinghouse 2012). Currently there are two key data sets on the project website: the Tokyo Metropolitan Expressway and the data from the Regiolab-Delft project. The aim is to attract more partners and researchers to share their data sets to improve the quantity and quality of traffic data available for traffic modeling. The website can be accessed at http://trafficdata.iis.u-tokyo.ac.jp/index.php. A map of Regiolab in the Delft region from the site is shown in Figure 3.16.

3.2.3.12 Traffic England, England

Traffic England provides live traffic information about the motorways and major all-purpose roads in England. The service is provided by the National Traffic Operations Center of the Highway Agency. Traffic data, traffic volume, speed, and travel time are collected from the motorways and major highways using sensors and readers (i.e., inductive loops and automatic license plate recognition cameras). The information is updated continuously.

Traffic England updates real-time traffic information by displaying speed and delays, roadway closures, major disruptions, incidents and congestion, adverse weather, and roadside

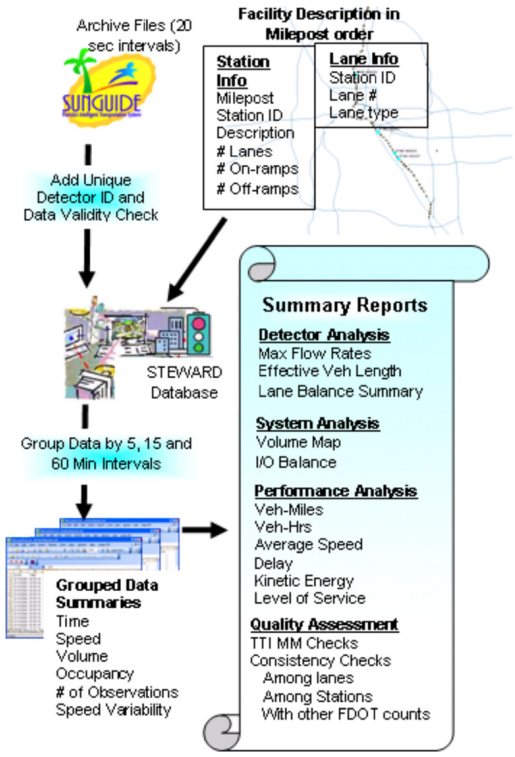


Figure 3.13. STEWARD overview.

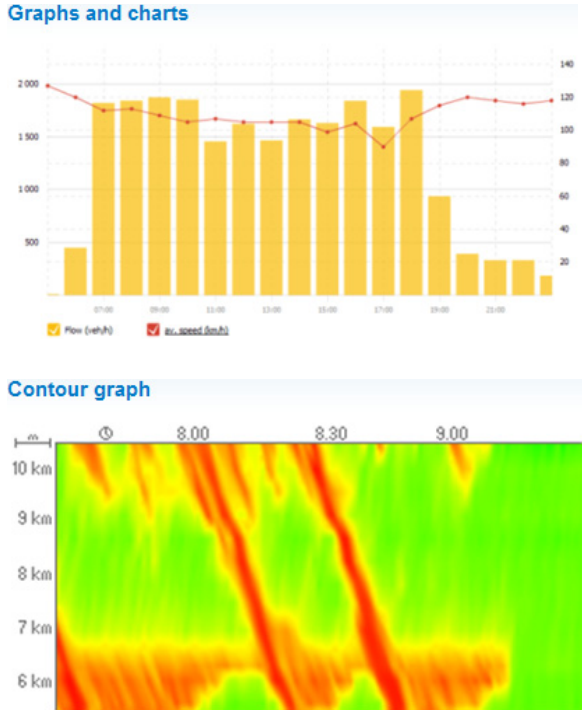


Figure 3.15. Sample chart and contour graph from Regiolab-Delft project website.

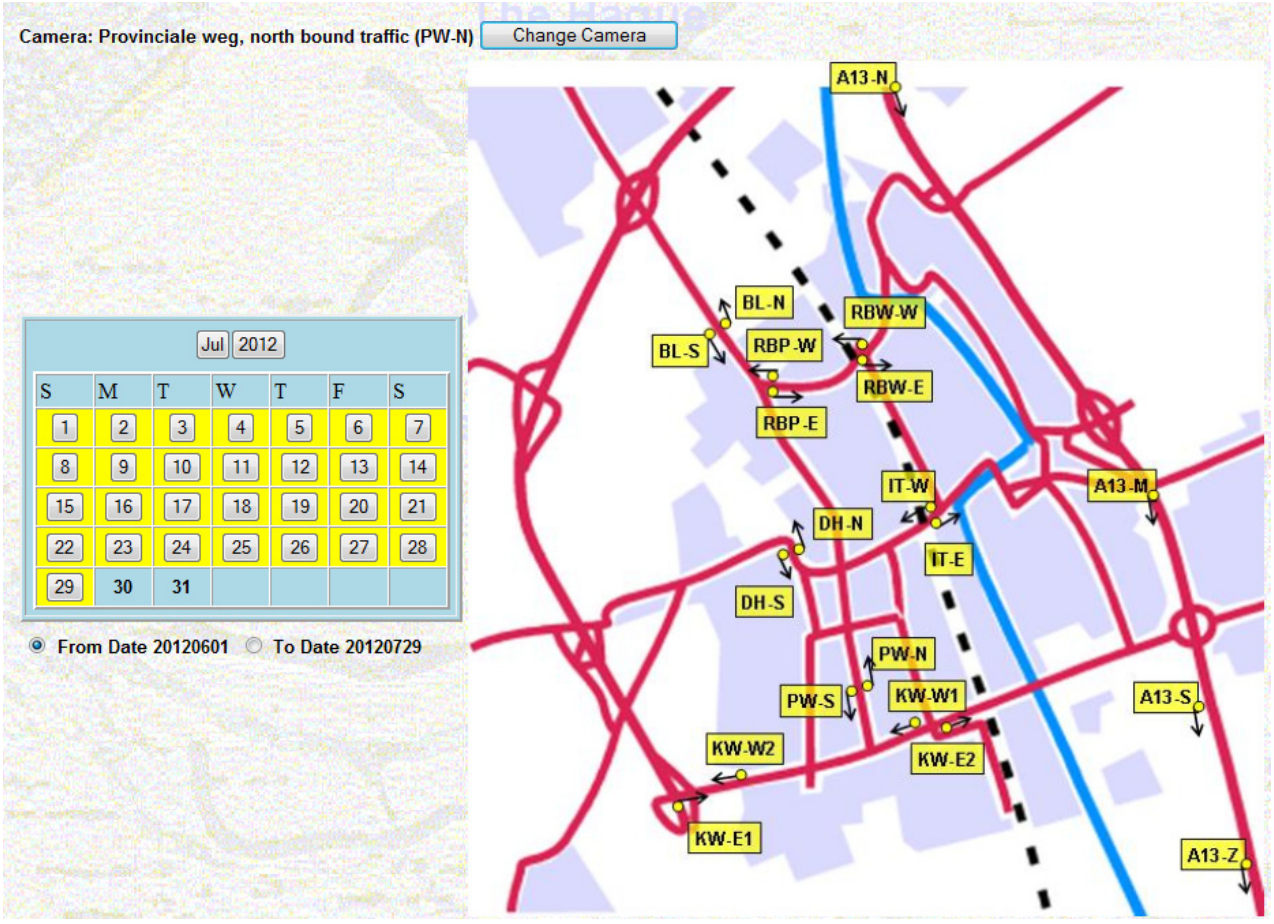


Figure 3.14. Map of camera locations from Regiolab's website.

Figure 3.16. Map of Regiolab-Delft.

message signs. The purpose of this service is to help the motoring public make informed decisions about their journey. Sample real-time information from the Traffic England website (http://www.trafficengland.com/) is shown in Figure 3.17.

3.2.3.13 Land Transport Authority, Singapore

Land Transport Authority (LTA), Singapore, developed a system that provides real-time traffic information including accidents, vehicle breakdowns, traffic signal status, current electronic road pricing rates, and work zones (Figure 3.18). The system can be accessed at http://interactivemap.onemotoring.com.sg/mapapp/index.html.

LTA provides real-time traffic updates by displaying speed, accidents, breakdowns, roadwork, other incidents, and traffic signals down. The purpose of this service is to optimize the road network efficiency and improve road safety for the benefits of all road users. LTA has deployed various ITS components as a part of advanced traffic management systems. The collected traffic data are aggregated, integrated, and disseminated at the ITS Center control room for traffic monitoring and incident management.

3.3 Online Archiving Systems

The L13A team reviewed transportation-related content management systems and existing online archiving systems. This section summarizes the results of the review.

3.3.1 Archived Data Levels

To understand the context of services other data archives provide, the L13A team looked into the five categories of information that were introduced by NASA's Committee on Data Management, Archiving, and Computing (CODMAC)

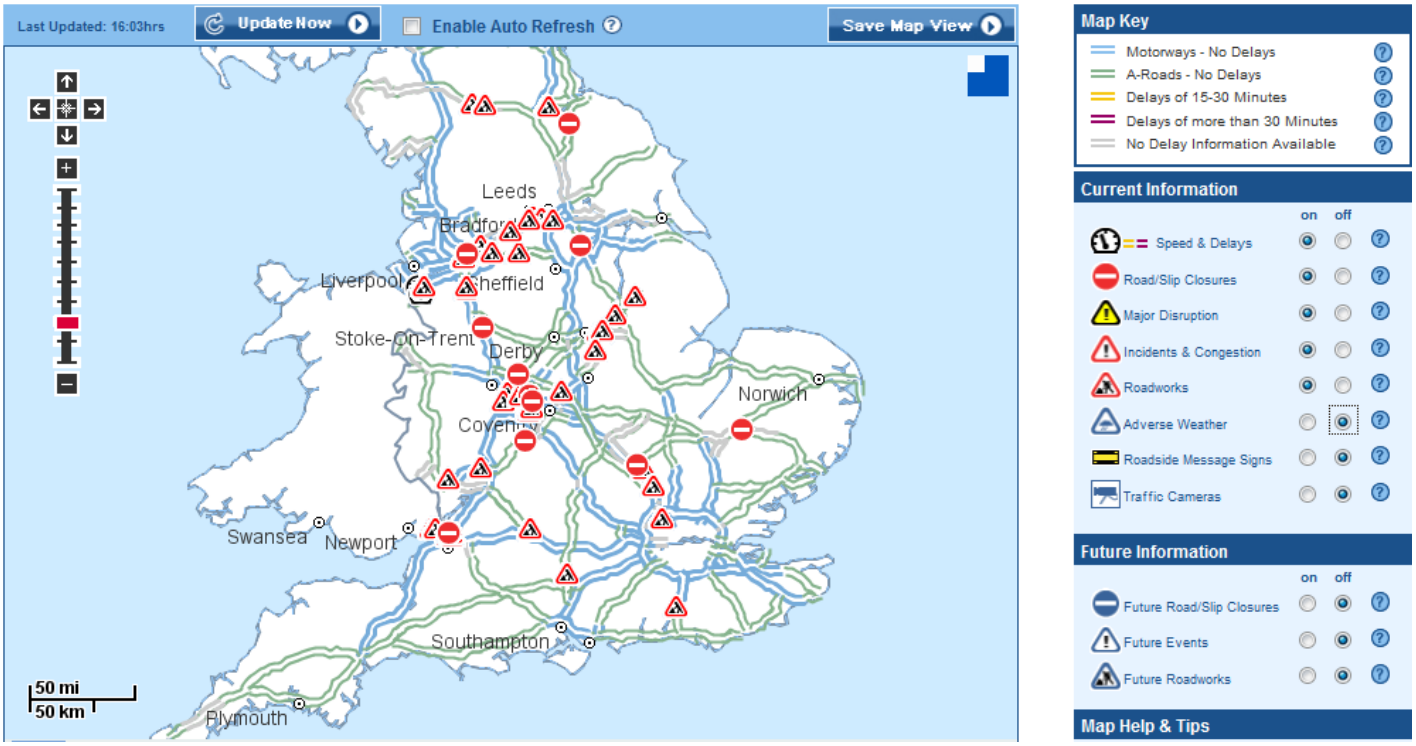


Figure 3.17. Traffic information map of road network, Traffic England.

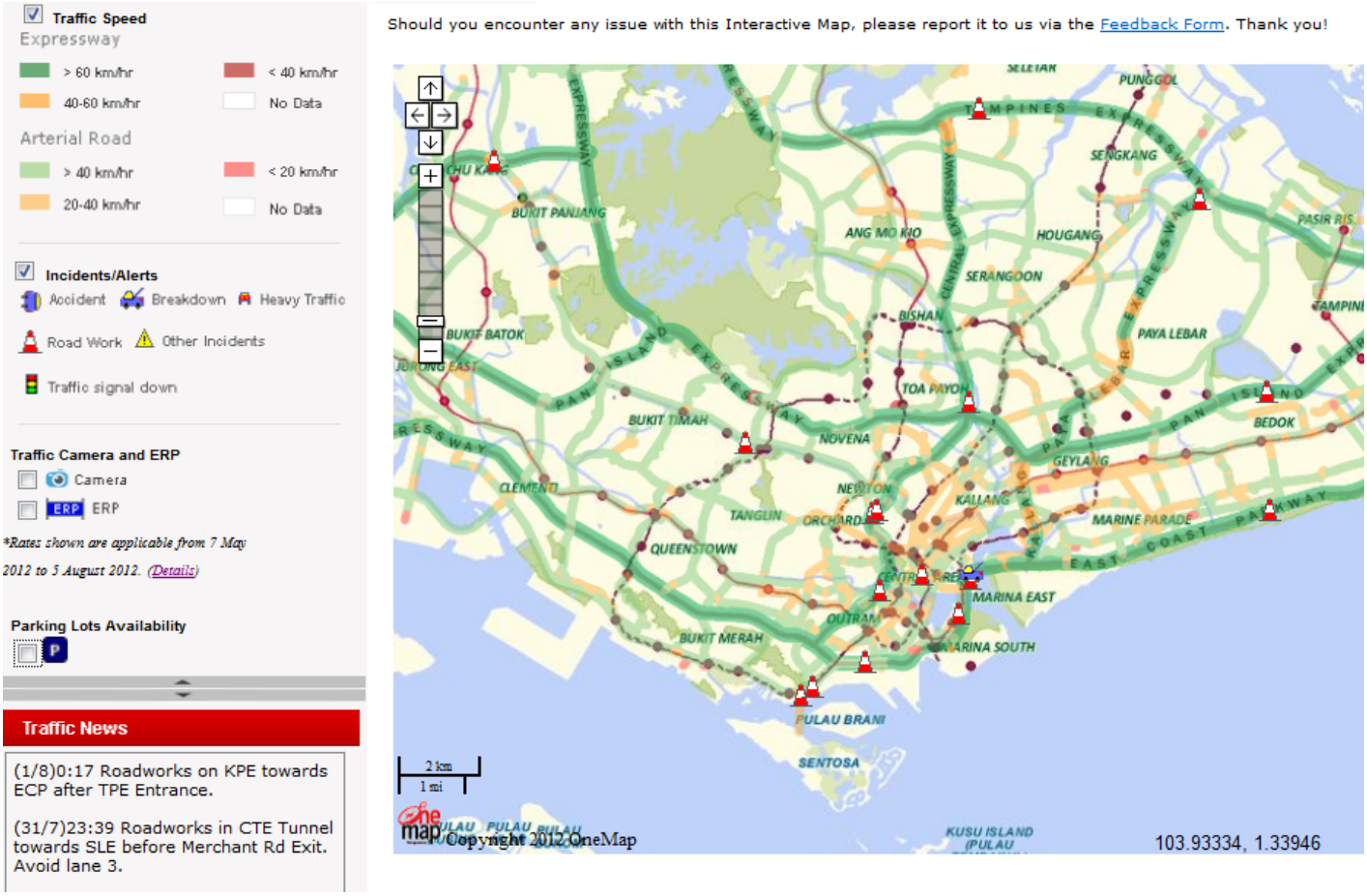


Figure 3.18. Traffic information map of road network in Singapore.

Archived Data Type Ontology, a well-established standard for the handling of archive data. Table 3.3 summarizes the archive data levels suggested by CODMAC.

3.3.2 Document Management Systems and Content Management Systems

Document management systems (DMS) and content management systems (CMS) provide much of the technological foundation for organizing, storing, controlling, and distributing data and results in a controlled environment. Both types of systems usually provide storage, version control, and distribution of electronic documents. CMSs typically provide more functionality, including publishing and editing of content. Both systems often include a centralized interface or portal through which all site content can be accessed.

DMS and CMS form a solid foundation for the handling of documents and web content. However, handling of data may require additional technologies. Data sets may include millions of individual records that may be related in multiple ways. One user's data needs may be vastly different from any others' needs. Storage of data in a fashion that supports individual user requirements implies that data are organized, catalogued, and stored such that they can be accessed according to what data are required by a given user. These are database or data warehousing functions.

3.3.3 Transportation-Related Document Management Systems

The project team reviewed numerous examples of transportation-focused document management systems, such as

1. The National Transportation Library and TRB's TRID, including the Transportation Research Information Services database (http://www.trb.org/InformationServices/InformationServices.aspx) and the Organisation for Economic Co-operation and Development's Joint Transportation Research Centre's International Transportation Research Documentation database (http://www.internationaltransportforum.org/jtrc/itrd/); and
2. The National Transit Agency database (http://www.ntdprogram.gov/ntdprogram/).

Table 3.3. Archived Data Levels

Level	Description	Example Formats
Level 0	Raw data, including raw traffic data such as volumes and speeds	Raw digital data and imagery
Level 1	Georeferenced data, such as speed associated with a specific route and direction	Individual records, processed images
Level 2	Derived variables at the same resolution and location as the Level 1 source data from which the variables are derived	Individual records, processed images
Level 3	Variables mapped on space-time grid scales	Imagery depicting the changes in time and/or space of variables
Level 4	Model output or results from analyses of lower-level data (i.e., variables derived from multiple measurements)	Model output files
Level 5	Reports and presentations using lower-level data	Abstracts, scientific papers, and presentations, typically in PDF, Word, or PPT format

Early examples of transportation document/data archives were relatively simple websites providing access to documents and data, such as the University of California, Berkeley, Freeway Service Patrol (FSP) project data archive.

It should be noted that many of the reviewed transportation archive systems are primarily focused on Level 5 information. They provide information about transportation projects and access to reports and documentation but not raw data (FSP is a notable exception to this, providing Level 0 data). By contrast, weather and social science archives focus more on providing the raw data in a form that researchers can use.

3.3.4 Comparison of Existing Online Archives

Existing online data archives were surveyed for their relevance to the L13 project. Table 3.4 lists a variety of climate, weather, social science, and transportation-related data archives. The nontransportation data archives provide the types of services (to varying degrees) that are envisioned under the L13 project. Transportation archives are noted for their domain relevance.

While many of these archives are referenced in Table 3.4, it should be noted that the Research Data Exchange (RDE) is very similar in scope to the L13A data archive. The RDE includes real-time data distribution and some additional capabilities regarding the management of data environments but is otherwise similar. At the time of this writing, the RDE was in development by FHWA. Lessons learned from the RDE project were not available because it was in the early stages of development; however, what is known is that the RDE will use a content management system such as Alfresco or Nuxeo and that it will include database and/or data warehousing functionality as required, depending on the characteristics of the data sets provided by the Connected Vehicle program.

Some data archives allow users to view data online using visualization tools. This is most relevant for data that can be organized geographically and overlaid on a map. Such visualization can enable rudimentary analysis and help the user determine if the data set may be of value. One large-scale example of this visualization is the one provided through the Earth Observing System Data and Information System (EOSDIS), which can be accessed at https://earthdata.nasa.gov/.

EOSDIS is several orders of magnitude larger in size than the L13 project envisages, but other than archive size and distribution rate it is remarkably similar to the L13 project in many ways. It includes collaborative information, project descriptions, data organized as individual files, and visualization of some of the data without download.

Similar visualization can be applied to traffic data, because such data are naturally organized geographically. Many transportation management systems use some kind of visualization to make traffic data easier to follow; a few, such as PeMS, maintain historical data online to permit visual analysis and trending.

3.3.4.1 Commercially Available Archiving Technologies

Technologies reviewed to help implement the L13A Archive—including content management, web services, and file distribution tools—are summarized in Table 3.5. These technologies were sorted roughly in order of priority. The L13A team assessed feasibility of the listed technologies before starting the development phase (Phase 3).

The objective of this assessment was to identify the best archiving or content management technology that

- Would provide the core functionality of the Archive; and
- Could be customized for delivery of special features like visualization.

In Table 3.5, the appropriateness value reflects the project team's assessment on how likely this system could be used in the Archive.

Table 3.4. Sample of Existing Online Data Archives Focused on Research

Archive	Domain	Size	Increase	Data Levels	Real-Time/ Near-Real-Time?	Visualization?	Collaboration?	Search?	Notes
National Environmental Satellite, Data, and Information Service (NESDIS) http://lwf.ncdc.noaa.gov/oa/climate/climatedata.html	Climate and weather	300 TB (digital)	80 TB/year	1–4	Some data are available NRT. NRT varies from minutes to weeks, depending on the data.	Maps with configurable layers	No	Queries entire site content	Privately hosted data centers, including digital and non-digital media
Clarus System http://www.its.dot.gov/clarus/	Weather	400 GB	80 GB/year	0–1	Hourly files	Map interface linking to data and quality flags; no visualization	No	No	
Earth Observing System Data and Information System (EOSDIS) http://earthdata.nasa.gov	Climate	4.8 PB	600 TB/year	0–4	Many data feeds available in NRT (minutes, hours)	Varies by research team; map interfaces and layer visualization	Projects, standards, and working groups	Queries entire site content; separate facilities for searching archives	Privately hosted, distributed data archival and distribution facilities
Data.gov http://www.data.gov/	Public data across a wide variety of domains	50 GB	20 GB/year	0–5	None	Depends on the data set, but much of the data are viewable in a visualization tool	Yes, forums, blogs, various RSS feeds	Yes, across the entire site or subsections	Uses Socrata Size based on current storage of roughly 250,000 data sets, each set averaging 200 KB in size. Rate of increase based on establishment in 2009.
Simple Online Data Archive for Population Studies (SodaPop) http://sodapop.pop.psu.edu	Social Sciences	>500 GB	[a]	0–4	None	None	[a]	Queries entire site content; separate facilities for searching archives	

(continued on next page)

Table 3.4. Sample of Existing Online Data Archives Focused on Research (continued)

Archive	Domain	Size	Increase	Data Levels	Real-Time/ Near-Real-Time?	Visualization?	Collaboration?	Search?	Notes
UCLA Social Science Data Archive http://www.sscnet.ucla.edu/issr/da/	Social Sciences	>500 GB	[a]	0–5	None	None	News posting, integration with Twitter and Facebook	Search for data only	Heavily hyperlinked between multiple universities
U.S. Census Bureau http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml	Social Sciences	>250 GB	[a]	4–5	None	Many data sets can be displayed on a map.	Feedback only	Very detailed and powerful search engines, global site search as well as detailed data search	Endeca (Oracle)–powered search
Bureau of Transportation Statistics http://www.bts.gov/	Transportation	[a]	[a]	3–5	None	Some data sets have predrawn visual summaries.	None	Global site search	
Next Generation Simulation Community http://ngsim-community.org/	Transportation	70 GB	[a]	0–5	None	None	User information and forums	Global site search	
PORTAL ITS data archive http://portal.its.pdx.edu	Transportation	>60 GB	~10 GB/year	0–4	Current traffic data are real time, all available through visualization. No external feeds.	Extensive map and performance measure-based plots	News, Facebook integration	Neither global nor data search. All data is accessed through a variety of intuitive interfaces.	
National Transportation Library (NTL) http://ntl.bts.gov/	Transportation	[a]	[a]	5	None	None	Interaction with librarian only	Search documents	
Caltrans Performance Measurement System (PeMS) http://pems.dot.ca.gov/	Transportation	11 TB	1 TB/year	0–4	Real-time data are included in the archive but not distributed.	Map-based	Map-based presentation of traditional traffic measures and incidents	Global site search	
Connected Vehicle Research Data Exchange (RDE) https://www.its-rde.net/home	Transportation	2 TB	Projected 500 GB/year	0–4	As available from external providers, will distribute real-time feeds	None	Forums, feedback to operators	Global site, real time, and archive data by metadata	Planning to use Alfresco or Nuxeo technologies; prototype uses Drupal Ongoing project

Note: [a] = Undetermined.

Table 3.5. Data Archival Technologies

Tool	Application	Appropriateness (scale of 1 to 10, 10 being highest)	Notes
WordPress	Content management	10	WordPress provides a flexible environment for the developers to easily modify the UI.
Alfresco	Enterprise Content Management (ECM)	8	Alfresco and Nuxeo are considered affordable. Drupal is capable but smaller scale. A detailed analysis of these tools should be performed to select one.
Nuxeo	Enterprise Content Management (ECM)	8	
DSpace	Data archive management	8	Capabilities of DSpace are close to those of Alfresco. Alfresco allows for content management functionality and thus flexible processing of the uploaded content, which is important for special treatment of data sets that are to be visualized.
Socrata	Service	6	Socrata provides a full range of capabilities but is not focused on archiving large data sets.
OpenKM	Document management	5	OpenKM provides document management, not content management, but could be used with additional work.
Drupal	Content management	7	Drupal and CKAN would have to be used together.
CKAN	Portal	7	
Cyn.in	Content Management	5	Cyn.in would need additional work to manage metadata.
S4PA	File management	3	S4PA would require web portal, version management, and other work; however, it is fast and simple.
OpenDocMan	Document management	2	OpenDocMan is not likely to be used in the Archive.
KnowledgeTree	Document management	1	KnowledgeTree is not likely to be used in the Archive.
Fedora-Commons	Data repository	1	Fedora-Commons is not likely to be used in the Archive.
EPrints	Electronic publishing	1	EPrints is not likely to be used in the Archive.
Nesstar	Data cataloging system	6	Nesstar is a system for data publishing and visualization. Nesstar does not have built-in collaboration. Evaluators did not identify how to integrate third-party tools.

3.4 Commercially Available Archiving Technologies

3.4.1 Overview of Applicable Supporting Technologies

The team reviewed each of the technologies in Table 3.5. They provide some components of content management, document management, and web portal functionality.

3.4.1.1 WordPress

WordPress is an open source blogging and content management platform based on PHP and MySQL that runs on a web hosting service. This system has been used widely by many websites. It has a web template system that facilitates UI task building. For more information on WordPress, see Section 3.5.4, Section 7.2, and the website http://www.wordpress.org.

3.4.1.2 Alfresco

Alfresco is a free ECM system written in Java and is distributed in two formats:

1. Alfresco Community Edition, Lesser General Public License, licensed open source; and
2. Alfresco Enterprise Edition, commercially licensed open source.

Alfresco's design is geared toward a high degree of modularity and scalable performance. While the system is free to obtain, an annual subscription is needed for certified patches, maintenance releases, and technical support. Therefore, there were some challenges in customizing the front end. Alfresco has effective content management functionality, which allows for the flexible processing of the uploaded content that is important for special treatment of data sets that are to be visualized. (See http://www.alfresco.com.)

3.4.1.3 Nuxeo

Nuxeo is a free ECM system written in Python that includes functionality, such as document management, social collaboration, case management, and digital asset management capabilities. Nuxeo is similar in scope, scale, and cost to Alfresco and was considered a viable alternative. There were some challenges in customizing the front end. (See http://www.nuxeo.com.)

3.4.1.4 Socrata

Socrata is a cloud-based data publication and collaboration service. Socrata is not a component used to build a service, rather it is the service. Socrata includes web-based management, publication, measurement, and some visualization tools. While Socrata does include a free version, the L13A project required functionality that was only available in the paid versions, including custom metadata. Current pricing plans put L13A beyond the most expensive tier based on the amount of storage required (Socrata's top tier offers only 2 TB). Using Socrata might still be practical but may require discussion with the service's sales staff. (See http://www.socrata.com.)

3.4.1.5 OpenKM

OpenKM is a free Java-based DMS providing web interface for managing files. It is distributed under GNU General Public License (GPL) v.2. OpenKM could be used to support L13A but would require additional development work beyond an Alfresco- or Nuxeo-based solution. (See www.openkm.com.)

3.4.1.6 Drupal

Drupal is an open source content management system. It provides database cataloging and storing of data sets, web front-end development, and an application programming interface (API). It is distributed under GNU GPL v.3. It is less extensive than Alfresco and Nuxeo but includes many of the features needed for L13A. It is a viable alternative, particularly if paired with a data portal such as CKAN (see below). (See http://www.drupal.org.)

3.4.1.7 CKAN

CKAN is an open source data portal system. It provides database cataloging and storing of data sets, web front-end development, and an API. It is distributed under GNU GPL v.3. CKAN could be a viable alternative for L13A if paired with a DMS such as Drupal. (See http://www.ckan.org.)

3.4.1.8 Cyn.in

Cynapse's digital asset management solution is a module of the Cyn.in ECM offering that enables it to leverage a number of inherent features already provided as part of the wider platform. Based on the project team's brief investigation of the promotional literature, support for embedded metadata is missing in this system. However, workflow and transcoding facilities as well as desktop clients are available. Cyn.in is written in Python and Zope. It also uses the Plone open source framework. It is distributed under GPL v.3. (See http://www.cynapse.com.)

3.4.1.9 S4PA

The Simple, Scalable, Script-Based Science Product Archive (S4PA) is a data archive and distribution system distributed under the National Aeronautics and Space Administration (NASA) open source agreement. It includes a data acquisition module suitable for real-time ingestion and a data distribution module that provides data files to users. Data are managed in a tightly organized UNIX file structure. Data storage and distribution are file-based. The S4PA kernel includes subscription services. Data distribution and acquisition use FTP or sFTP (Secure FTP).

S4PA does not include its own web-based front end or any collaboration tools. NASA uses an online visualization tool called Giovanni (http://disc.sci.gsfc.nasa.gov/giovanni/overview/index.html) to allow researchers to visualize and examine aspects of data without having to download entire data sets.

Use of S4PA would require the development of a data portal front end or integration with another tool such as CKAN, the feasibility of which was not clear. (See http://disc.sci.gsfc.nasa.gov/additional/techlab/s4pa.)

3.4.1.10 OpenDocMan

OpenDocMan is a free, open source web-based PHP DMS distributed under GPL. It is not a CMS—it only allows users to upload files with limited metadata description; tag them; maintain revision history; classify documents by category, department, or author; and search by category, department, or author.

OpenDocMan runs with PHP 5, MySQL 5, and Apache HTTP server. The system has some simple user management. The team decided OpenDocMan did not have sufficient capabilities for L13A. (See http://www.opendocman.com.)

3.4.1.11 KnowledgeTree

KnowledgeTree provides a cloud-based service for document management and workflow. Its representational state transfer

(REST) and Simple Object Access Protocol (SOAP) APIs allow integration into third-party websites. This solution did not have sufficient capabilities for L13A as it was not highly customizable and does not handle user metadata. (See http://www.knowledgetree.com.)

3.4.1.12 Fedora-Commons

Fedora defines a set of abstractions for expressing digital objects, asserting relationships among digital objects, and linking services to digital objects. The Fedora Repository Project implements the Fedora abstractions in an open source software system under Apache license. Fedora provides a core repository service (exposed as web-based services with well-defined APIs). Fedora is not an out-of-the-box product that can be installed and run as an application. It is a repository framework, which requires an extensive software development to be able to run simple examples. Fedora lacks UI; a third-party tool such as DSpace would have to be integrated to provide a collaboration engine, such as user forums, and community/user group management. (See http://www.fedora-commons.org.)

3.4.1.13 EPrints

EPrints is open source software under GPL v.3 and Lesser General Public License (LGPL) v.3 for building open access repositories that provide UI as well as a repository engine. Although EPrints allows metadata and UI customization, its focus is on publishing collections of online journals. Thus, it is mostly suitable for document-type content. EPrints does not provide a collaboration engine and does not have detailed instructions about integration with third-party tools. (See http://www.eprints.org.)

3.4.1.14 Nesstar

Nesstar is a free software system designed for online publication and dissemination of data and metadata. The system also includes data analysis and visualization tools, including maps. Survey data, multidimensional tables, and text documents are all supported; and the system software allows users to search, browse, and visualize the data online. Nesstar has limitations in UI customization. All Nesstar catalogs on the web look the same. The deployment of Nesstar requires three products: (1) Publisher, a tool for uploading the data and preparing it for publication; (2) Server, a data repository; and (3) WebView, a UI that allows searching, browsing, and visualization. Nesstar does not have built-in collaboration. Evaluators could not determine how to integrate third-party tools and thus data upload capability for collaborating users. (See http://www.nesstar.com.)

3.4.1.15 DSpace

DSpace is open source repository software distributed under BSD license for storing digital content. It manages digital files of any format. DSpace allows for customization of metadata, as well as the user interface. The software is continually expanded and improved by a community of developers. Its capabilities are close to those of Alfresco. DSpace focuses on the approval of content rather than wider workflow customization. (See http://www.dspace.org.)

3.5 Summary of the Preparatory Analysis

As part of the project, the team reviewed the L13 report, existing ADUS systems, and past work on data archiving systems. The major goal of the preparatory analysis was to select an appropriate core CMS engine with which the Archive system would be built. This section summarizes the outcomes of the review effort and goes over the factors that the project team considered for choosing WordPress as the core CMS engine.

3.5.1 L13 Report Review

The L13 final report mostly described the system requirements and proposed a web-based solution, using cloud-computing services and COTS software (Alternative 3). It also estimated the cost of this solution at $5,530,132 over 25 years. Based on the L13 report, the cost of implementation would be $173,425 per year.

The report did not provide any system design other than the high-level concept shown in Figure 3.5. In addition, it did not specify any particular technology for a CMS, although it recommended COTS over open source and in-house development.

The project team generally agreed with the analysis performed by the L13 researchers except for their deemphasis on high-level data visualization and their 70-TB storage requirement. The L13A team concluded that including a high-level visualization tool would provide both a flexible way for users to view objects and a standardized way for visualizing and aggregating objects. The project team's preliminary assessment of SHRP 2 Reliability artifacts confirmed that the proposed 70 TB storage requirement seemed excessive.

Additional details on the L13 report can be found in Section 3.1.

3.5.2 Archived Data User Service Analysis

3.5.2.1 Federal ADUS Guidelines

The L13A Archive has more diverse and unstructured content than a typical ADUS archive. However, there were still

lessons that could be drawn from the review of the ADUS guidelines:

- Ensuring that institutional issues like privacy concerns, liability, and confidentiality of privately collected data were taken care of in the data provided by SHRP 2 project teams; and
- Incorporating training and outreach in the project. The key to successful outreach will be to show that ADUS systems help perform common tasks faster and more easily and accurately.

3.5.2.2 Summary of ADUS Systems

Other than the Washington State TDAD database, most of the ADUS systems reviewed have been successful in engaging users even beyond the transportation community. A key element in engaging users has been the incorporation of analysis tools and map-based displays (which were included in the L13A Archive).

The L13A team noted that the University of Maryland was able to use an iterative user engagement process, as proposed for SHRP 2 L13A, in the development of its ADUS. This process helped the university develop a final product that met the needs of the target audience. The project team also learned that all the state ADUS systems included quality measures to ensure a high level of data accuracy and integrity.

An overview of existing ADUS systems can be found in Section 3.2.3.

3.5.3 Online Archiving Systems Analysis

The data archives surveyed have a number of features in common that appeared successful and pertained to the L13A Archive:

1. Comprehensive site search that allows the user to query across all site content aside from data archives. This makes it easy to find information about how to use the site and to collaborate with other users.
2. Data archive search by any and all available metadata. This is one of the primary tools that users can use to identify data that may be of interest to them.
3. Data visualization to help users grasp the potential value and applicability of data sets. Many of the archives identified here lack visualization. While they do serve large communities and provide much information, the lack of visualization is a barrier to use; it makes initial investigation of these archives more difficult. It is not clear, but is conceivable, that the data are similarly obfuscated to the ostensible users. By contrast, the EOSDIS systems integrate visualization with search functionality, which provides convenient data preview and engages the user. If practical and affordable, inclusion of some visualization is desirable.
4. Provision of system performance characteristics, so that contributors can see how their data are being used and thus quantify the benefits of sharing their data.
5. Collaboration tools with feedback mechanisms, such that researchers can provide information about their use of data sets to other researchers. Constructive criticism can yield more useful data in the future, foster additional collaboration, and encourage use of the Archive.
6. Feedback on archived artifacts. This feature is similar to the previous point but includes a notion of quality to entice or discourage (as appropriate) use of data. Without an understanding of data quality it is hard to determine with confidence how seriously any research should be taken.
7. Following best practices in clean and simple web design. Some of the studied data archives have been around for a long time and have varying degrees of complexity and artistic standards applied to their designs.

3.5.4 Commercially Available Archiving Technologies Analysis

The project team considered the following technologies as potential candidates for implementing L13A Archive data:

1. WordPress,
2. Socrata, and
3. Alfresco.

The team ruled out the possibility of using Socrata after a cost analysis. The team then analyzed the WordPress and Alfresco systems by building small prototypes. The team tested functionalities and features provided by each platform to check which fit the Archive needs well. Features that the development team looked into included the flexibility of each platform for front-end and back-end customization, complexity of content management standards, XML content modeling, required learning curve, ability to access the database directly, risk and cost, and extensibility. In the end, the team decided to use WordPress.

CHAPTER 4

System and User Needs and Requirements

This chapter defines key system and user requirements that were used to build the Archive. These requirements were captured as user stories and developed on the basis of two workshops, conducted as part of the L13A project, which aimed to collect SME feedback on essential features of the Archive. This chapter begins with the results of the feasibility study conducted under L13. Note that, while the feasibility report did include a set of requirements, those requirements were written from the perspective of analyzing the feasibility of building such a system and for determining whether various high-level architectural concepts were practical.

Systems engineering (Haskins 2007) would suggest the creation of a Concept of Operations that scopes the goals and needs of the system. However, much of the intent of that work was accomplished in the L13 report; what was required was a set of needs that can be used as the basis for system development and validation. From these needs a set of requirements could be generated, which in turn would form the basis for system verification. For Project L13A, the needs were defined in a series of user stories. Each user story describes how a user interacts with the system. In systems engineering this is partially the function of documenting scenarios. In software development as practiced using agile methodology, user stories take the place of requirements and form the basis for system verification, since the verification of the user story indicates the system accomplishes the described task.

4.1 System Overview

At the highest level, the Archive system provides a web-based interface to users, from which it provides access to a variety of tools to search the data archives, acquire data, and submit data for inclusion in the Archive. A similar interface is provided to administrators, who have additional abilities, including review of submitted data and maintenance of the data and metadata in the Archive. Data are submitted by users and held separately until an administrator can examine and validate their format and contents; at that time those data may be moved into the data archive, where they will be available to the web server and thus other users. Figure 4.1 illustrates this very high-level view of the Archive system.

4.1.1 Types of Artifacts

The term *artifact* in this document covers any stored digital objects (e.g., document, data set, computer code) in the Archive system that will be viewed and used by researchers. This document will refer to two types of artifacts: (1) primary SHRP 2 artifacts, and (2) user-submitted artifacts. Primary SHRP 2 artifacts (or project artifacts) are those from the SHRP 2 Reliability-related project teams and are the primary focus of the Archive system. User-submitted artifacts are secondary files, typically derived from the primary artifacts, which Archive users can submit to the Archive system as part of the collaborative research process. Primary SHRP 2 artifacts and their related metadata are the foundation of the Archive system. As of the date this report was written, the L13A team decided not to let regular users submit any non–project artifacts due to security and privacy concerns. As a result, no user-submitted artifact has been uploaded into the system.

4.2 Roles

Four major user roles are envisioned for artifact ingestion: administrator, principal investigator (PI), registered user, and guest. To support different roles for different artifacts or projects, these roles are associated with individual user accounts. Table 4.1 compares user roles for the Archive system.

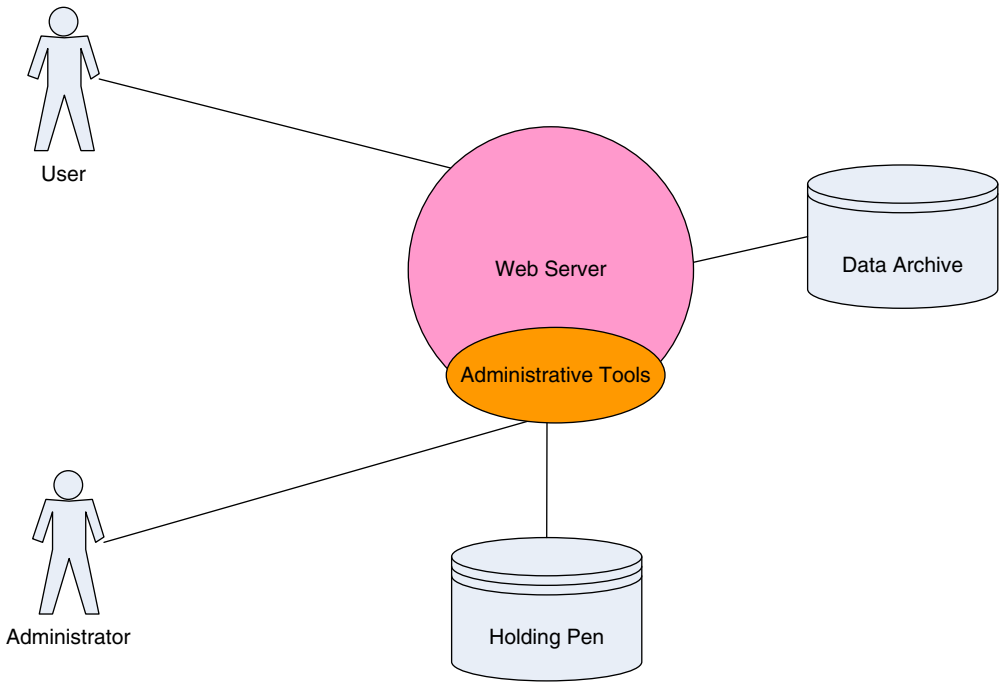


Figure 4.1. Archive system high level.

4.2.1 Administrator

The administrator role is performed by the project management members of the Archive system management team. The administrator's role serves the following functions:

- *Archive setup.* Administrators define the system processes (e.g., access policies, ingestion process), specify metadata attributes and relationships, and define user roles. Administrators also create the Archive folder structure that stores the artifacts, generate initial metadata describing Archive folders, and test the artifact and metadata submission UI process.
- *Access and content moderation.* Administrators work with project teams to determine the list of artifacts to be submitted for each project; perform quality/content control checks of submitted artifacts and metadata; and moderate user profiles, roles, and comments. Administrators have the final authority to approve and accept artifacts (if needed) and metadata.
- *System monitoring.* Administrators are responsible for regularly monitoring system health and usage statistics, allocating information technology (IT) resources to address areas of concern, providing technical support to users on an as-needed basis, and creating regular preservation planning reports (~ every 3 to 5 years).
- *Preservation of PII.* Administrators have to make sure no PII data are shared with the public via the Archive.

4.2.2 Registered User

Each registered user creates a user account, which contains a minimum of information: userID, password, and e-mail address. The e-mail address of the user is verified during the

Table 4.1. User Roles

Role	Guest	Registered User	Principal Investigator	Administrator
Download artifacts	✓	✓	✓	✓
Search and visualize artifacts	✓	✓	✓	✓
Participate in the discussion	✗	✓	✓	✓
Upload artifacts from the web page	✗	✗	✓	✓
Delete artifacts	✗	✗	✓	✓
Permanently delete artifacts from Archive	✗	✗	✗	✓

registration process. Users may choose to store additional information in their user account.

Registered users can view pages, download artifacts, and write comments. Their user account keeps track of their comments and file uploads. These users can edit metadata only on artifacts that they have uploaded.

4.2.3 Principal Investigator

Principal investigators are registered users who can upload primary SHRP 2 artifacts. A *PI* takes on the role only for those artifacts that he or she uploads to the Archive system. There can be only one PI per artifact in the Archive. The PI is responsible for preprocessing the artifact, creating the mandatory metadata associated with the artifact, and completing the artifact submission process.

The PI is also responsible for performing a formal quality control check of the artifact and metadata after completing the ingestion process, to ensure high quality of submitted data and to minimize the burden of quality assurance on the administrator. The PI can also monitor comments and other collaboration on the artifact over time and update metadata to address any concerns raised by the user community.

4.2.3.1 Creator

The notion of *creator* is conceptual and has not been used as a distinct user type in the Archive because, from the Archive system's point of view, the creator and the PI are the same. They are registered users of the Archive system who can upload artifacts. The only difference is that a creator can be someone other than the PI. The creator is assigned by the PI or SHRP 2 to perform preprocessing and metadata information collection tasks on behalf of the PI.

4.2.4 Guest

To create the most open system possible, the Archive was set up so that any Internet visitor may access the system as a guest. However, in an attempt to minimize spam and archive abuse, this role is limited to "read only" and is limited to viewing pages and downloading data. Without a user account to tie other activities back to, guest users cannot write comments or upload files.

4.3 System Needs

This section discusses the Archive operational goals, system needs, and user needs. Operational goals establish parameters and targets for system performance. *Needs* identify what the system needs to do:

- System needs describe what the system needs to do to meet operational goals.
- User needs describe what the system needs to do to satisfy users.

The needs were developed on the basis of the comments gleaned from the June 2012 stakeholder workshop and project revised work plan. They were written in a language compatible with Institute for Electrical and Electronics Engineers (IEEE) Standard 1362. While this document does not follow that standard completely [it is not a fully fledged concept of operations (ConOps)], maintaining some of the structure within the traditional ConOps allows cleaner traceability to requirements and subsequent needs validation and requirements (user story) verification.

4.3.1 System and User Needs

1. *Accept data.* The system needs to accept data from users. Data may be provided electronically through an upload of individual files, or loaded into the system by an administrator from physical media (e.g., a portable hard drive, optical media, memory key). This allows producers of test data to share their data with other users, and also allows users of that data to provide transformed versions of that data back to the community. This enables collaboration, corroboration, verification, and other research activities, without requiring researchers to enter into a direct relationship.
2. *Store data.* The system needs to store data provided by users in an organized manner so that other users may access that data. Such storage needs to maintain the data in its original form and with access restrictions stipulated by the provider, until such time as the data are no longer needed. This provides the essence of a long-term data archive.
3. *Distribute data.* The system needs to provide users the ability to download the data they select from the Archive. This includes the ability to download complete test data sets. Downloads will occur electronically and may be limited in speed by available communications bandwidth both at the Archive and at the data requester.
4. *Maintain metadata.* The system needs to associate metadata with data sets and files. This data about the data allows users to associate characteristics with data sets and, in turn, enables multiple dimensionality of association between data sets, access rights to data sets, and search functionality.
5. *Search for data.* The system needs to provide a mechanism for users to search through available data. This mechanism allows the user to input criteria, and the system will then provide a list of data that are relevant to those search criteria, which the user can then examine and download.
6. *Site navigation.* The system needs to provide a logical organization of the data archive and collaboration website. This

allows users to browse available data and discussions. Essentially, this means developing the data archive site in a modern style that is easy to navigate.

7. *Allow modification of metadata.* The system needs to allow users to modify the metadata associated with data sets and files. This function may be restricted based on user permissions. It allows enhancement to and correction of metadata. It also provides for the possibility of expanded definitions of metadata, which is relevant for those data sets whose formats have not yet been defined.
8. *Allow data review.* The system needs to provide a staging area for data input, where an authorized user can review submitted data and determine if it is acceptable for distribution. This allows quality control over data inputs. This implies an administrator or content manager role needs to exist to examine the data and determine its acceptability.
9. *Organize data.* The system needs to provide tools that allow users to organize data, so that data sets and files may be located, associated, and downloaded. This includes linking data to external data sets. This is subtly different from metadata associations, though the implementation may use similar or the same mechanisms if appropriate. Data set and file association should be available only to users permitted to manage site content.
10. *Backup.* The system needs to provide a backup of itself—including its configuration, functionality, and all stored data—to a remote location, so that in the event of a hardware failure the system may be restored.
11. *Predictable performance.* The system needs to provide its services in a predictable fashion commensurate with similar services offered by other facilities. This manages user expectations.
12. *Expandable.* The system needs to be able to expand its storage capacity to accommodate more and larger data sets (at least 50 TB of data). This allows the system to grow and be maintained over the projected system lifetime.
13. *Service.* The system must be able to expand its storage capacity without disrupting services to users. This allows system upgrades without downtime and enables the system to cope with changing storage requirements as noted in Need 12.
14. *Extensible.* The system must be able to support new data management technologies as those become available. This means it must be architected in such a way that data management functionality may be isolated and replaced. This allows the administrator of the data archive to extend the system's capabilities over time.
15. *Encryption.* The system needs to provide the ability to encrypt any data that it exchanges, including administration exchanges, data storage, and data dissemination. This secures data and information exchange, which may be a requirement for some data sets in the future.
16. *Compression.* The system needs to provide the ability to (losslessly) compress any data that it exchanges, including during storage and dissemination. This frees up network bandwidth, thus reducing overall system costs.
17. *Logging.* The system needs to log auditable system configuration and performance information. This information may also be presented to permitted administrative users. Events to be logged include artifact ingestion process errors, system warnings, and system statuses.
18. *Availability.* The system needs to be designed to be 100% available, operating with no scheduled downtime, except in the case of short outages for system updates and longer outages due to an occasional rebuild.
19. *Data submission testing.* The system needs to provide tools that analyze data sets for administrator-specified criteria. This allows the administrator to analyze the type, format, and size of submitted data and provides a measure of quality assurance.
20. *Maintain users with individual permissions.* The system needs to provide a means to distinguish user roles and permissions. This allows different users and user classes to be created with varying degrees of control over the system.
21. *Robustness.* The system needs to continue to operate during a single failure instance. This implies the use of failover or fault-tolerant implementation and ensures continued availability.
22. *Administration.* The system needs to provide management capabilities to administrator accounts. These abilities allow the administrator to configure the system; to manipulate files by moving them, deleting them, or modifying their metadata; and to edit other user accounts.
23. *Facilitate collaboration.* The system needs to provide facilities that encourage collaboration between test data users. These collaboration facilities are intended to foster communication between researchers and to improve the quality of research and dissemination of relevant results.
24. *Visualization.* The system needs to provide visual mechanisms for viewing subsets of the data it stores. This should be available from search or other navigation means and should also present the user with images of geographically referenced data.

4.3.2 Assumptions and Constraints

It is assumed that ingestion would be accomplished with an API. Constraints would be the following:

- *Ingestion constraints.* User-data ingest must support HTTP.
- *Metadata standards.* Metadata must be developed according to a documented standard.

- *System backups.* System backup formats must follow a documented industry standard.
- *Administration security.* Administration functionality must be provided over a secure connection except for those instances when such a connection is unavailable.
- *Interface documentation.* All interfaces and backdoors must be documented.
- *Parallelism.* The system must allow multiple files to be accessed simultaneously.
- *System hardware constraints.* The system's environmental footprint must be quantifiable in terms of power, floor space, and cooling if a dedicated (non–cloud-based) solution is chosen.

4.4 High-Level System Features

Not simply a data warehouse, the SHRP 2 Archive does much more. It is a user-interactive repository that enables uploading of artifacts, searching with results arranged by list or map, and bulk and subset downloading of artifacts (see Figure 4.2). Also, the Archive system is a user-friendly toolset that facilitates visualizations of user-selected data and collaborations between multiple researchers. Although it is being designed to ingest all artifacts created by the Reliability focus area–related projects, it is purposefully not limited to only those projects. Indeed, the system has the capability to share other researchers' new/transformed data and/or work products of any origin that are related to travel time reliability.

4.4.1 Upload

The SHRP 2 Reliability Archive's ingestion wizard allows Reliability project leaders to upload artifacts along with their related metadata and data dictionaries. The system categorizes the artifacts into two general groups: data sets and non–data sets (e.g., documents, computer codes, video, pictures). Artifacts submitted under each category need to meet certain requirements. Therefore, the producer of an artifact has to preprocess it before submitting it into the system.

4.4.2 Search and Download

The Archive provides faceted and text search tools to help users look for artifacts. The text search feature enables users to conduct content search within artifacts and metadata. In addition, the faceted search tool allows users to explore the Archive's repository. Users can filter the search results by selecting various related criteria. The system provides the search results on a map (Figure 4.3) and in a list.

Once archived artifacts are found, users can download the desired objects along with their metadata and metadata documents (if available). Users also can download a subset of a data set.

4.4.3 Visualization

The Archive system accommodates visualization schemes that provide valuable information interactively to users. Users will benefit from two types of visualization tools: map visualization and data visualization.

4.4.3.1 Map Visualization

The system has the capability to map the geolocation information provided in a data set. The geofilter and map view capabilities can help users explore sensor locations (Figure 4.4).

4.4.3.2 Data Visualization

Users can explore and filter the content of a data set. They can make various 2-D plots (e.g., lines, points, bar, and column)

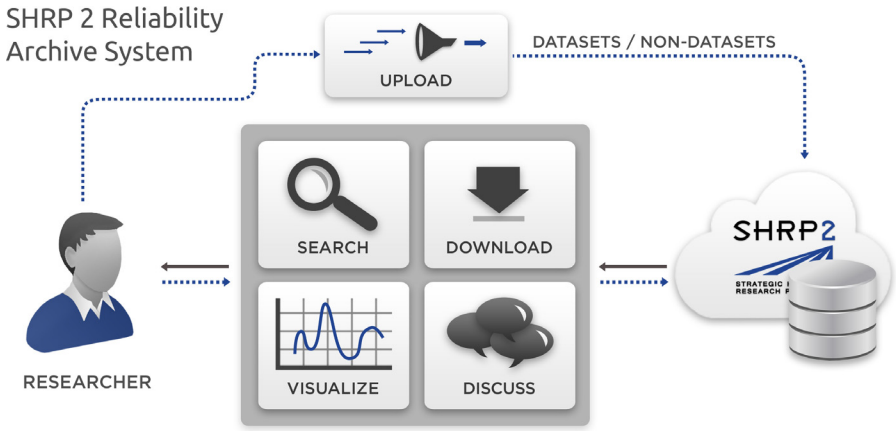


Figure 4.2. Archive features.

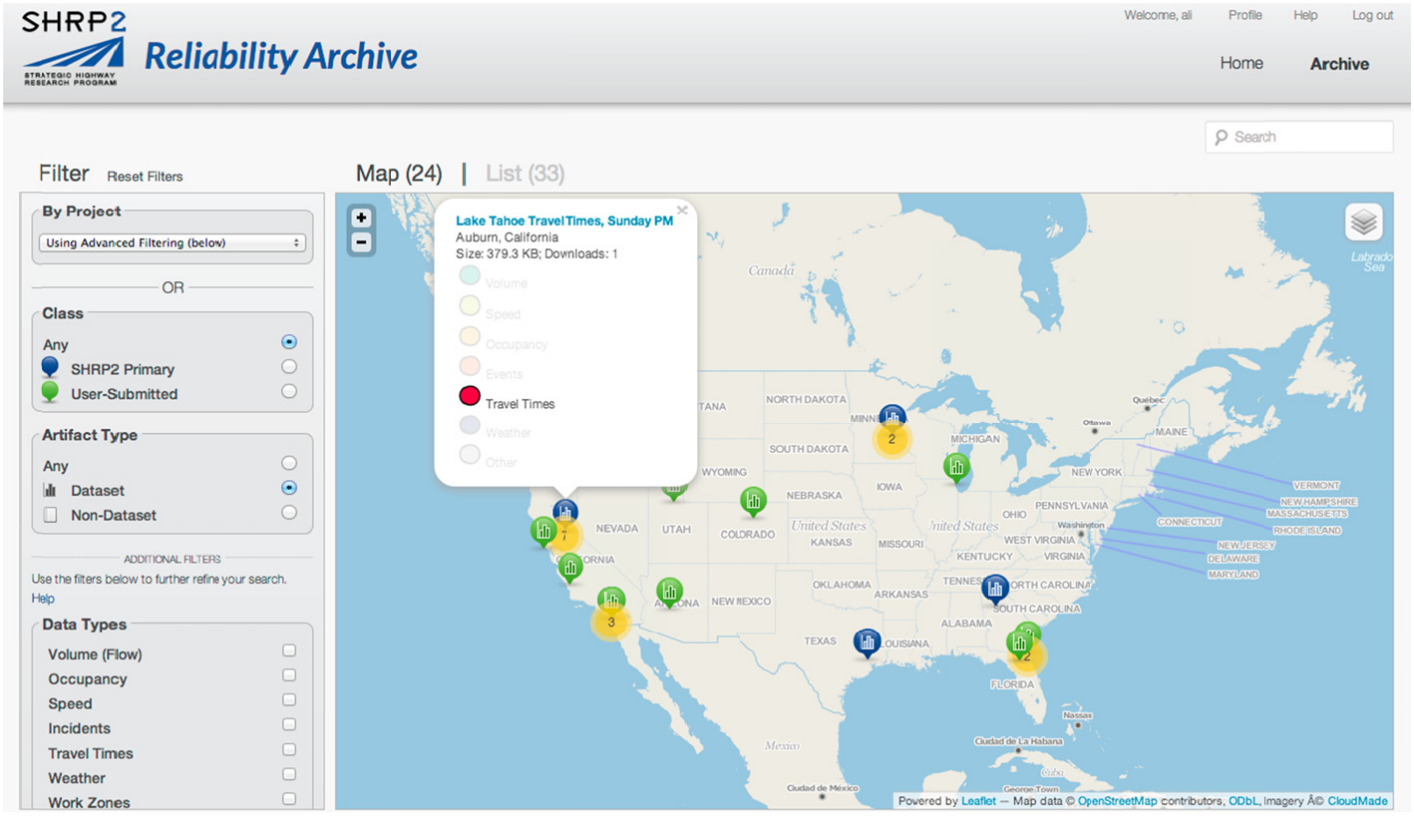


Figure 4.3. Map search.

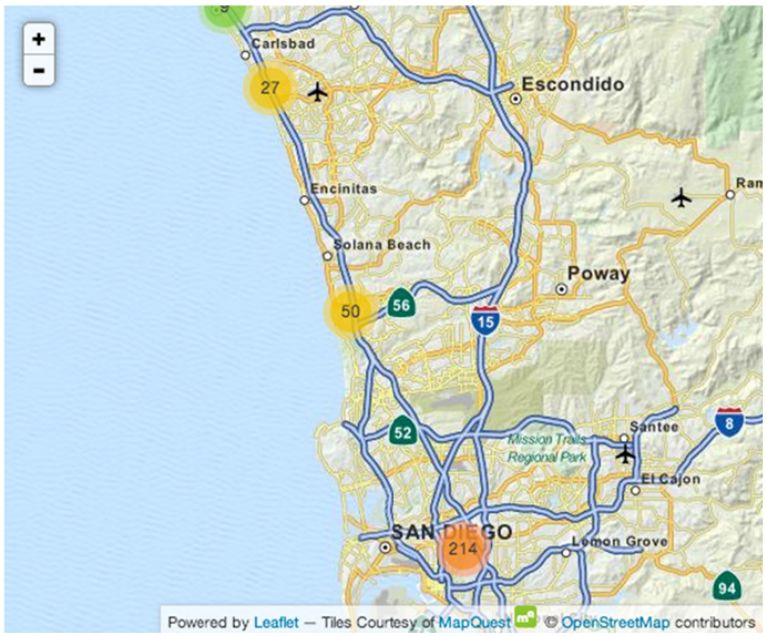


Figure 4.4. Map visualization.

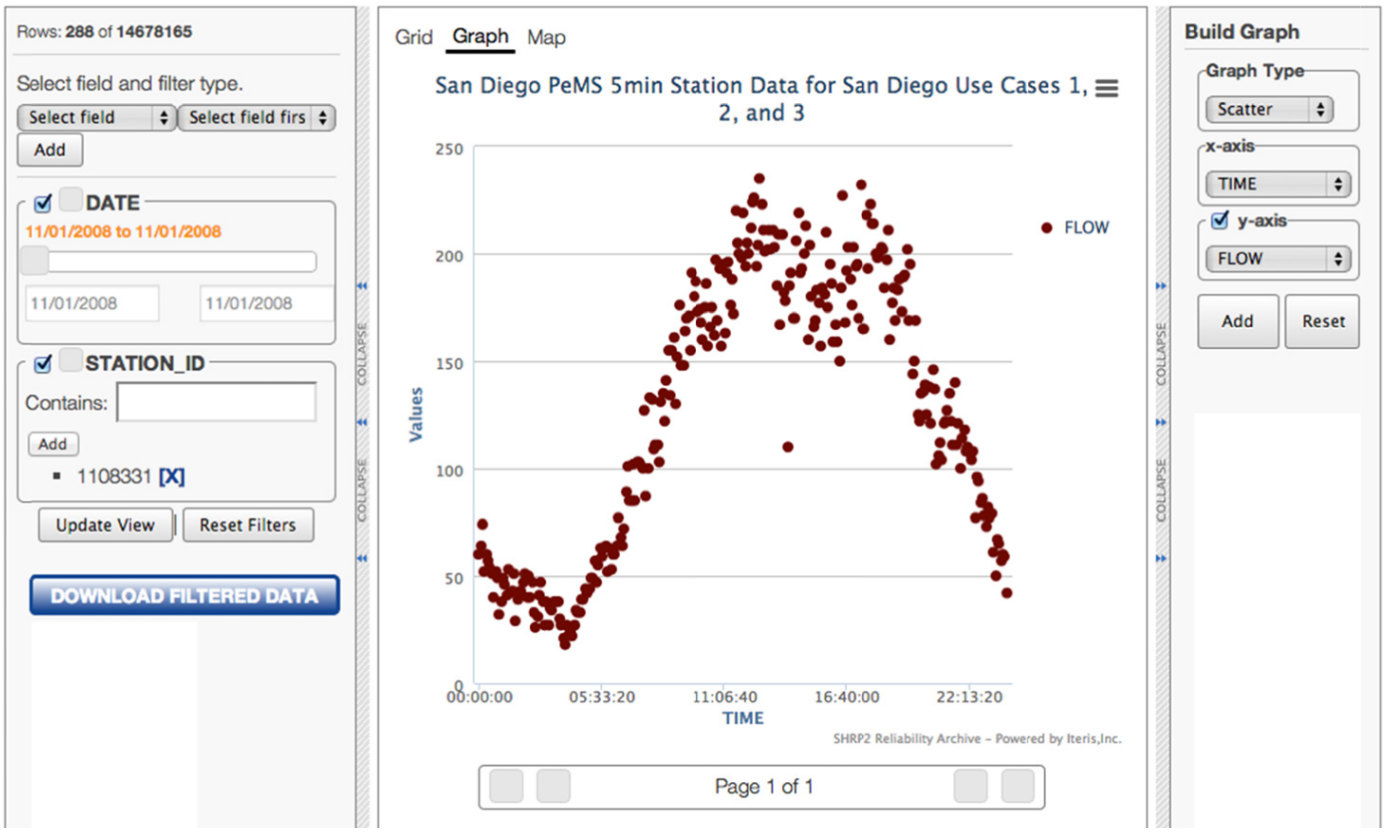


Figure 4.5. Data visualization.

of any fields in the data set or a subset of a data set. The system also has the capability to plot different series on a common plot. Furthermore, users can save and print the plots for future use (Figure 4.5).

4.4.4 Collaboration

Registered Archive users can comment on project pages and artifact pages. They also can rate projects and artifacts. To comment on Archive pages and rate artifacts, users have to be registered.

4.4.5 Administration and User Profile Interface

The Archive administrator is capable of granting access to users as well as adding, editing, and deleting site content. Through the administrator interface, the Archive administrator can also monitor the artifacts being ingested and moderate the comments being submitted by the users.

Using the user profile interface, users are able to change their password, modify and delete the artifacts that they submitted to the Archive, and edit a portion of their profile information.

4.5 Scenarios

Scenarios are documented as user stories. To understand the system from the perspective of a user, user (U) terms are defined as follows:

U1. *Archive user.* The Archive user wants services from the Archive system. These users may provide data to be shared, may download data others have provided, and may comment on their own or others projects and data sets.

U2. *Archive administrator* (administrator access control level). The Archive administrator operates the Archive system, manages data content, including metadata, and handles all IT administrative chores.

U3. *Archive system.* The Archive system provides services to Archive users and responds to commands from the Archive administrator.

U4. *Other archive systems.* Other archive systems interact with the Archive system by providing or acquiring data in automated fashion.

4.5.1 User Stories/Requirements

A *user story* (US) is text written in everyday language that captures what a user does or needs to do as part of his or her job

function. User stories are employed in agile software development methodologies as the basis for defining the functions a business system must provide and to facilitate requirements management. A user story captures the "who, what, and why" of a requirement in a simple, concise way. User stories are written by or for the business user as that user's primary way to influence the functionality of the system being developed.

User stories are written for the Archive system to provide a bridge between needs and requirements. They enable a more intuitive understanding of how users will interact with the system than is traditionally possible when examining requirements.

User stories are traceable to needs, forming the basis for acceptance of the Archive system. The user stories were categorized into seven groups: general, ingestion, administration, search, download, collaboration, and visualization.

4.5.1.1 General

US1. The Archive system provides information about the SHRP 2 Reliability focus area, SHRP 2 Reliability-related projects, and the artifacts associated with each project.

US1.1. The user can search for an artifact at any point while on the website.

US1.2. The user can browse through SHRP 2 Reliability-related projects.

US1.3. The Archive system separately provides information about each Reliability-related project.

US1.4. The Archive system provides general information about the SHRP 2 Reliability program and the Archive system.

US1.5. The Archive system provides a thorough help document to guide the users.

4.5.1.2 Ingestion

US2. The Archive user can submit artifacts to the Archive system.

US2.1. The Archive user submits artifacts to the Archive system along with associated metadata. The user needs to meet the requirements specified in the Archive's ingestion procedures. The Archive system accepts artifacts and associated metadata and storage parameters as noted through the portal interface. The process of preparing and submitting an artifact to the Archive is called the *ingestion process.*

US2.2. During the ingestion process, the user can save the submitted information at any step before continuing to the next step. The user can also save and exit the ingestion process at any time.

US2.3. If needed, the user can also attach any related document that provides extra information about the artifact (e.g., data dictionary).

US2.4. The Archive system provides confirmation to the Archive user of a successful submittal.

US2.5. The Archive system processes the artifacts submitted by the user and logs both successful and unsuccessful submissions.

US2.6. The Archive system notifies the administrator once an artifact is submitted.

US3. The Archive administrator approves/rejects artifacts for archiving.

US3.1. The Archive administrator verifies the content (making sure the content is appropriate) and approves/rejects the artifact. The Archive administrator also verifies content of the metadata and makes sure the artifact does not contain PII.

US3.2. On successful submission of the artifact and approval by the administrator, the Archive automatically checks and verifies the artifact (data structure and format).

US3.3. The provider will be notified of the result of the verification process.

US4. The Archive system stores the digital objects.

US4.1. The Archive system stores the artifacts in the permanent storage.

US4.2. The storage archive must be organized so that Archive users can locate digital objects and determine associations between artifacts and projects.

US4.3. The users should be able to visualize selected portions of data sets.

US4.4. The permanent storage archive must protect against data loss.

US4.5. The storage archive must not allow any changes to the SHRP 2 primary artifacts by the Archive user. Only the administrator and creator have the authority to delete a SHRP 2 primary artifact and modify the metadata. The administrator receives a notification when a creator deletes one of his or her artifacts. The archival storage may be physically and logically distributed but must appear to Archive users as one archival system.

4.5.1.3 Administration

US5. The Archive user registers with the Archive system.

US5.1. The Archive user acknowledges a terms of use agreement and provides a userID, password, and contact e-mail. The Archive system creates an account for the Archive user and provides an e-mail confirmation. The Archive user is then able to use the Archive system to view and download artifacts and comment on artifact pages. Registering as a user is only needed to make comments (i.e., read and write access). No registration is needed to search, view, and download artifacts as a guest (i.e., read-only access).

US6. The Archive administrator deletes/updates artifacts.

US6.1. The Archive administrator must be able to delete/replace/update artifacts and their metadata.

US7. An Archive PI cannot permanently delete SHRP 2 primary artifacts.

US7.1. Under no circumstances can an Archive user permanently delete SHRP 2 primary artifacts. A PI can only request permanent deletion.

US8. An Archive PI updates only artifacts that he or she submitted.

US8.1. A PI must be able to update metadata.

US8.2. A PI can request deletion of an artifact.

US9. The Archive administrator manages added capacity.

US9.1. The Archive administrator adds additional storage capacity to the Archive system. The addition of the storage capacity will not disrupt any ongoing Archive system operations. The additional storage capacity must be available to use immediately. If storage capacity must be replaced, it must be possible to move data off of the storage capacity to be replaced and onto another storage media.

US10. The Archive system protects the artifacts against power disruptions and failures.

US10.1. The archive system protects the digital objects in the permanent storage archive against power disruption and failure of a single storage component (e.g., one hard drive).

US11. The Archive system replicates the artifacts.

US11.1. The Archive administrator is responsible for configuring the parameters of artifact replication. The Archive system must be able to replicate artifacts to one or more additional storage instances to provide resilience to site-level disasters.

US12. The Archive system backs up the artifacts.

US12.1. The Archive system is able to back up artifacts to one or more additional storage instances. The Archive system can provide a subset of artifacts to back up based on various parameters (e.g., timescale, names, area). The Archive system must provide the ability to save system configuration and system metadata to reconstitute system recovery and reconstruction.

US13. The Archive administrator restores digital objects from backup.

US13.1. The Archive administrator is able to restore digital objects and their metadata from a backup.

US14. Users can create an account.

US14.1. Users can sign up to the system by providing their name and a valid e-mail address.

US15. The Archive administrator manages user accounts.

US15.1. The Archive administrator is able to create, modify, and delete Archive user accounts.

US16. The Archive administrator manages user permissions.

US16.1. The Archive administrator sets up the user permissions for users that have responsibilities with the Archive system. This includes the ability to submit SHRP 2 primary artifacts or user-submitted artifacts only. The Archive administrator also can grant full control of the system to another user.

US17. The Archive administrator manages the Archive system.

US17.1. The Archive administrator configures the system and manages the Archive activities of digital object ingestion and metadata modification. These activities include

- Starting and stopping Archive system services;
- Updating and patching application and system software;
- Moderating content;
- Providing customer service;
- Controlling access; and
- Verification.

US18. The Archive system determines and maximizes performance of the system.

US18.1. The Archive system is responsible for determining and then maximizing the overall performance of the Archive system. To verify this, performance is monitored and recorded, performance-related actions are initiated and recorded, and subsequent performance is monitored and recorded. The Archive administrator will be able to access the system performance report.

US19. The Archive system can be easily expanded without service interruption.

US19.1. The Archive system's storage must be expandable while maintaining services to Archive users and the Archive administrator.

US20. The Archive administer defines a project and focus area.

US20.1. The Archive administrator defines a project and focus areas within the Archive system. This includes setting up the project namespace, location in the Archive, and directory structure.

4.5.1.4 Search

US21. The Archive user can search and discover Archive artifacts in the Archive system.

US21.1. The Archive user can search for artifacts on the following criteria:

- Metadata entry,
- Content of documents, and
- Location associated with artifacts within their metadata.

4.5.1.5 Download

US22. The Archive user can download Archive objects.

US22.1. Once archived artifacts are found, the Archive user can download the archived objects along with headings metadata.

US23. The Archive system can compress artifacts.

US23.1. The Archive system must have the capability to compress digital objects. This feature will be used for downloading large data sets.

4.5.1.6 Collaboration

US24. An Archive user may comment on projects.

US24.1. A registered Archive user can provide comments on a project page within the Archive system. The comments can be part of a comment thread. The comments contain the Archive user's name for reference.

US24.2. A registered Archive user can rate a project or an artifact and provide feedback.

US25. The Archive administrator may delete comments on a project.

US25.1. The Archive administrator has the capability to delete comments on a project.

US26. An Archive user may comment on an artifact within a project.

US26.1. An Archive user can provide comments on an artifact page within the Archive system. The comments can be part of a comment thread. The comments contain the Archive user's name for reference.

US26.2. An Archive user can rate a project or an artifact and provide feedback.

US26.3. A registered Archive user can contact the administrator to report an inappropriate artifact. The administrator will be notified. The administrator needs to review the content immediately.

US27. The Archive administrator may delete comments on an artifact.

US27.1. The Archive administrator has the capability to delete comments on a project or an artifact.

4.5.1.7 Visualization

US28. The Archive system allows an interface for visualization of the digital objects.

US28.1. The Archive system needs to be able to provide an interface for visualization of the digital objects that can be used by a third-party front-end visualization application.

US29. The Archive user can select and manipulate the visualization of Archive artifacts in the archive system.

US29.1. The Archive system accommodates visualization schemes that provide valuable information interactively to the users. The users will benefit from three types of visualization tools: grid, graph, and map visualization.

Filter

US29.2. Users can filter data columns using number and text filters. For a given data set, filters can be activated on one or more columns simultaneously. The number filter allows users to specify a value or ranges of values for filtering.

US29.3. Once filtered, users can download the filtered subset of the data file.

Grid view

US29.4. The user can explore the content of the data and can sort a data set by fields (columns) in ascending or descending order.

US29.5. In grid view, the user can show and hide fields of data and arrange the fields on screen for easy viewing. Users can scroll down a page and navigate to other pages within the data set.

Graph view

US29.6. The user can make 2-D plots of any fields in the data set or any subset of the data set.

US29.7. Users can choose the type of graph to plot from the following options: line, spline, area spline, column, bar, pie, scatter plot.

US29.8. To allow comparison of different fields, the user can choose one field to be displayed on the x-axis and up to five fields to be displayed on the y-axis. Users can show/hide each data field displayed on the y-axis and remove any y-axis field from the plot.

US29.9. The user can print the graph and save the graph as one of the following formats: PNG image, JPEG image, PDF document, or SVG vector image.

Map view

US29.10. If geoinformation is provided on the metadata, the user can search for artifacts on the map.

US29.11. If geodata is included in a data set, the Archive system is capable of displaying the references on the map.

CHAPTER 5

Artifact Upload

5.1 Types of Artifacts Archived

The Archive is designed to collect all project related data as shown below.

5.1.1 Data Sets

- Include, for example, traffic engineering data such as travel time, flow, and occupancy (an example data set is shown in Figure 5.1).
- Must be in comma-separated values (.csv) format.
- Can be used for visualization.
- Can be queried.
- Include metadata. (While metadata is needed for all files, it is especially important for data sets. Metadata must correctly identify every column of data within the file and precisely locate the geographical location where the data were collected. Otherwise, the data are less usable.)
- Require special and general metadata.
- Require a data dictionary.

5.1.2 Non–Data Sets

- Include, for example, documents, computer codes, simulation models, spreadsheets, presentations (see Figure 5.2).
- When documents, must be in .pdf format.
- Require only general metadata.
- Are not for visualization.
- Include Excel spreadsheets.
- Require data dictionaries only for Excel spreadsheets.

5.2 Artifact Ingestion Process

An artifact passes through different steps from the time it is being collected from the researchers until it becomes available to the Archive user. Figure 5.3 summarizes these steps.

5.2.1 Step 1. Artifact/Metadata Collection

In this step, the person responsible for uploading collects project artifacts provided by researchers, along with their associated metadata and data dictionary (if needed). For data sets, providing a data dictionary is mandatory. A standard data dictionary template has been developed to help the researcher provide required information related to data sets.

5.2.2 Step 2. Preparation

Each artifact needs to meet some basic requirements before being uploaded into the Archive. There are two types of requirements: general and specific. The general requirements are common for all types of artifacts. For instance, file size should not exceed 1 GB. Specific requirements apply only to data sets and are mandated by database and visualization tool constraints. They are checked by the system during the upload process. The creator of the artifact and the person uploading the artifact are responsible for making sure that the upload criteria are met (e.g., number of columns, column type, column name, location information, data/time formatting). The Archive rejects any artifact that fails to meet the requirements. Section 5.6 provides detailed information on the artifact preparation process.

5.2.3 Step 3. Upload

Using the upload wizard interface (Figure 5.4) PIs, creators, or administrators can upload artifacts and provide metadata information. The interface guides the user through all the upload steps. In this step, the system asks the user to provide appropriate metadata information after the user uploads the file. Some of the metadata fields are mandatory (see Figure 5.5).

If the user is uploading a data set or an Excel spreadsheet file, a data dictionary needs to be attached as well. Also, the

TIME_ID	mean_tt	FROM_NAME
4/22/2011 18:00	3.1	I-80E at Rest Area
4/29/2011 15:00	203.7	I-80W at Kingvale
4/29/2011 15:00	19.3	I-80W at Kingvale
4/29/2011 15:00	248.9	I-80W at Kingvale
4/29/2011 15:00	336.1	I-80W at Kingvale
4/29/2011 15:00	8.2	I-80E at Donner Lake
4/29/2011 15:00	207.3	I-80E at Donner Lake
4/29/2011 15:00	19.9	I-80E at Donner Lake
4/29/2011 15:00	284.8	I-80E at Donner Lake
4/29/2011 15:00	339.9	I-80E at Donner Lake
4/29/2011 15:00	45.1	I-80W at Prosser Village
4/29/2011 15:00	323.7	I-80W at Prosser Village
4/29/2011 15:00	9.3	I-80E at Rest Area

Figure 5.1. Data set example.

user can define column type and modify column labels for data sets. The system produces an error message if this is not completed properly and does not proceed to the next step. For more information, the user may review Chapter 6 or the online Help section.

5.2.4 Step 4. Back-End Processing

On completion of the upload task, the administrator and user receive an e-mail confirming the upload. Then the uploaded artifact appears on the administrator's workflow page under the "Artifact to Be Processed" list. The administrator then needs to review the artifact and accept or reject the upload. Administrator credentials are required to access the page.

The artifact is then processed internally in the back end. This postupload process is called *back-end processing*, in which the artifacts get prepared to support search and visualization features. For security reasons, the administrator needs to approve any further processing of an artifact in the back end by clicking on the "Process" link. This intermediate step gives the administrator the ability to check the artifact

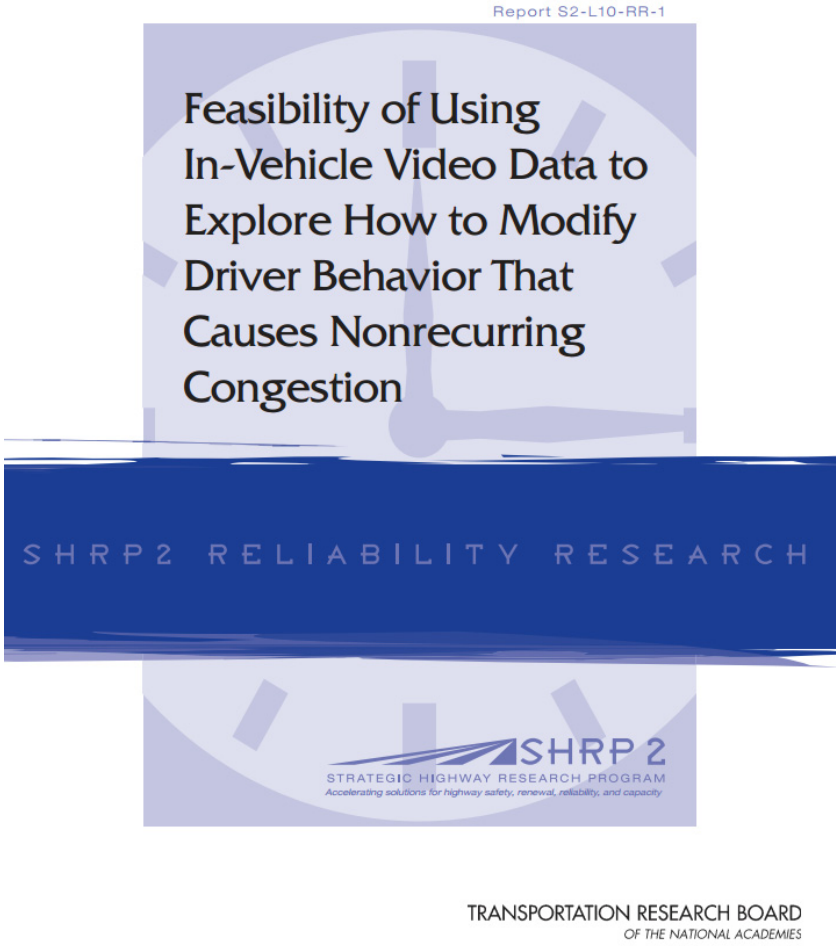


Figure 5.2. Non–data set example.

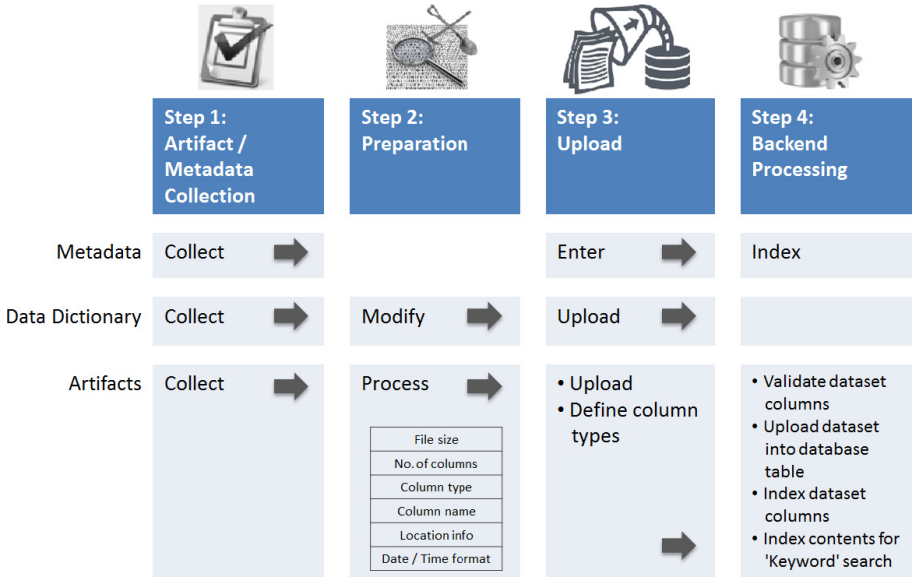


Figure 5.3. Artifact ingestion process.

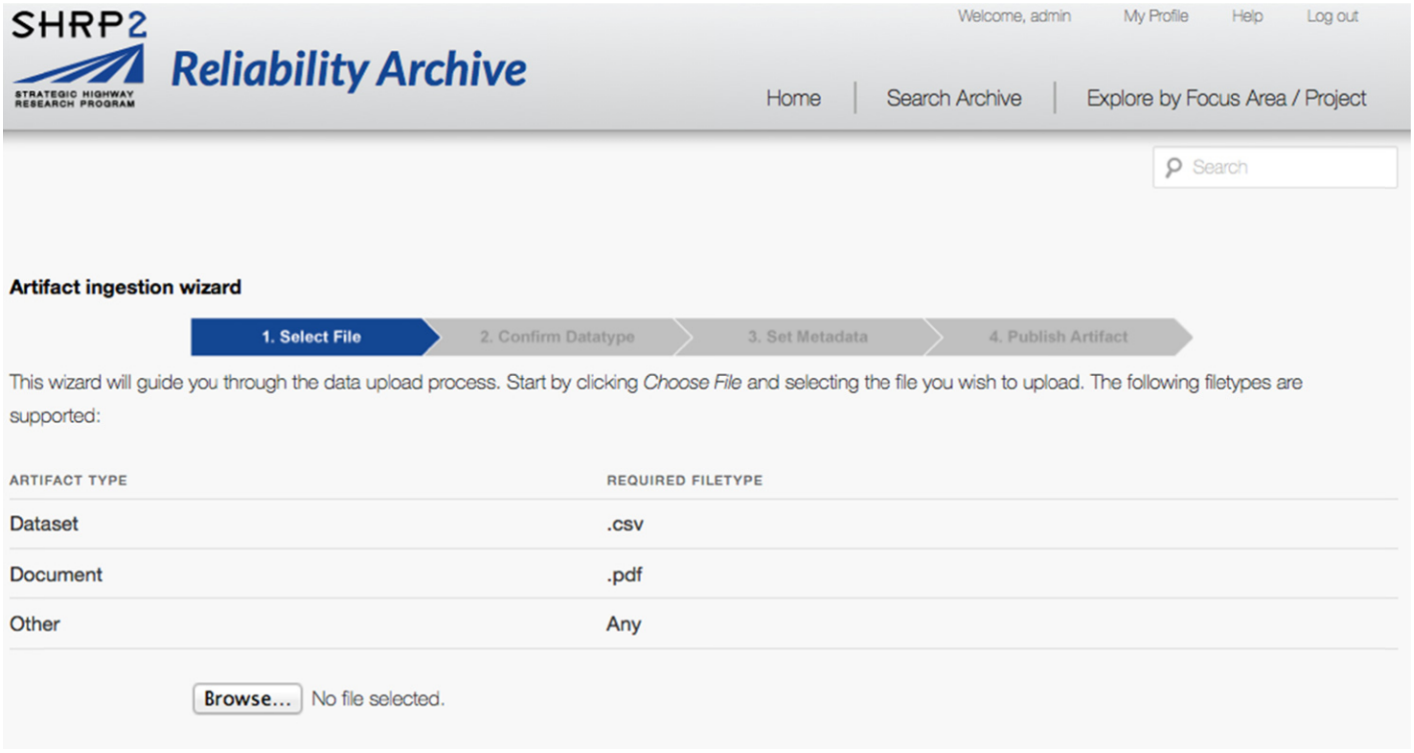


Figure 5.4. Upload wizard.

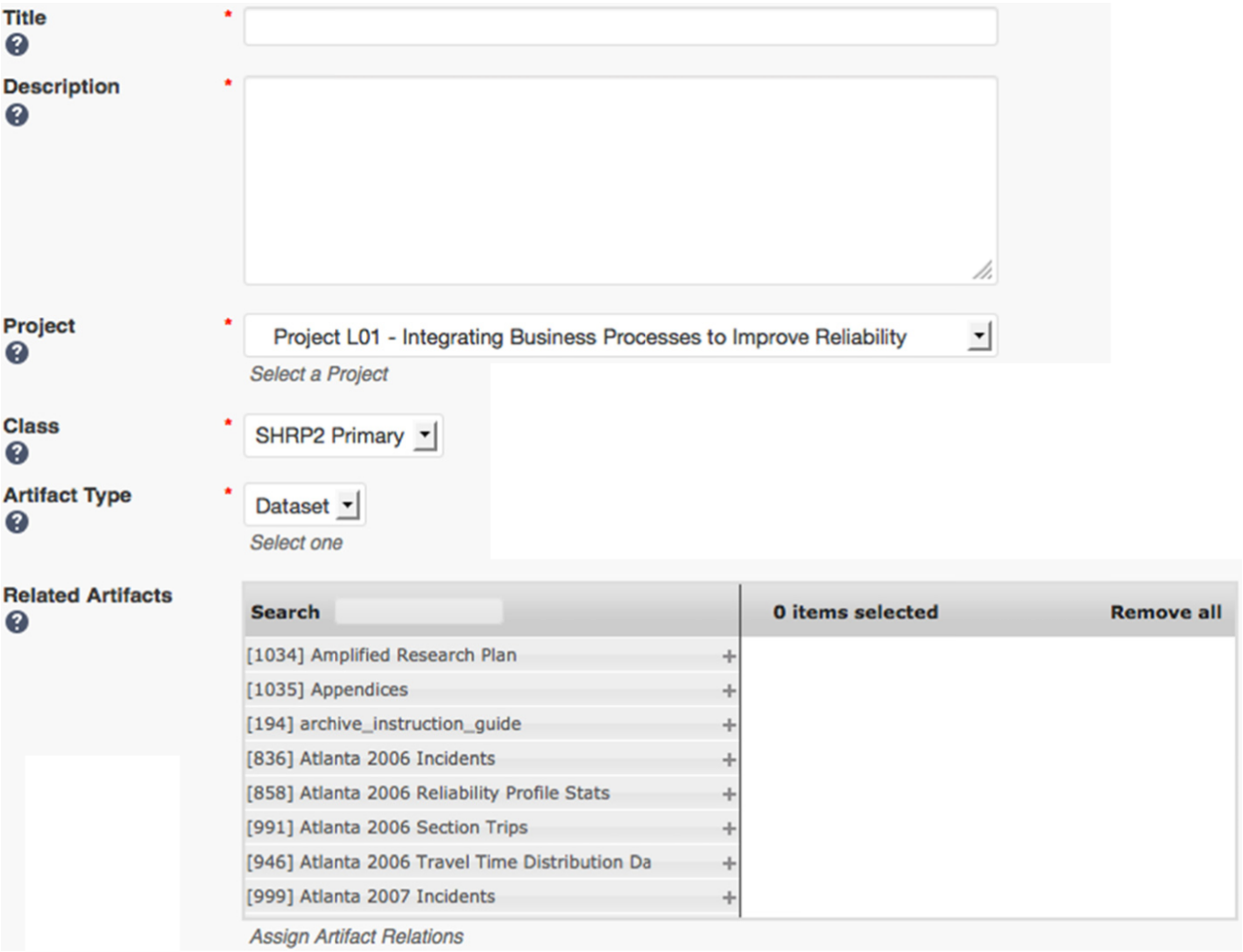


Figure 5.5. Metadata page.

content. The postupload workflow process consists of the following steps:

5.2.4.1 For Data Sets

Step 4.1: Validation. The back end runs a checklist on the data set to make sure that it meets certain criteria that are mandated by database and visualization tool constraints.

Step 4.2: Database upload. Once the data set passes the validation phase, the system starts uploading the fields of the data set into a database table.

Step 4.3: Database indexing. In this phase, the system indexes each column to enable the use of queries.

Step 4.4: Metadata keyword indexing. The system indexes the metadata text for keyword search.

5.2.4.2 For Non–Data Sets

Step 4.1: Keyword indexing of the content. In this step, the system indexes the text content of the artifact to support the full-text search feature.

Step 4.4: Metadata keyword indexing. The system indexes the metadata text for keyword search.

After successful completion of Step 4, the artifact along with its metadata will show up on the Archive webpage.

5.3 Data Dictionary

A data dictionary is a companion document that describes the data stored in the data set. It is a user guide about the data set file. It should contain the following information:

- Data collection methodology;
- Data processing techniques that were applied;
- Column headings for the data set;
- Units of measurements for each column;
- Any other relevant information about the data in each column; and
- Acknowledgment to the people who contributed to creating the data set, such as the road authority that owns the

vehicle detector or individuals/organizations that helped process the data.

Submission of a data dictionary is mandatory along with any data set and Excel spreadsheet.

5.4 Metadata

The most common definition of *metadata* is "data about data." Metadata describes the original data. Metadata in the SHRP 2 Reliability Archive provides information about the artifacts, including title, description, file size, type of artifact, how the data were collected (data sets only), and much more.

Metadata is used throughout the Archive to describe various objects as follows:

- Overall site;
- Focus area;
- Projects;
- Users; and
- Artifacts.

5.4.1 Metadata Relating to the Overall Site

Metadata is used to describe both the structure of the Archive and the artifacts stored within it. Figure 5.6 shows the hierarchical structure of the Archive. The design of the site is flexible and more folders can be added later under the "Focus Area" category, if needed.

Descriptive metadata was attached to each of the site elements—site, focus area, project, artifacts, and user (collaboration)—and this metadata is of critical importance. As part of the metadata scheme design, the L13A team defined element sets, lists of metadata attributes, and relationships that apply to each site element. Attributes are descriptive elements such as title, abstract, and artifact type. Relationships are links between archive elements, such as the link between an artifact and its creator.

For each element in an element set, the project team determined the controlled vocabulary, cardinality (1:1, 1:many, many:1, or many:many), generator (system versus user), whether the element is user-editable, and whether each element is mandatory or optional. Mandatory metadata must be filled in to complete the artifact submission process, while optional metadata can be left blank. Any mandatory or optional metadata that is editable can be updated by a creator; the administrator can later correct any errors or add in missing information. User-generated metadata must be completed by a user (typically the creator); system-generated metadata will be generated by the Archive system, typically by scanning the submitted artifact for embedded metadata. Controlled vocabularies (e.g., a list

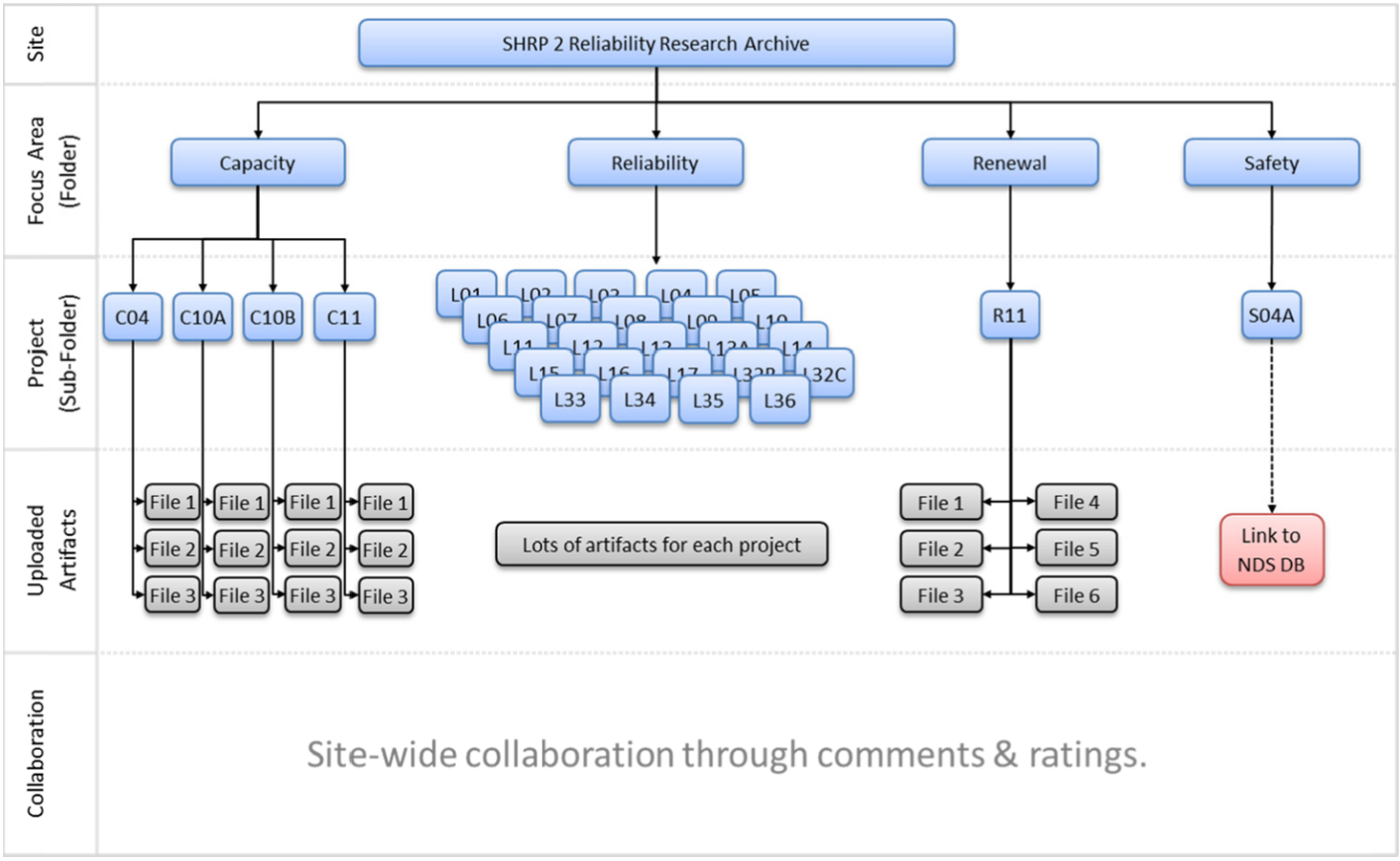


Figure 5.6. Archive site structure.

Table 5.1. Site Metadata

Element Name	Type	Mandatory?	Editable?	Multiple?	Generator	Format	Controlled Vocabulary?
Title	Attribute	Yes	Yes	No	Administrator	Text	None
Description	Attribute	No	Yes	No	Administrator	Text	None
URL	Attribute	Yes	No	No	Administrator	URL	None
Child focus areas	Relationship	Yes	No	Yes	Administrator	Focus area	
Administrator(s)	Relationship	Yes	Yes	Yes	Administrator	User	
Creator	Relationship	Yes	No	No	System-generated	User	
Viewer(s)—registered user PI	Relationship	Yes	Yes	Yes	Administrator	User	
Viewer(s)—registered user	Relationship	Yes	Yes	Yes	System-generated	User	
Viewer(s)—guest user	Relationship	Yes	Yes	Yes	System-generated	User	

of state names to select from) and encoding schemes (e.g., YYYY-MM-DD format) prevent unintended metadata entry errors and help ensure that artifacts can be found by users using the search functionality of the system.

Site, focus area, and project metadata was entered into the system by the L13A project team administrator as part of the Archive software development process. Before the Archive system went live, the SHRP 2 team reviewed the site, focus area, and project metadata. Comments and requested changes were submitted to the L13A team. Similarly, project metadata was reviewed by the relevant team, and comments with any requested changes were submitted to the L13A team. The L13A team responded to the comments and made final changes to the site.

5.4.2 Site Metadata

Table 5.1 provides a brief summary of the site metadata elements.

5.4.3 Focus Area Metadata

Table 5.2 provides a brief summary of the focus area metadata elements.

5.4.4 Project Metadata

Table 5.3 provides a brief summary of the project metadata elements.

5.4.5 User Metadata

User metadata is handled within each user's account profile. The only required elements of a user profile are a userID, password, e-mail, and display name. All other elements of a user profile are optional or set by the site administrator (e.g., site roles). Table 5.4 provides a brief summary of the user metadata elements.

Table 5.2. Focus Area Metadata

Element Name	Type	Mandatory?	Editable?	Multiple?	Generator	Format	Controlled Vocabulary?
Name	Attribute	Yes	Yes	No	Administrator	Text	Yes
Description	Attribute	No	Yes	No	Administrator	Text	No
Date created	Attribute	Yes	No	No	System-generated	Date-time	Yes
Date modified	Attribute	Yes	No	No	System-generated	Date-time	Yes
Child project(s)	Relationship	Yes	Yes	Yes	Administrator	Project	
Administrator(s)	Relationship	Yes	Yes	Yes	Administrator	User	
Creator	Relationship	Yes	No	No	System-generated	User	
Viewer(s)—registered user PI	Relationship	Yes	Yes	Yes	Administrator	User	
Viewer(s)—registered user	Relationship	Yes	Yes	Yes	System-generated	User	
Viewer(s)—guest user	Relationship	Yes	Yes	Yes	System-generated	User	

Table 5.3. Project Metadata

Element Name	Type	Mandatory?	Editable?	Multiple?	Generator	Format	Controlled Vocabulary?
Name	Attribute	Yes	Yes	No	Administrator	Text	Yes
Title	Attribute	Yes	Yes	No	Administrator	Text	No
Background	Attribute	No	Yes	No	Administrator	Text	No
Objectives	Attribute	No	Yes	No	Administrator	Text	No
Research agency	Attribute	No	Yes	No	Administrator	Text	No
Primary investigator	Attribute	No	Yes	No	Administrator	Text	No
Keywords/tags	Attribute	No	Yes	Yes	Administrator	Text	No
Date created	Attribute	Yes	No	No	System-generated	Date-time	Yes
Date modified	Attribute	Yes	No	No	System-generated	Date-time	Yes
Parent site	Relationship	Yes	Yes	No	Administrator	Site	
Parent focus area	Relationship	Yes	Yes	No	Administrator	Focus area	
Child artifact(s)	Relationship	Yes	No	Yes	Administrator	Artifact	
Administrator(s)	Relationship	Yes	Yes	Yes	Administrator	User	
Creator	Relationship	Yes	No	No	System-generated	User	
Viewer(s)—registered user PI	Relationship	Yes	Yes	Yes	Administrator	User	
Viewer(s)—registered user	Relationship	Yes	Yes	Yes	System-generated	User	
Viewer(s)—guest user	Relationship	Yes	Yes	Yes	System-generated	User	

5.4.6 Artifact Metadata

Artifact metadata is fundamental to the Archive's search function and is the supporting information users rely on to determine the applicability and utility of a data set to their needs. Therefore, the quality of this metadata is also very important. However, artifact metadata quality is subject to an important trade-off between producers and consumers. On the one hand, consumers (the Archive users) demand complete and accurate descriptions of each artifact. On the other hand, producers (the users submitting data) have a limited amount of time and resources to devote to metadata gathering and submission. Asking for, or requiring, too much metadata may cause producers to rebel, either by entering poor quality

Table 5.4. User Metadata

Element Name	Type	Mandatory?	Editable?	Multiple?	Generator	Format	Controlled Vocabulary?
Username	Attribute	Yes	Yes	No	Registered user	Text	Yes
Password	Attribute	Yes	Yes	No	Registered user	Text	Yes
Display name	Attribute	Yes	Yes	No	Registered user	Text	No
E-mail address	Attribute	Yes	Yes	No	Registered user	E-mail	Yes
First name	Attribute	No	Yes[b]	No	Registered user	Text	No
Last name	Attribute	No	Yes[b]	No	Registered user	Text	No
Biographical information	Attribute	No	Yes	No	Registered user	Text	No
Site role	Attribute	Yes	Yes	No	Administrator	Role	Yes
Submitted artifact(s)	Relationship	Yes[a]	No	Yes	System-generated	Artifact	
User comment(s)	Relationship	Yes[a]	Yes	Yes	System-generated	Text	No

[a] If applicable.

[b] Editable only by administrator.

metadata or by abandoning the process altogether (specifically for user-submitted artifacts). Asking for, or requiring, too little metadata will not provide enough information to help consumers find the data they want. The list of requested metadata and the metadata submission method represent a compromise between these two competing interests.

The artifact submission UI collects a limited amount of metadata and any remaining metadata is uploaded as a separate document (e.g., a data dictionary or user guide). In this way, the metadata burden is minimized for both submitters and administrators, while still providing valuable information for users of the system.

Requested metadata, both mandatory and optional, depends on the artifact type. The artifacts that have been—or will be—generated by Reliability projects are categorized into two types for the purposes of this archive:

1. Data sets. These are structured data sets in .csv file format.
2. Everything else. Documents fall within this category.

Table 5.5 provides a brief summary of the elements in each set; Table 5.6 summarizes the additional metadata requirements for data sets.

Table 5.5. General Metadata Element Set

Element Name	Type	Mandatory?	Editable?	Multiple?	Generator	Format	Controlled Vocabulary?
Filename	Attribute	Yes	No	No	System-generated	Text	No
Title	Attribute	Yes	Yes	No	Creator	Text	No
Abstract/description	Attribute	Yes	Yes	No	Creator	Text	No
ID	Attribute	No	Yes	No	System-generated	URL	No
Primary SHRP 2 artifact?	Attribute	Yes	Yes	No	Creator	Text	Yes
Focus area	Relationship	Yes	Yes	No	Creator	Focus area	Yes
Project	Relationship	Yes	Yes	No	Creator	Project	Yes
Artifact type	Attribute	Yes	Yes	No	Creator	Text	Yes
Artifact relation	Relationship	No	Yes	Yes	Creator	Text	No
Location(s)	Attribute	No	Yes	Yes (up to 10)	Creator	City, state	Yes
Latitude, longitude	Attribute	No	Yes	Yes (up to 10)	System-generated	Decimal degrees	Yes
Year range	Attribute	No	Yes	No	Creator	YYYY–YYYY	Yes
Date uploaded	Attribute	Yes	No	No	System-generated	Date-time	Yes
Date last modified	Attribute	Yes	No	No	System-generated	Date-time	Yes
File format	Attribute	Yes	No	No	System-generated	Text	Yes
File size	Attribute	Yes	No	No	System-generated	Number	No
Number of downloads	Attribute	Yes	No	No	System-generated	Number	No
Related comments	Relationship	Yes	No	No	System-generated	Comment	Yes
Review status	Attribute	Yes	Yes	No	System-generated	Number	Yes
Workflow status	Attribute	Yes	No	No	System-generated	Number	Yes
Validation status	Attribute	Yes	No	No	System-generated	Number	Yes
Indexing status	Attribute	Yes	No	No	System-generated	Number	Yes
Administrator(s)	Relationship	Yes	Yes	Yes	Creator	User	Yes
Creator	Relationship	Yes	No	No	System-generated	User	Yes
Viewer(s)—registered user PI	Relationship	No	Yes	Yes	Creator	User	Yes
Viewer(s)—registered user	Relationship	Yes	Yes	Yes	System-generated	User	Yes
Viewer(s)—guest user	Relationship	Yes	Yes	Yes	System-generated	User	Yes

Table 5.6. Data Set Metadata Element Set

Element Name	Type	Mandatory?	Editable?	Multiple?	Generator	Format	Controlled Vocabulary?
Column name	Attribute	Yes	Yes	No	Creator	Text	No
Column index	Attribute	Yes	Yes	No	Creator	Number	Yes
Column type	Attribute	Yes	Yes	No	Creator	Text	Yes
Column label	Attribute	Yes	Yes	No	Creator	Text	No
Latitude 1 column	Attribute	Yes[a]	Yes	No	Creator	Column	Yes
Longitude 1 column	Attribute	Yes[a]	Yes	No	Creator	Column	Yes
Latitude 2 column	Attribute	Yes[a]	No	No	Creator	Column	Yes
Longitude 2 column	Attribute	Yes[a]	No	No	Creator	Column	Yes
Data set dictionary		No	Yes	No	Creator	URL	No
Data source(s)	Attribute	No	Yes	Yes	Creator	Text	No
Data type(s)	Attribute	No	Yes	Yes	Creator	Text	Yes
Corridor(s)	Attribute	No	Yes	Yes	Creator	Text	No
Collection technology(ies)	Attribute	No	Yes	Yes	Creator	Text	Yes
Collection frequency	Attribute	No	Yes	Yes	Creator	Text	Yes
Days of the week	Attribute	No	Yes	No	Creator	Text	Yes
Holidays included?	Attribute	No	Yes	No	Creator	Text	Yes

Note: These data are required in addition to the metadata requirements from Table 5.5.
[a]If applicable.

5.5 Artifact Relationships

An important feature of the Archive is its ability to relate one artifact to another, providing the user with links to other artifacts that may be of interest. Ideally, links between artifacts would be bidirectional. For example, if Artifact A is related to Artifact B (i.e., there is a link to Artifact B in Artifact A's page), then Artifact B also contains a related link back to Artifact A. The database software does not do this automatically, instead relying on PIs or creators to keep track of link relationships.

5.6 Preparing Artifacts for Upload

Preparation is required for every artifact that is uploaded into the Archive (see Section 5.2). This section discusses the preparation work for data sets and non–data sets.

5.6.1 Data Sets

5.6.1.1 Need for Preparation

A standardized format for data sets increases the usability of the research data by future users and maximizes the distribution and impact of the research. In addition to downloading SHRP 2 Reliability artifacts, users of data sets will be able to visualize research data in a grid layout, as a graph, or on a map. They will be able to create filter queries that customize the data set to their needs and quickly preview it before downloading.

Enabling these features in the Archive requires some preparation of the data sets into a standardized format. The system will accept other types of data files (including spreadsheets), although the data will not be classified by the Archive as a data set. Visualization functionality is available only on data sets.

5.6.1.2 Data Set Preparation Checklist

Table 5.7 shows the checklist for preparing data sets.

5.6.1.2.1 Data Set Size Restrictions

Data sets should be less than 500 MB; however, the system will accept data sets as large as 1 GB. Data sets larger than 1 GB should be split into multiple files less than 1 GB each and uploaded separately.

5.6.1.2.2 Data Set Format

Data set files must be in comma-delimited (.csv) format. The first row should contain column names, and each row must contain the same number of fields (i.e., the same number of commas).

It is recommended that the user prepare data sets using a spreadsheet program or database tool, then save as or export

Table 5.7. Data Set Preparation Checklist

Data Set Size	The size of the data set is: (please tick) ❑ Less than 1GB—proceed to Step 2 ❑ Greater than 1GB—contact the Administrator	
Data Set Format	Check that the data set meets the following conditions: ❑ CSV format—see extra information ❑ Has at least 1 column of data ❑ Has less than 60 columns of data ❑ Each row of data has the same number of columns (i.e., same number of commas) ❑ The first row contains header names ❑ Each column has at least one non-null field	
Column Headings	Check that each heading name: ❑ Is between 1 and 80 characters long ❑ Is unique ❑ Contains only permitted characters: a–z, A–Z, 0–9, dash, space and underscore	
Data Type	Check that each column of data conforms to the requirements for that data type:	
	Text fields? Number fields? Date & time fields? Data collection points?	❑ Contains any character (e.g., US101) ❑ Contains only number characters, a minus sign or a period (e.g., –1.1) ❑ Must be in a permitted date/time format—see extra information ❑ Must be accompanied by location coordinates ❑ Latitude and longitude coordinates must be in decimal format

to a .csv file (Figure 5.7). Larger files may require manipulation using a programming language.

5.6.1.2.3 Column Headings

Headings must be between 1 and 80 characters long, unique, and contain only permitted characters, including a to z, A to Z, 0 to 9, dashes, spaces, and underscores. Example column headings may include

- Time;
- Date;
- Volume;
- Speed;
- Travel time;
- Station ID;
- Latitude; and
- Longitude.

The user should try to avoid column names that match SQL reserved words. The system will insert underscore characters if the column name is an exact match. For example "select" is changed to "_select_".

The user should place latitude and longitude columns after the second column of the data set. The first two columns should not include latitude and longitude information.

5.6.1.2.4 Data Type

The user should be methodical about the type of data in each column. The system will process each column as one of the following data types:

- Text: a text string of any length such as "US101";
- Number: an integer (1, 2, 3) or real number (1.1, 1.2, 1.3);
- Date and time format (see below); or
- Latitude and longitude (see below).

5.6.1.2.5 Date and Time Format

Data sets that include date and/or time information must conform to one of the formats shown in Figure 5.8. "Date only" and "time only" fields are preferred.

Example spreadsheet file

Date time	Volume	Travel time	Latittude	Longitude
4/25/2011 17:00	7	248.6	39.311977	-120.495774
4/25/2011 17:01	57	37.5	39.311977	-120.495774
4/25/2011 17:02	5	204	39.311977	-120.495774
4/25/2011 17:03	71	4.1	39.329142	-120.292934
4/25/2011 17:04	9	261.4	39.329142	-120.292934

Example .csv format

Date_time,Volume,Travel time,Latitude,Longitude
4/25/2011 17:00,7,248.6,39.311977,-120.495774
4/25/2011 17:00,57,37.5,39.311977,-120.495774
4/25/2011 17:00,5,204,39.311977,-120.495774
4/25/2011 17:00,71,4.1,39.329142,-120.292934
4/25/2011 17:00,9,261.4,39.329142,-120.292934

Figure 5.7. Examples of file formats.

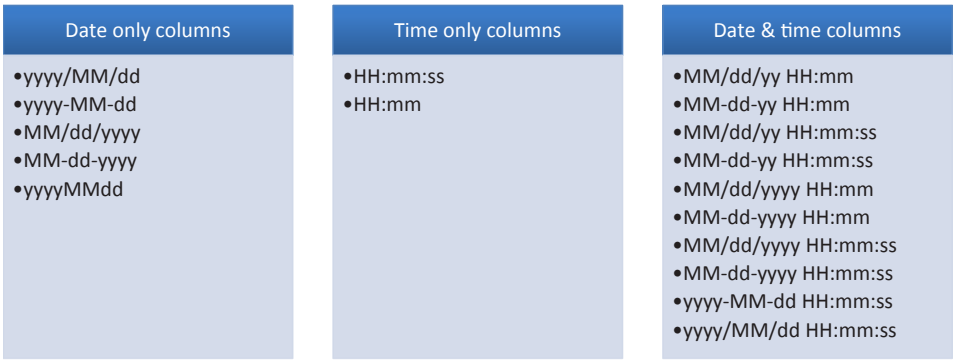


Figure 5.8. Date and time formats.

5.6.1.2.6 Latitude and Longitude Format

Latitude and longitude are used to identify the geographical location of data collection points, such as detector stations. Latitude and longitude values should be in decimal format. The latitude values range from –90 to 90; north is positive and south is negative. The longitude values range from –180 to 180; east is positive and west is negative.

Example: 37.8716667, –122.2716667

5.6.1.3 Common Errors When Preparing Data Sets

The Archive has been designed to allow future users to visualize data sets as grids or graphs and on maps. As such, data should be formatted using the specified convention; otherwise, the data set may not successfully pass through validation or may not be compatible with the visualization features. Solutions to common errors are shown below.

5.6.1.3.1 Incorrect Data Type

- Number columns should contain integers or real numbers only.
- Latitude and longitude should be in decimal format. The following formats will be processed as text or cause a validation error: 41 25 01N, 41°25′01″N.
- Date and time should be in the specified format (see Figure 5.8).

5.6.1.3.2 Missing Values

- Missing numbers or latitude/longitude values are substituted with a zero.
- Missing text values are replaced by no character.
- Missing date/time values will cause an error for the entire column.

5.6.2 Non–Data Sets

5.6.2.1 Artifact Size Restrictions

Non–data sets must be less than 1 GB. Non–data sets larger than 1 GB should be split into multiple artifacts less than 1 GB each and uploaded separately.

5.6.2.2 File Format

Reports and documents should be in .pdf format. Many other file types are accepted. The complete list (in alphabetical order) is as follows:

7z, asc, asf, asx, avi, bmp, c, cc, class, co, css, csv, divx, doc, docm, docx, dotm, dotx, exe, flv, gif, gz, gzip, h, htm, html, ics, jpe, jpeg, jpg, js, m4a, m4b, m4v, mdb, mid, midi, mka, mkv, mov, mp3, mp4, mpe, mpeg, mpg, mpp, odb, odc, odf, odg, odp, ods, odt, oga, ogg, ogv, onepkg, onetmp, onetoc, onetoc2, pdf, png, pot, potm, potx, ppam, pps, ppsm, ppsx, ppt, pptm, pptx, qt, ra, ram, rar, rtf, rtx, sldm, sldx, swf, tar, tif, tiff, tsv, txt, wav, wax, wma, wmv, wmx, wp, wpd, wri, xla, xlam, xls, xlsb, xlsm, xlsx, xlt, xltm, xltx, xlw, zip.

5.7 Supplementary Documents to Assist Principal Investigators and Creators

One of the challenges of gathering project deliverables from over 30 different projects, conducted by a similarly large number of PIs and consultants, is providing consistency in the experience of the Archive's users. For that purpose, the project team and SHRP 2 staff have drafted a series of documents to help PIs complete their task and to foster an organized and internally consistent archive.

- *Data Dictionary Template.* The project team produced the Data Dictionary Template (Appendix A) to help PIs get started on documenting their data set artifacts. This template helps ensure that the format and quality of the metadata are consistent between different data sets, PIs, and SHRP 2 projects.
- *Archive Ingestion and Visualization Guide.* This document was drafted by the project team to provide PIs and Archive users with instructions for uploading artifacts and visualizing data sets. The document contains valuable information regarding the Archive's limitations and the correct column formatting to be used when uploading data sets. The user

guide, in the form of an online Help section, is available on the Archive (http://shrp2archive.org/?page_id=155).
- *SHRP 2 Policy on Software Version Control for Reliability-Related Projects.* SHRP 2 created this policy document to establish a convention for version control pertaining to software developed by contractors within the Reliability focus area. In summary, the document requires that all software contain a descriptive label unambiguously identifying the version of the software. For example, the software convention is as follows: SHRP 2_Softwarename_LXXA_Contractor_Vn.n_dd/mm/yyyy. In addition, a label that is visible and easily read on opening shall be included with software and spreadsheets.

5.8 Quality Assurance After Uploading Artifacts

The team has developed checklists for ensuring adherence to upload requirements and other SHRP 2 guidelines, with the intent of creating an archive with high-quality, consistent, and well-documented artifacts. These checklists focus on listing the steps necessary to conduct a quality check for each type of artifact (data sets and non–data sets) after completion of the upload process.

5.8.1 Checklist for Data Sets

Data sets are the most time-consuming artifact type to upload. They are typically large files with many rows and columns and must be carefully formatted to be correctly processed by the Archive. Once the data set has been satisfactorily processed, the following checks are performed:

1. After downloading and opening the data dictionary PDF file, does the file adhere to the data dictionary template? Does it accurately describe the data set artifact?
2. If the data set contains a data collection stationing column (e.g., "Loop Count Station #"), is there an adjacent column with a related artifact that provides the geographical location of stations (i.e., a stations configuration file) or latitude/longitude coordinates?
3. On the artifact's Data tab, does the grid content load correctly? While it may take a few seconds to load, it should not hang for minutes.
4. If a time and date field is available, does the data grid update readily after filtering?
5. Switching to the Graph subtab and adding an x-axis and y-axis field, does the graph plot the first 300 points appropriately?
6. If the artifact contains latitude and longitude information, after navigating to the Map subtab and using the controls on the Build Map pane to plot latitude and longitude, does the map plot the first 300 points appropriately?

5.8.2 Checklist for Non–Data Sets

5.8.2.1 Documents and Presentations

Microsoft Word, Adobe PDF, and Microsoft PowerPoint files are the quickest items to upload to the Archive. They do not require detailed metadata and can be uploaded in the format provided by the PI. Accordingly, the quality check for this type of artifact is quick:

1. Is the document or presentation a final deliverable? The Archive should only be populated with final versions.
2. After uploading and processing the artifact, then downloading and opening the document or presentation from the Archive, does the file open OK?

5.8.2.2 Spreadsheets and Computer Code

The complexity in uploading and processing spreadsheets and computer code lies between documents/presentations and data sets. Although considered non–data sets, spreadsheets and computer code typically benefit from documentation in the form of metadata, a user guide, or a data dictionary file. The following checks are performed for this type of artifact:

1. Is computer code named in accordance with the software version control policy? If the spreadsheet is a computational tool, it should also be named in accordance with this policy.
2. Does the computational spreadsheet have a readily visible label in accordance with the software version control policy? For software, does it contain a "readme" text file with the label in the main folder directory?
3. Does the computational spreadsheet have secure macros? If so, they must be unlocked before uploading to the Archive to avoid indexing errors.

Metadata and data dictionaries shall be included with all spreadsheets. However, the process for uploading data dictionaries for spreadsheets and computer code is different than that for data sets. While the data set specifically requires a data dictionary when uploaded, for spreadsheets the data dictionary must be uploaded as a separate, non–data set artifact and related to the spreadsheet or computer code using the Related Artifacts tool. A user guide for a software artifact can be provided in a similar manner.

CHAPTER 6

User Guide—Working with the Archive

This chapter reviews the Archive functionalities and explains how to use the Archive. In other words, this chapter serves as a user guide, describing system features and providing a step-by-step guide to using the system. The web address to access the Archive is http://www.shrp2archive.org. The latest version of the user guide can be found as the online Help section at http://shrp2archive.org/?page_id=155.

6.1 Creating and Managing Your User Account

Any user can access the Archive through the Internet. An anonymous user (guest) can search for artifacts, use the visualization tools, and download artifacts—all without logging in. However, to participate in the online discussion forum, the user will need to create an account.

Anyone who is interested in travel time reliability research is eligible to create a user account in the Archive. Creating an account is a quick process, taking 3 to 8 min.

6.1.1 How to Create an Account?

An account can be created by finding the Log In or Register link at the top right of the Archive (Figure 6.1) and completing the following steps:

1. The user clicks Log In or Register.
2. The user follows the link and then clicks Register to open the following registration page.
3. The user completes the form and clicks Register.
4. An e-mail will be sent to the user to validate his or her identity. This e-mail contains a temporary password and a link to the login page.
5. The user enters his or her e-mail address and temporary password into the login screen. The user can change the temporary password if desired.

6.1.2 Update Profile Details

Registered users can update their profile details at any time. Updating a profile involves the following steps:

1. The user logs in to the system and finds the My Profile link at the top right of the Archive (Figure 6.2).
2. The user follows the link and updates his or her details on the My Profile tab.
3. The user clicks Update Profile to confirm the changes. If the user changes an e-mail address, the user will be automatically logged out. To resume activities, the user needs to log in using the updated e-mail address.

6.1.3 Reset Password

Resetting a password involves the following steps:

1. The user finds the link to Log in or Register at the top of the Archive.
2. The user follows the link to the login page and then follows the Lost Password link below the login boxes (Figure 6.3).
3. The user enters the e-mail address used to register, and a new password will be sent to the registered e-mail address.

6.2 Remove an Account

To remove an account, the user contacts the administrator using the feedback page. The user can find the link to the feedback page at the bottom of any page on the Archive website.

The user completes the feedback page (Figure 6.4) and, if possible, includes a reason for the removal of the account. The user then clicks the Send Feedback button to notify the administrator. The administrator will contact the user to confirm the removal of the account.

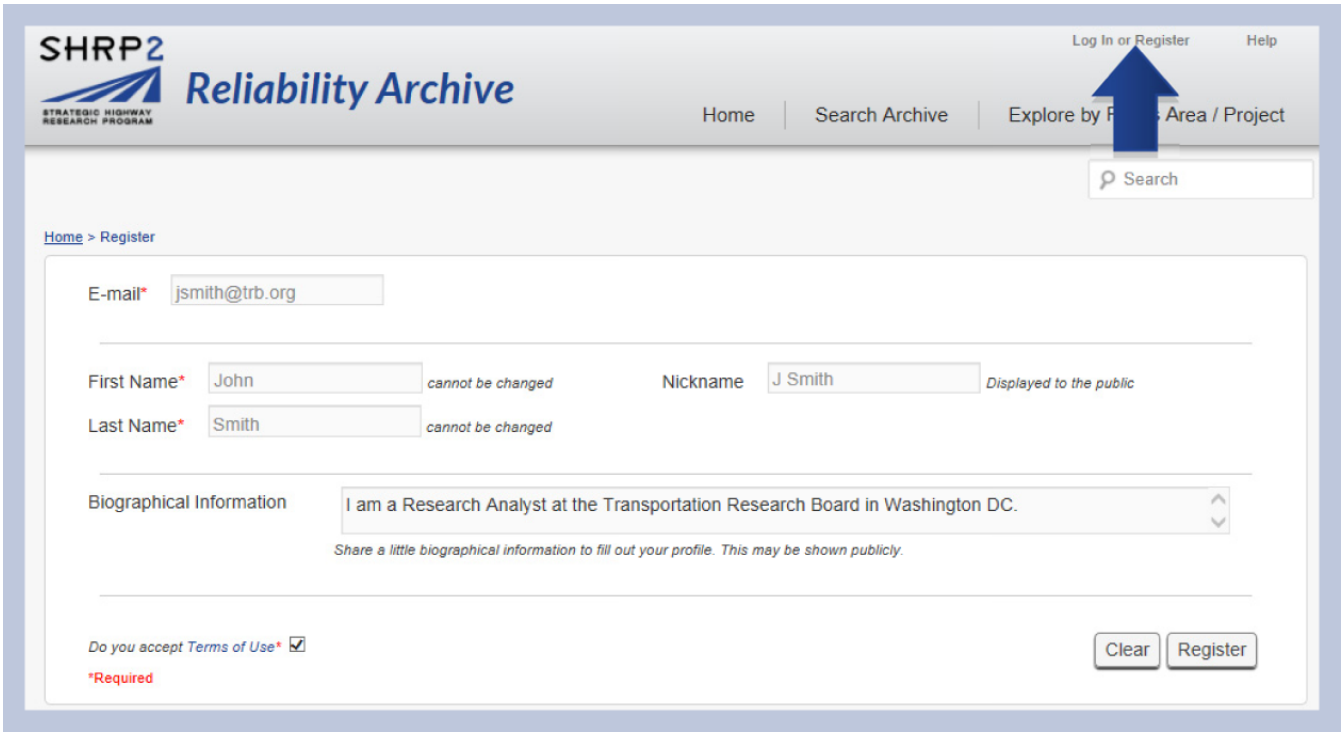


Figure 6.1. Log In or Register link.

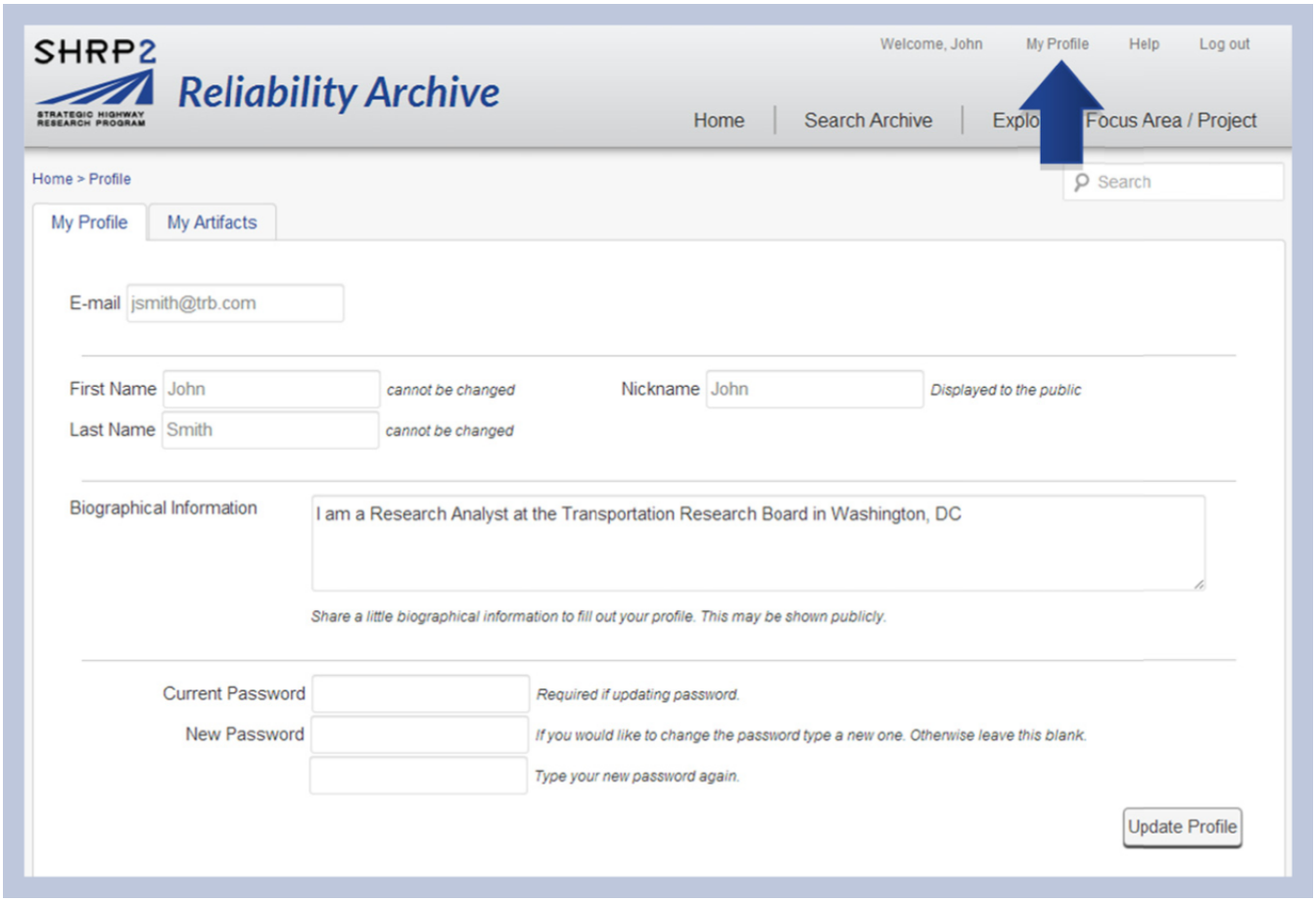


Figure 6.2. My Profile link.

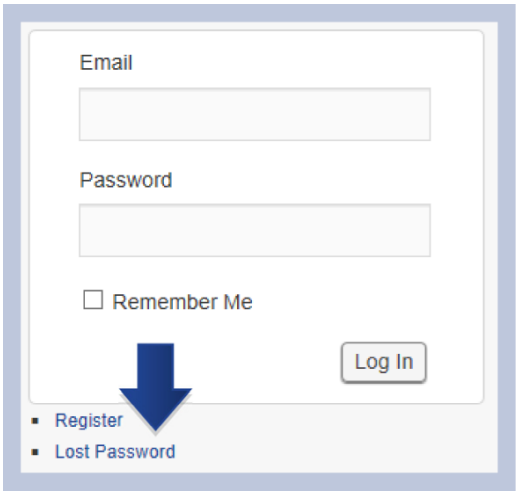


Figure 6.3. Lost Password link.

6.3 Search for Artifacts

6.3.1 Search by Keyword

The Archive includes a search box at the top right of the page that lets the user enter keywords to find relevant artifacts (Figure 6.5). Keywords may include the project title, artifact tags, artifact ID, or location.

6.3.2 Search Archive

The Search Archive feature at the top right of the Archive pages is similar to an advanced search feature in a library

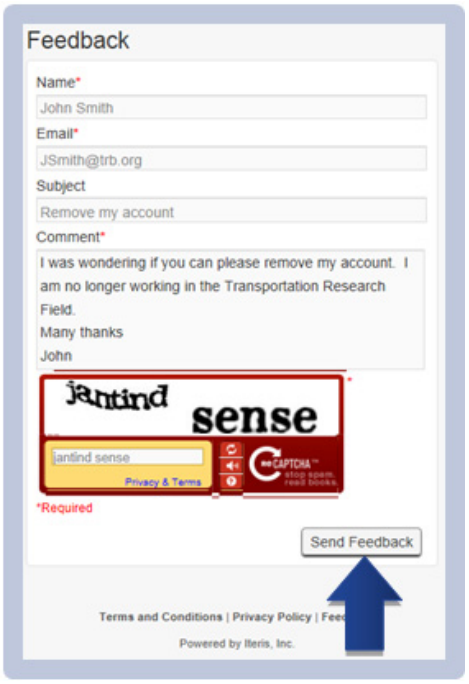


Figure 6.4. Feedback page.

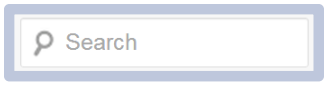


Figure 6.5. Keyword search box.

catalog (Figure 6.6). The user should use this feature when searching through all artifacts to find specific artifacts that meet certain criteria. The user can look for

- Data sets or non–data sets;
- Artifacts on different SHRP 2 projects;
- Data sets with a specific type of data (e.g., volume, occupancy, speed, incidents, travel times, weather information, or work zone information);
- Data sets that were collected using a particular collection technology (e.g., Bluetooth, loop, radar, video);
- Data sets that were aggregated at different time intervals (e.g., 30-s to daily);
- Data sets that were collected on a specific day of the week;
- Data sets that include or exclude holidays; and
- Specific types of non–data sets, including analysis tools, computer code, data dictionaries, final reports, guide books, presentations, RFPs, or surveys.

After users choose conditions to filter the search by, they can display the remaining artifacts geographically on a map or in a list. See below.

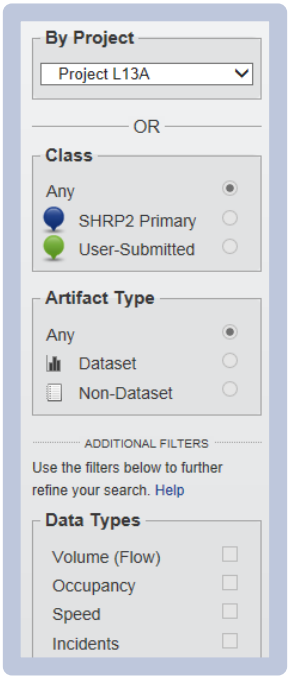


Figure 6.6. Search Archive feature.

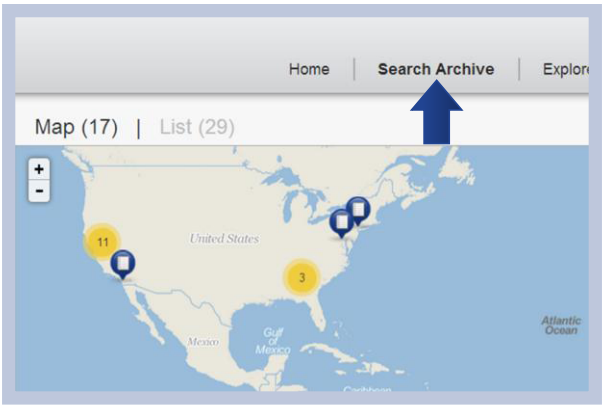


Figure 6.7. Search by geographical location.

6.3.3 Search by Geographical Location

When the user is looking for artifacts at a specific geographical location, the map searching feature is most useful.

To view artifacts on a map, the user can click the Search Archive link at the top right of the page (Figure 6.7). The map shows the locations of all artifacts (if they were submitted with location information). The user can filter the artifacts by project, class, artifact type, and additional filters. To open a data set, the user clicks on it on the map.

6.3.4 Search Through a List of Artifact Titles

Users can use the list view to view a list of artifact titles (Figure 6.8). The list of the results can be sorted and filtered by project, class, artifact type, and additional filters. Artifacts can be opened by clicking the title.

6.3.5 Explore by Focus Area/Project

The user can explore by SHRP 2 focus area and project when interested in viewing all the artifacts produced under a specific SHRP 2 Reliability-related research project. To explore the projects and artifacts by focus area,

1. At the top of an Archive page, the user clicks Explore by Focus Area/Project (Figure 6.9). The user follows this link to a page, which shows all the focus areas.
2. The user finds the focus area of interest and clicks the heading to expand. The user will see a description of the focus area and a further option to expand.
3. The user clicks List of Reliability Focus Area Related Projects to view the list of projects.

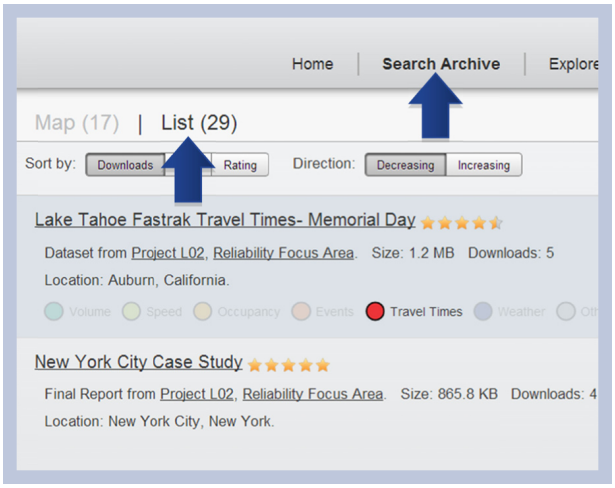


Figure 6.8. List of artifact titles.

6.3.6 View Latest Artifacts

The list of latest artifacts is shown on the Archive homepage (Figure 6.10).

6.3.7 View Your Own Artifacts

Users who are a principal investigator or creator on a project can use this option to view all the artifacts that they have uploaded (Figure 6.11).

1. At the top of an Archive web page, the user who has already logged in clicks My Profile to view his or her profile information.
2. The user chooses My Artifacts to view a list of the artifacts that he or she has uploaded.

6.4 Working with Artifacts

This section provides information about working with artifacts. Once users have found the artifact they are interested in, they can view metadata (or the artifact page), visualize the information in a data set, download the artifact, and contribute to the discussion.

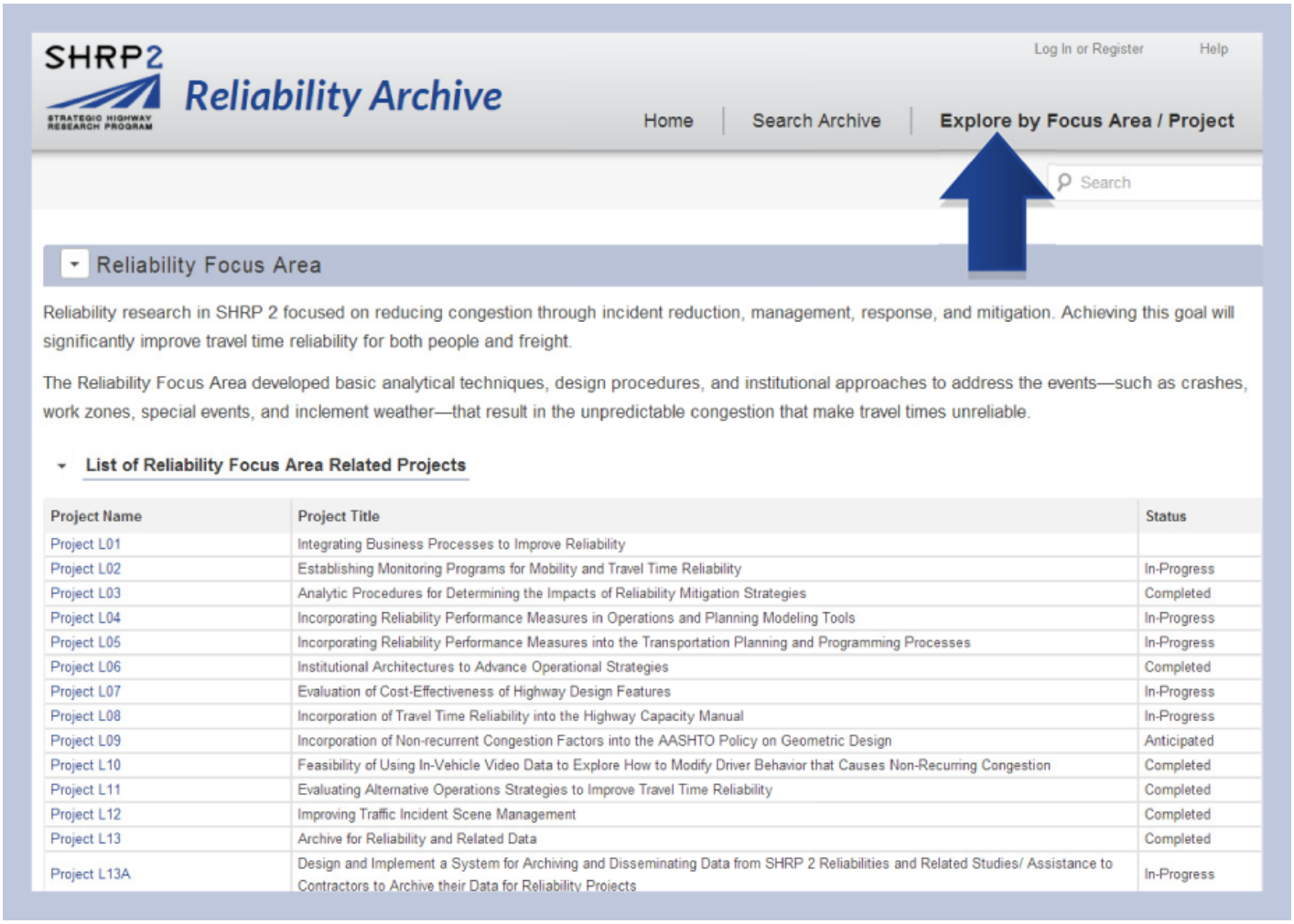
6.4.1 View Metadata

The Metadata tab gives general information about the artifact, including title, description, location, file size, author, type of artifact, and more. If the artifact is a data set, the tab provides a data dictionary and specifies other information such as data sources, road corridors covered, collection technology, and collection frequency.

The original file can be downloaded by clicking Download File at the top right of the Metadata tab. To download the data dictionary, the user can click the data dictionary file name.

6.4.2 Visualize Data

The user can use the Data tab to visualize the data in a data set. There are three ways of visualizing data: in a grid, on a graph, or plotted on a map.



Project Name	Project Title	Status
Project L01	Integrating Business Processes to Improve Reliability	
Project L02	Establishing Monitoring Programs for Mobility and Travel Time Reliability	In-Progress
Project L03	Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies	Completed
Project L04	Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools	In-Progress
Project L05	Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes	In-Progress
Project L06	Institutional Architectures to Advance Operational Strategies	Completed
Project L07	Evaluation of Cost-Effectiveness of Highway Design Features	In-Progress
Project L08	Incorporation of Travel Time Reliability into the Highway Capacity Manual	In-Progress
Project L09	Incorporation of Non-recurrent Congestion Factors into the AASHTO Policy on Geometric Design	Anticipated
Project L10	Feasibility of Using In-Vehicle Video Data to Explore How to Modify Driver Behavior that Causes Non-Recurring Congestion	Completed
Project L11	Evaluating Alternative Operations Strategies to Improve Travel Time Reliability	Completed
Project L12	Improving Traffic Incident Scene Management	Completed
Project L13	Archive for Reliability and Related Data	Completed
Project L13A	Design and Implement a System for Archiving and Disseminating Data from SHRP 2 Reliabilities and Related Studies/ Assistance to Contractors to Archive their Data for Reliability Projects	In-Progress

Figure 6.9. Explore by focus area and project.

Figure 6.10. List of latest artifacts.

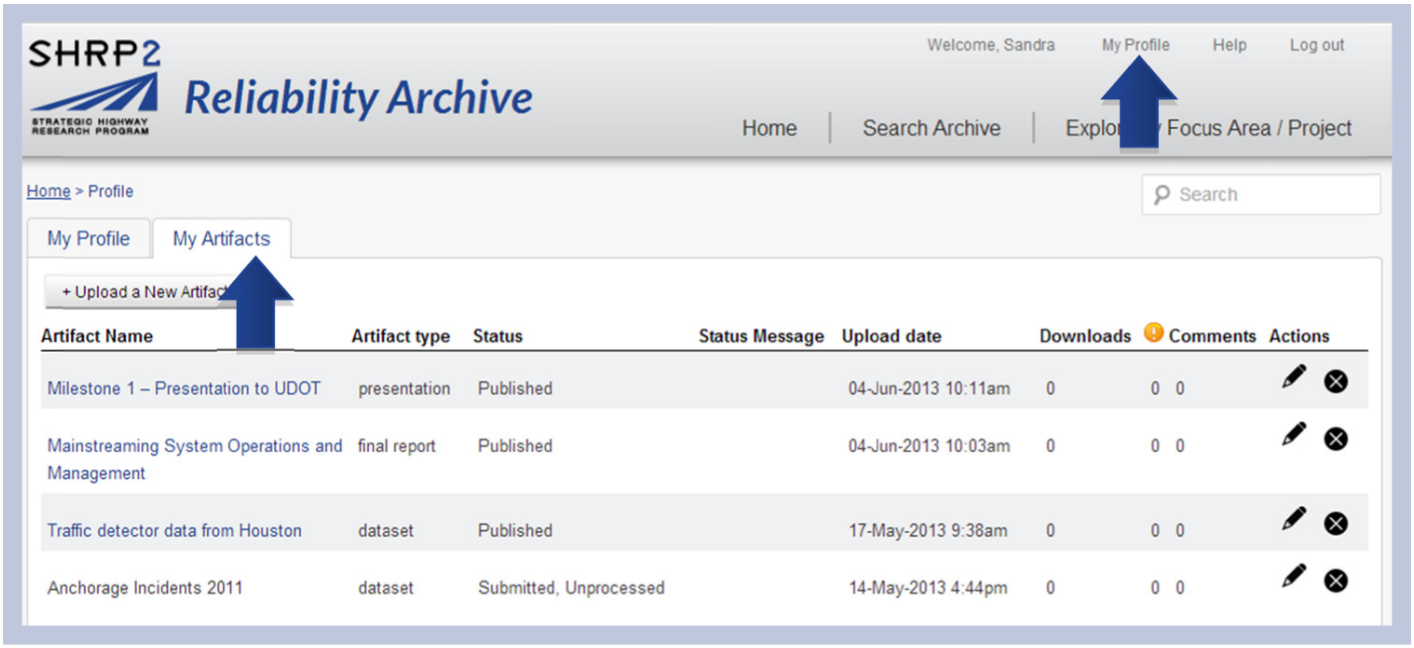


Figure 6.11. My Artifacts tab.

6.4.2.1 General Navigation of Data Tab

Before explaining how to visualize data sets, this section provides some general navigation tips for the Data tab. Users opening this tab will see central, left-, and right-side panels (Figure 6.12).

- *Central panel.* This panel displays the data set visualization. The user can choose to visualize the data set in a grid layout, on a graph, or plotted on a map. At the top of the central panel, the user can choose Grid, Graph, or Map.
- *Left-side panel.* The left-side panel gives all the options to filter a data set using text or number filters. For example, the user can create a filter that finds all the data points with speeds between 40 and 50 mph. For more information about filters, read Section 6.4.2.5.
- *Right-side panel.* This panel displays extra options specifically associated with the visualization type that the user has selected. For example, if the user is graphing, then this panel will give options to build a graph.

The left and right panels can be opened and closed, so that the user can give more space to the main central panel. The

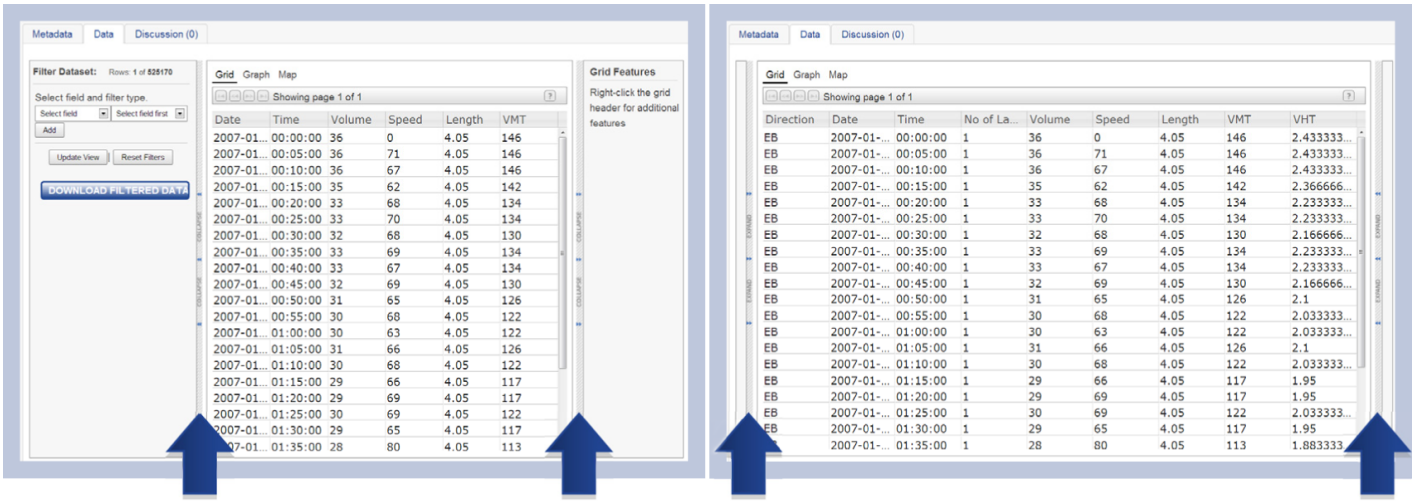
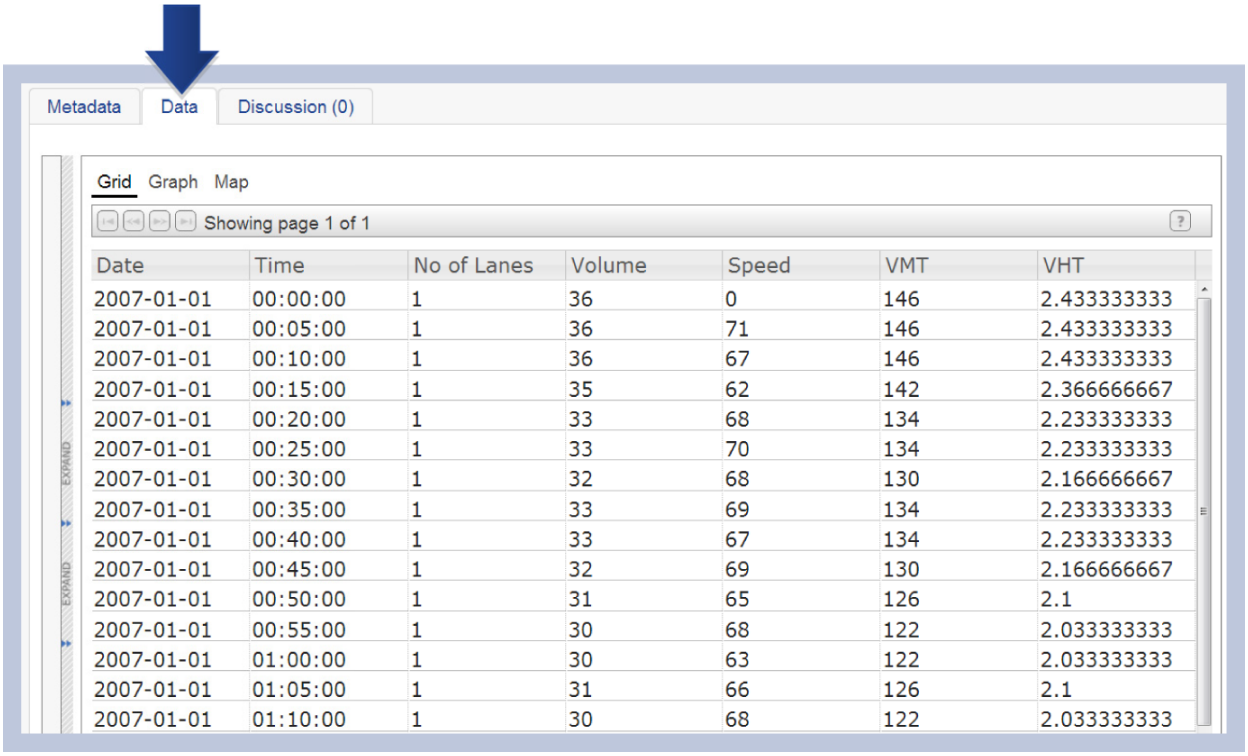


Figure 6.12. General navigation of Data tab.



Date	Time	No of Lanes	Volume	Speed	VMT	VHT
2007-01-01	00:00:00	1	36	0	146	2.433333333
2007-01-01	00:05:00	1	36	71	146	2.433333333
2007-01-01	00:10:00	1	36	67	146	2.433333333
2007-01-01	00:15:00	1	35	62	142	2.366666667
2007-01-01	00:20:00	1	33	68	134	2.233333333
2007-01-01	00:25:00	1	33	70	134	2.233333333
2007-01-01	00:30:00	1	32	68	130	2.166666667
2007-01-01	00:35:00	1	33	69	134	2.233333333
2007-01-01	00:40:00	1	33	67	134	2.233333333
2007-01-01	00:45:00	1	32	69	130	2.166666667
2007-01-01	00:50:00	1	31	65	126	2.1
2007-01-01	00:55:00	1	30	68	122	2.033333333
2007-01-01	01:00:00	1	30	63	122	2.033333333
2007-01-01	01:05:00	1	31	66	126	2.1
2007-01-01	01:10:00	1	30	68	122	2.033333333

Figure 6.13. Grid view.

side panels can be collapsed or expanded by clicking on the vertical gray bars.

6.4.2.2 Grid

The grid view is used when a user wants to see the data in tabular format (Figure 6.13). To view a data set in a grid, the user navigates to the Data tab. By default, the Grid page is then displayed.

To sort the data in ascending or descending order, the user can left click the header row of a column. To show or hide columns of data, the user can right click the header row. On the menu that appears, the user can make selections to show or hide columns (Figure 6.14).

6.4.2.3 Graph

The graph view is used when the user wants to look at the relationships between the various columns of data and have better insight about the data set (Figure 6.15). To view the data set on a graph,

1. The user navigates to the Data tab and then the Graph link.
2. On the right-side panel, the user can build a graph by using the drop-down boxes. The user first selects the type of graph, then chooses data series to plot on the x- and y-axes. The user can edit the columns of data plotted on the graph using the drop-down boxes for the x-axis and y-axis.

The first 300 data points will be plotted on the graph. To plot all the data points, the L13A team recommends downloading the data set and visualizing all of the fields on a local computer.

6.4.2.4 Adding Multiple Series

The graphing tool allows users to plot more than one column of data against the y-axis. For example, users can plot the time of day on the x-axis and then both volume and speed on the y-axis. Five columns of data can be plotted on the y-axis at the same time. Once two or more series are selected, the visualization tool provides new options for displaying the data series (Figure 6.16):

- *Show/hide series.* The user can show or hide each series of data plotted against the y-axis in two ways: (1) click on the legend of the data series, or (2) check or uncheck the tick box above the y-axis drop-down menu.
- *Remove a data series from the plot.* To remove a data series from the plot, the user clicks the [X] next to the y-axis data series.

6.4.2.5 Types of Graphs

The user can choose from scatter, line, spline, area spline, column, bar, and pie graphs (Figure 6.17).

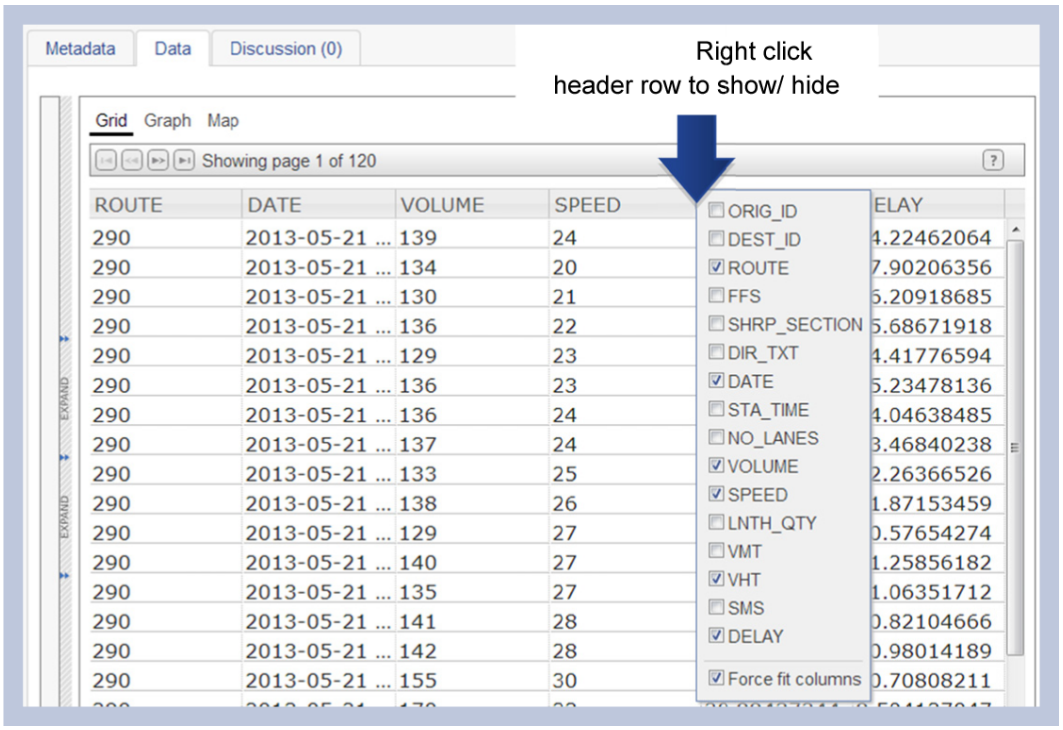


Figure 6.14. Show or hide columns of data.

Figure 6.15. Graph view.

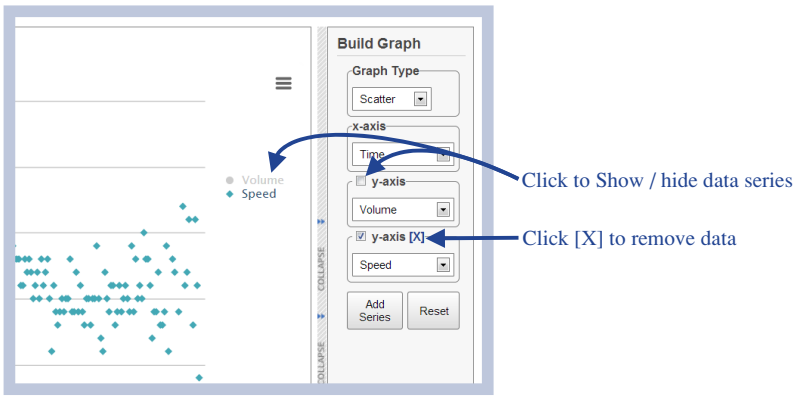


Figure 6.16. Displaying data series.

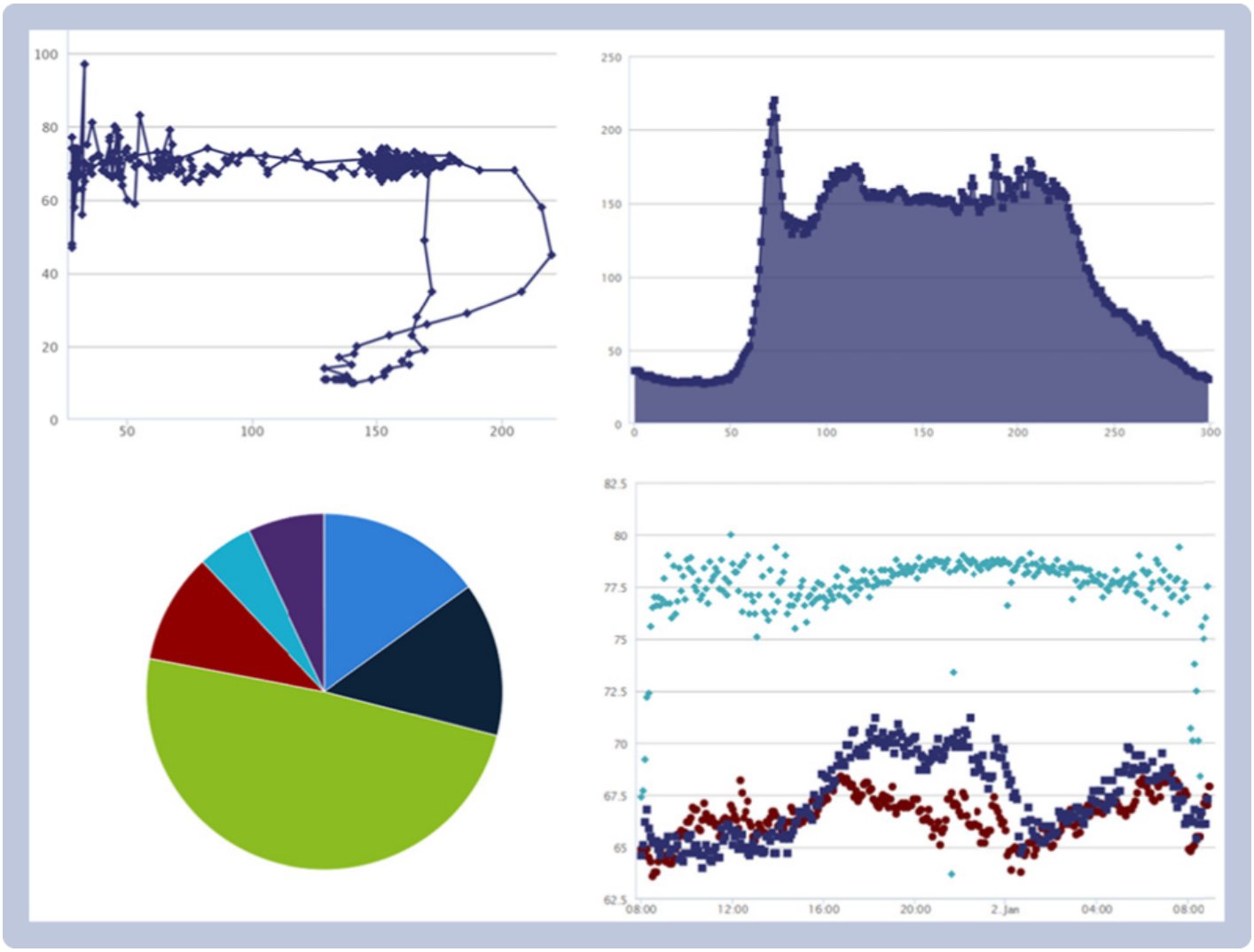


Figure 6.17. Types of graphs.

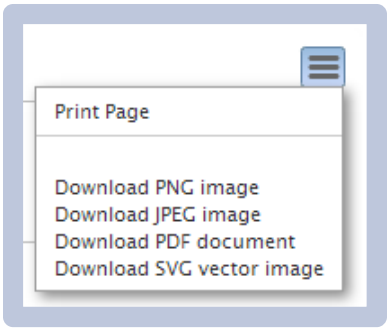


Figure 6.18. Printing and downloading options.

6.4.2.6 Printing and Downloading Graphs

Once a graph is created, the user can print the page or download images of the graphs. Related options can be selected by clicking the icon located at the top right of the graph (Figure 6.18).

The following are constraints in graphing:

- Only numerical values can be graphed.
- Columns of data in a date and/or time format can only be plotted on the x-axis.
- Up to five columns of data may be plotted against the y-axis.

6.4.2.7 Map

The user can use the map view to visualize the location of data collection stations on a map (Figure 6.19). This option is available for data sets with latitude and longitude information stored within the data set.

To view the data set on a map, the user navigates to the Data tab and then the Map link. On the right-side panel, the user can build a map by using the drop-down box to find the latitude and longitude coordinates. If latitude and longitude coordinates are not available, then unfortunately, mapping will not be possible on this artifact. Assuming the coordinates are available, then the first 300 points are plotted on the map.

Options to display are

- *Auto zoom.* The user can check the auto zoom box, and the map will automatically zoom in to the detector locations.
- *Cluster.* The user can check the cluster box to group detector stations together, making the map easier to view. To view the detector stations individually, clear this box (Figure 6.20).

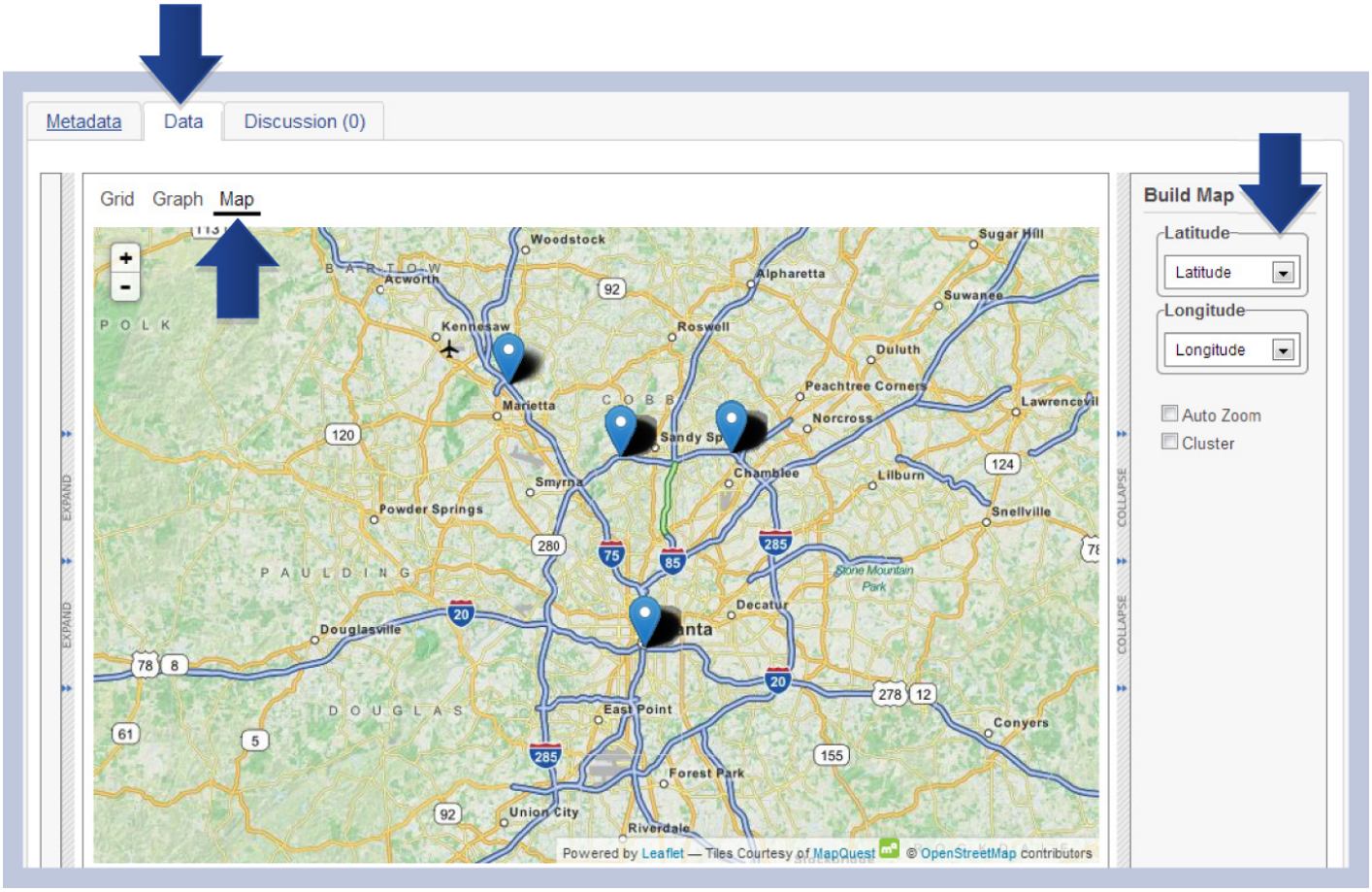


Figure 6.19. Map view.

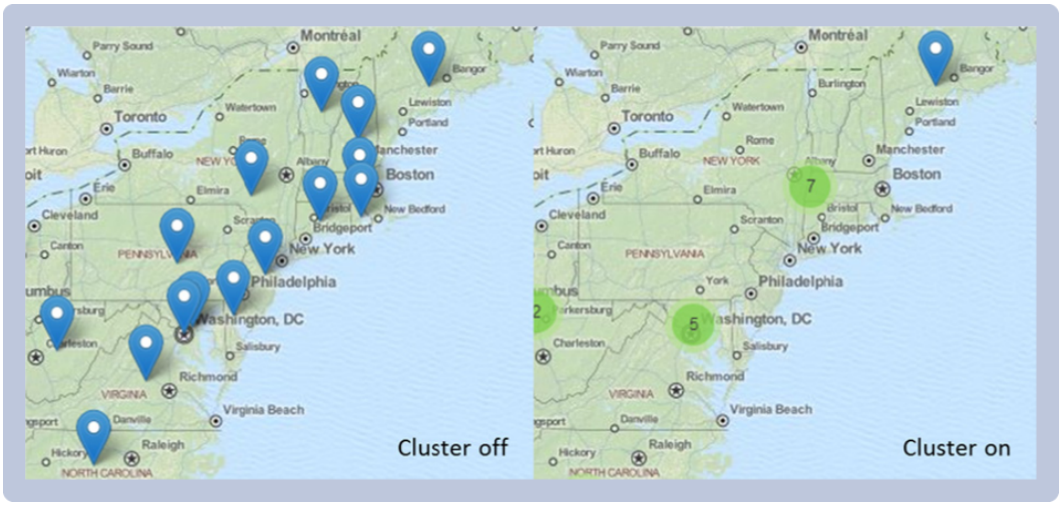


Figure 6.20. Cluster detector stations.

6.4.2.8 Filter

The Archive can accept artifacts which can be tens of millions of rows long. They can contain many years of data from multiple detector stations. The filter function is used to customize the data set to include only the information the user is interested in.

As an example, a user can create a filter query on

- A time column to view data in the a.m. peak between 7 a.m. and 9 a.m.;
- A date column to view data for July 2011 only;
- A speed column to view speeds below the free flow speed (e.g., 0 to 55 mph);
- Latitude and longitude coordinates to view data at one location only; and
- A weather information column, to view data gathered under snowy conditions only.

6.4.2.9 Selecting a Filter

There are three types of filters available:

- Slider filter,
- Number filter, and
- Text filter.

Slider and number filters apply to data columns with numerical values. The text filter can be applied to columns containing text only.

To define the filtering criteria (Figure 6.21),

1. The user selects the field using the drop-down box and then the type of filter desired.
2. The user clicks Add.

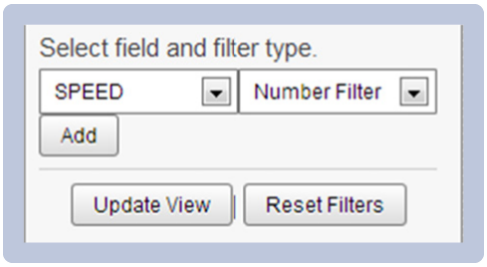


Figure 6.21. Defining filter criteria.

3. The user chooses filter conditions and clicks Update View to update the visualization in the central panel.

6.4.2.10 Using Slider

Slider filters constrain the data set between minimum and maximum bounds and are only available on columns of data with numerical values (Figure 6.22). To use slider filters,

1. The user slides the square boxes to add a filter with new maximum or minimum values.
2. Alternatively, the user can type new maximum and minimum values in the white text boxes.

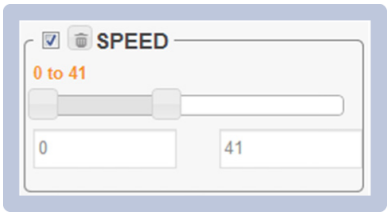


Figure 6.22. Slider filter.

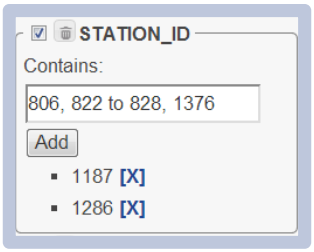


Figure 6.23. Number filter.

6.4.2.11 Number Filter

When using number filters, the user can be more specific about which rows to include in the filter by typing the values and/or filter ranges separated by commas (Figure 6.23). To use number filters,

1. The user types "to" to define a range.
2. The user uses commas to select different values.

In the example shown in Figure 6.23, a filter was created that selects rows in the data set with station IDs equal to 806, 1376, 1187, 1286 and all station IDs between 822 and 828.

6.4.2.12 Text Filter

On data columns containing text, the text filter can be implemented to constrain the data set to only the text values that the user specifies (Figure 6.24). The user just needs to type the text filter and click Add. In the example shown in Figure 6.24, a filter was created that selects rows in the data set in the U.S. states of DE, CO, DC, AL, ID, MD, or FL. The text criteria can be separated by commas.

6.4.2.13 Finer Details About Filters

The results of a search can be enhanced by adding multiple filters. Therefore, it is important to understand the logical operators involved in filtering: AND versus OR. OR logic is applied within a filter, while AND logic operates between

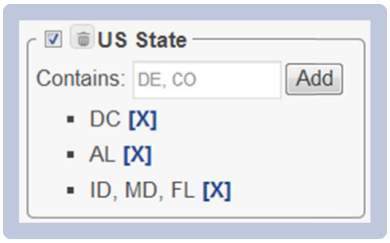


Figure 6.24. Text filter.

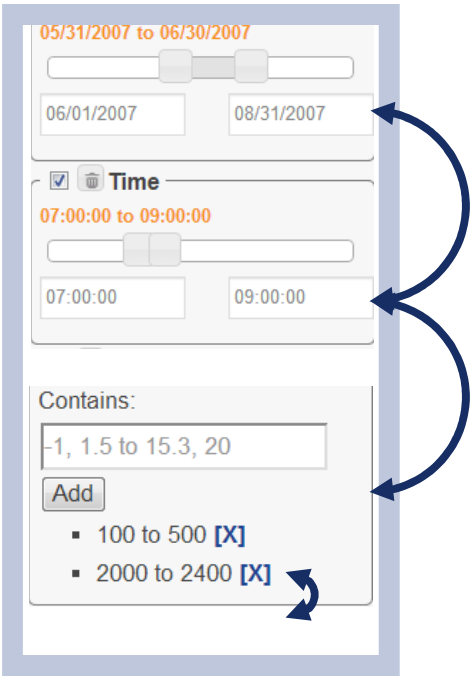


Figure 6.25. Example of search with AND and OR operators.

filters. For instance, in the example shown in Figure 6.25, a filter was created that selects rows in the data set with

- Dates in summer 2007 AND
- Times in the a.m. peak period AND
- Volumes between 100 to 500 OR 2000 to 2400.

Other details about filters include the following:

- To reset a filter, the user clicks Reset Filters.
- To remove an individual filter, the user clicks the trash can next to the filter heading.
- To temporarily show/hide filters, the user clicks the tick box next to the filter heading.

6.5 Download an Artifact

There are two types of artifact downloads: (1) full download and (2) subset download.

6.5.1 Full Download

This feature enables the user to download the file originally uploaded by the creator/PI to the Archive. This feature is accessible from the top right of the Metadata or Artifact page. This option is available for all artifacts (i.e., both non–data sets and data sets). For data sets, this feature will download the .csv file.

6.5.2 Download Filtered Artifact

Users who have just filtered a data set can download the remaining rows by choosing the Download Filtered Data option at the bottom of the filter panel. If the user has not applied any filters, then the original data set will be downloaded.

6.6 Discussion

The last tab on the artifact page is the Discussion tab that is used for providing comments on artifacts and projects. It is also used to rate artifacts and projects.

6.6.1 Comment

This feature gives users an opportunity to express their opinions or ask questions about an artifact. The comment feature is also capable of blocking inappropriate words and phrases automatically. However, it is not a foolproof method to stop inappropriate comments. Users can report inappropriate content by contacting the site administrator via the feedback form.

6.6.2 Rating Artifacts

Users can rate an artifact with up to five stars (Figure 6.26). The same rating will be applied to all future comments on that artifact unless the user decides to change his/her rating or remove the rating. To remove the rating, the user can click the red circle on the left side of the stars.

Users will be required to provide a supporting comment to justify any rating.

Figure 6.26. Rating artifacts.

6.7 Uploading Artifacts

6.7.1 Who Can Upload Artifacts?

The Archive allows PIs and creators to upload artifacts from their research projects. In the future, all registered users may be able to upload other supporting material, but this option is not available at the time of this writing. Therefore, the remainder of this section is applicable to PIs and creators only.

To gain the PI level of access,

1. The user registers as a user.
2. The user contacts the administrator via the feedback form. The user will need to provide his or her name, e-mail address, and the name of the SHRP 2 project.

6.7.2 Artifact Upload Wizard

Before uploading an artifact, the file needs to be prepared using the required formatting (see Section 5.6). To upload artifacts, the uploading user must be logged in and the SHRP 2 system must recognize the account as belonging to a SHRP 2 PI.

To start the upload process, the user should navigate to the My Profile page, select the My Artifacts tab, and then click + Upload a New Artifact (Figure 6.27).

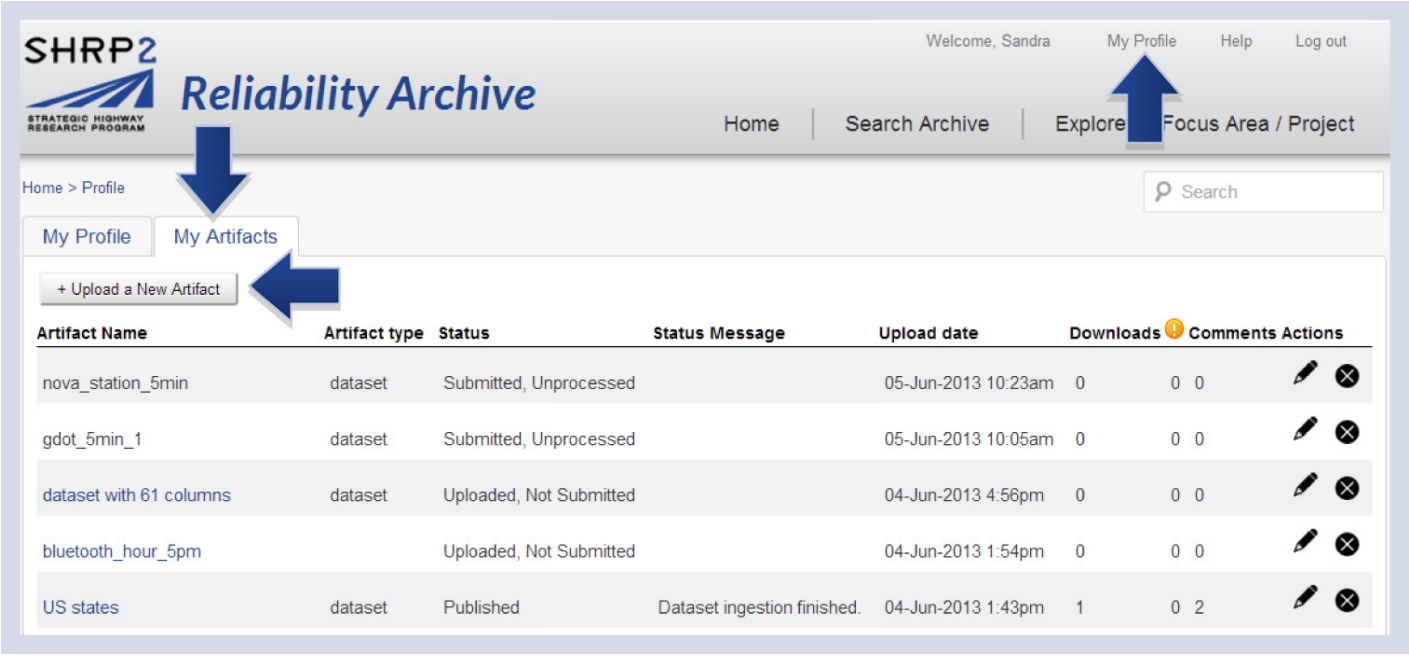


Figure 6.27. Uploading an artifact.

6.7.2.1 Step 1: Select File

The purpose of this step is to choose an artifact to upload. The user clicks the button to choose a file and then clicks Save and Continue. Large files may take some time to load at this step. For the list of acceptable file formats, please read Section 5.6.2.2.

6.7.2.2 Step 2: Confirm Data Type

Users uploading a data set (i.e., a .csv file) then should complete Step 2 of the Upload Wizard. If the user chose a non–data set in Step 1 (i.e., anything other than a .csv file), then the wizard will skip Step 2 entirely.

In Step 2, the wizard will display the headings and a few rows of the data set for the user to review. Then Step 2 asks for some information about each column of data (Figure 6.28).

- First, the user gives each column a heading. Headings should be between 1 and 80 characters long, unique, and contain only permitted characters including a to z, A to Z, 0 to 9, dashes, spaces, and underscores. The heading should be user friendly.
- Next, the user should choose the type of data in each column. The options are number, text, date-time, date, and time.
- Alternatively, the user may choose to exclude a column of data entirely. The Archive will accept a maximum of 60 columns of data, so the user should reduce the number of columns if in excess of this number.

The system does not allow edits to Step 2 in the Upload Wizard after the file has been submitted. Therefore, the user should review the column headings and data types thoroughly before submitting the artifact for processing.

6.7.2.3 Step 3: Set Metadata

Step 3 of the Upload Wizard is all about entering other information about the artifact (or the metadata). Entering these data allows the search functions of the Archive to find the uploaded artifact (Table 6.1).

6.7.2.3.1 Related Artifacts

This field acknowledges that relationships can exist between artifacts. For example,

- A data set may have been used to determine the final recommendations in a report;
- A raw data set may have been cleaned up into a processed data set; and
- Someone on one project may use a data set from a different project.

This section can be used to create relationships among artifacts. In the example shown in Figure 6.29, the Georgia DOT data set is related to the Atlanta Case Study, which in turn is related to the artifacts from Northern Virginia.

Related artifacts can be selected in the left side by searching for their title or artifact ID number. The ID number of any artifact can be found on the Metadata tab of a published artifact. To select the related artifact, the user clicks the + sign to move it to the right into the selected items side. To deselect any artifact, the user clicks the – sign on the right side.

6.7.2.4 Step 4: Publish Artifact

The user reviews the metadata selections before submitting the artifact for processing. To make changes before submitting, the user presses the Back button.

6.7.3 What Happens After the User Submits an Artifact?

After submission, the artifact is processed and validated by the Archive back end and then approved by the administrator. A number of checks are undertaken both by the back end (S2A server) and by the administrator to review the artifact.

Large files will take longer to publish. The status of any artifact upload can be viewed on the My Artifacts tab on the My Profile page.

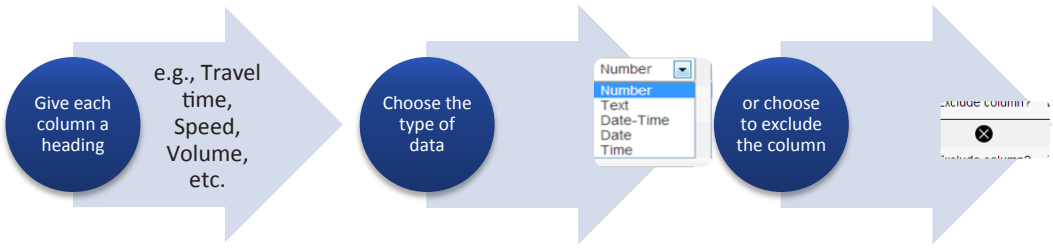


Figure 6.28. Artifact upload wizard, Step 2.

Table 6.1. Metadata Information Required During Upload

Name	Instructions	Data Sets?	Non–Data Sets	Input Required
Title	Short descriptive title	✓	✓	✓
Description	5–10-line description which may include: purpose, origin of data, processing techniques, observations, findings or acknowledgments. Be concise.	✓	✓	✓
Project	Select a SHRP 2 project that your artifact is associated with.	✓	✓	✓
Class	Select SHRP 2 Primary for files that were produced as part of a SHRP 2 research project. Select User-submitted for all other files.	✓	✓	✓
Artifact Type	Select file type. CSV files will automatically be nominated as 'Data sets.' For non–data sets please choose the type of artifact, e.g., final report.	✓	✓	✓
Related Artifacts	This field acknowledges that relationships can exist between artifacts. Further information provided below the table.	✓	✓	
Years described	Year range described in this file. For example, the traffic data in your data set may be collected during 2011 and 2012.	✓	✓	
Locations	This field gives you the opportunity to specify up to ten locations for the artifact. You may determine the location based on the location of the traffic data included in a data set. For non–data sets, your final report or presentation may be written about data collected at a particular location. Type a City, State, or Open Street Map ID. Then click the "Lookup" to validate the location.	✓	✓	
Data dictionaries	A data dictionary is a document that describes the data stored in the data set. For each column heading describe the data stored and the units of measurement.	✓		✓
Data source(s)	The organization that provided the data. This could include government bodies or third parties, e.g., traffic.com.	✓		
Corridors	This field captures the names of the road in a data set. For example, US101.	✓		
Data Types	Check the types of information that is included in this data set.	✓		✓
Collection Technologies	How was the data collected? Choose the 'on site' field data collection technology.	✓		
Collection Frequency	Time interval of data collection.	✓		
Days of Week	The day of week that data was collected.	✓		
Holidays	Indicate whether this data includes holidays or not.	✓		

Figure 6.29. Example of related data sets.

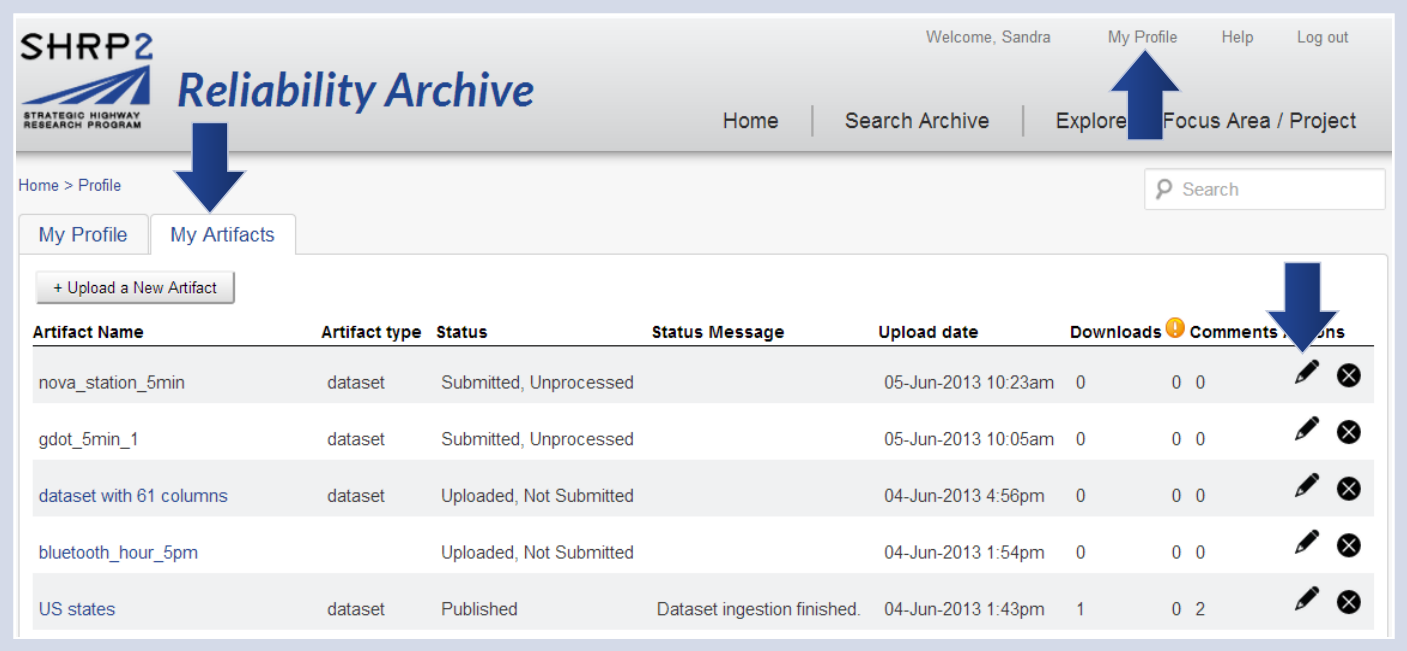


Figure 6.30. Editing an artifact's metadata.

6.7.4 Unsuccessful Artifact Processing

At times, it may not be possible for the Archive to process a data set. This is normally related to the preparation of data sets. In this case, the upload process needs to be completed again.

6.7.5 Editing Artifact Metadata

The metadata of a processed or in-process artifact can be edited anytime by clicking the edit pencil on the My Artifacts tab on the My Profile page (Figure 6.30).

The system does not allow edits to Step 2 in the Upload Wizard after the file has been submitted. Therefore, the user should review the column headings and data types thoroughly before submitting the artifact for processing.

6.7.6 Deleting Artifacts

The PI or creator can delete an uploaded artifact via the My Artifacts tab by clicking the black and white cross next to the artifact (Figure 6.31). The My Artifacts tab is located on the My Profile page.

6.8 Automatically Generated E-mail Notifications

This section lists the automatically generated e-mail notifications that are used in the Archive. Usually these e-mails are used to (1) validate the identity of a registered user, (2) notify the user of the status of an uploaded artifact during the ingestion process, and (3) inform PIs and creators when someone has commented on their artifact. The following subsections illustrate the various automatic e-mail notifications.

6.8.1 Validate the Identity of New User

Subject: SHRP 2 Reliability Archive: Registration validation

Dear [First Name],

Welcome to the SHRP 2 Reliability Archive. Your registered email address and temporary password are below:

Email Address: [email address]
Password: [temporary password]

Please complete your registration by clicking the following link and entering your temporary password [link to login screen].

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

This message has been automatically generated. To contact the site administrator, please complete the feedback form [link to feedback screen].

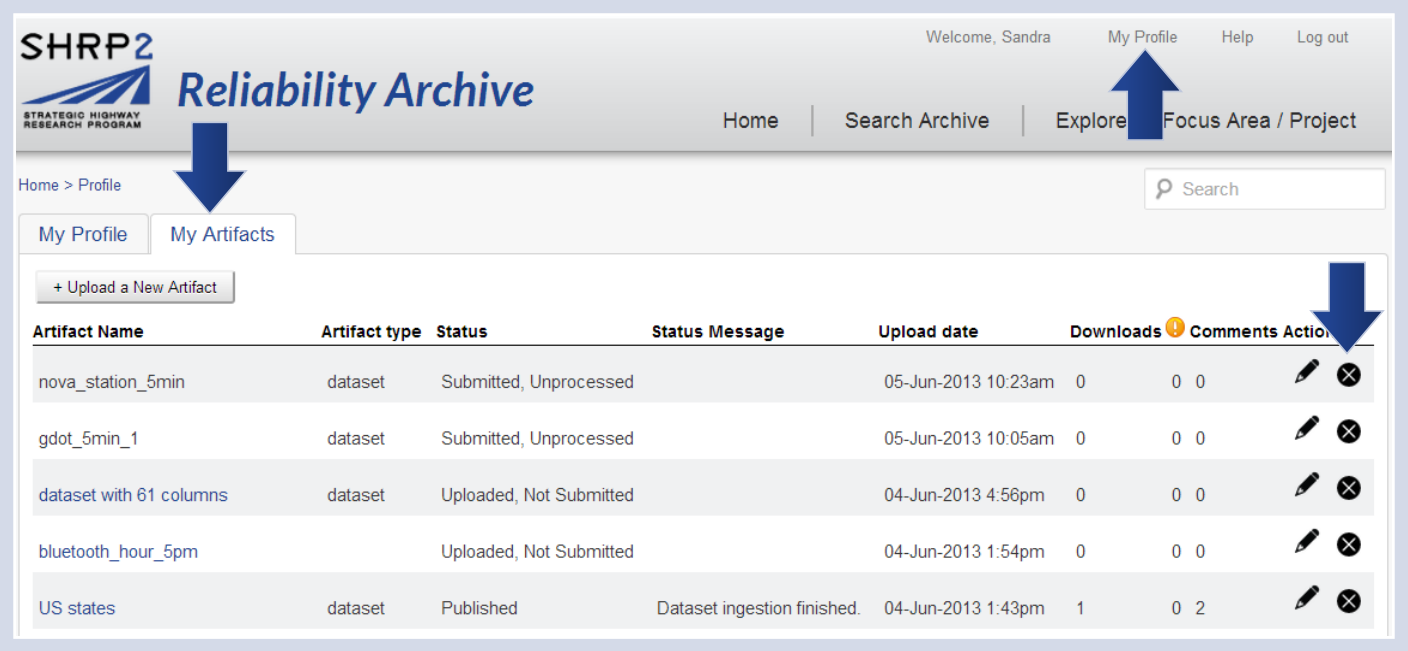


Figure 6.31. Deleting an artifact.

6.8.2 Artifact Is Processing

Subject: SHRP 2 Reliability Archive: Your artifact is being processed

Dear [First Name],

This email is to confirm that your artifact is being processed. Please note that processing may take some time for large data sets.

Artifact Title: [Artifact title]

Artifact ID: [Artifact ID if possible?]

Once processing is finished you can edit the artifact's metadata and view all the artifacts you've uploaded, click My Profile and then My Artifacts on the SHRP 2 Reliability Archive web page.

Thank you for your contribution to the SHRP 2 Reliability program Archive!

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

This message has been automatically generated. To contact the site administrator, please complete the feedback form [link to feedback screen].

6.8.3 Artifact Has Finished Processing

Subject: SHRP 2 Reliability Archive: Artifact processing is complete

Dear [First Name],

This email is to confirm that your artifact has been processed.

Artifact Title: [Artifact title]

Artifact ID: [Artifact ID]

Artifact URL: [link to artifact]

To edit the artifact's metadata and to view all the artifacts you've uploaded, click My Profile and then My Artifacts on the SHRP 2 Reliability Archive web page.

Thank you for your contribution to the SHRP 2 Reliability program Archive!

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

This message has been automatically generated. To contact the site administrator, please complete the feedback form [link to feedback screen].

6.8.4 Artifact Processing Was Unsuccessful

Subject: SHRP 2 Reliability Archive: Artifact processing was unsuccessful

Dear [First Name],

The SHRP 2 Reliability Archive could not process your data set and unfortunately you will need to complete the upload process again.

Artifact Title: [Artifact title]

Reason for processing error: [Reason for processing error]

Possible fixes: [Corresponding fix to the problem]

Please be assured that help is available. Our site administrator can assist you with the upload process and provide helpful information about pre-processing data sets. To contact the site administrator, use the Archive's feedback form [link to feedback screen].

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

This message has been automatically generated. To contact the site administrator, please complete the feedback form [link to feedback screen].

6.8.5 Artifact Has Been Removed

Subject: SHRP 2 Reliability Archive: Your artifact has been removed

Dear [First Name],

This email is to notify you that one of your artifacts has been removed from the SHRP 2 Reliability Archive.

Artifact Title: [Artifact title]

Artifact ID: [Artifact ID]

If you did not remove this artifact and you'd like to understand why it was removed, please contact the site administrator via the Archive's feedback form [link to feedback screen].

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

This message has been automatically generated. To contact the site administrator, please complete the feedback form [link to feedback screen].

6.8.6 Someone Has Commented on a Principal Investigator's or Creator's Artifact

Subject: SHRP 2 Reliability Archive: Your artifact has received a comment

Dear [First Name],

Your artifact has generated interest amongst the community and someone has made a comment!

Artifact Title: [Artifact title]

Artifact ID: [Artifact ID]

Comment: [The most recent comment]

Feel free to respond to the comment on the Discussion tab of your artifact. If you believe the comment contains inappropriate material, please contact the site administrator via the Archive's feedback form [link to feedback screen].

Thank you for your contribution to the SHRP 2 Reliability program Archive.

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

This message has been automatically generated. To contact the site administrator, please complete the feedback form [link to feedback screen].

CHAPTER 7

System High-Level Architecture

The SHRP 2 Archive system consists of the following components:

- Amazon Web Services (AWS);
- Apache HTTP server;
- WordPress system with specific SHRP 2 plugins and themes;
- MySQL database;
- Tomcat application server;
- Solr search engine; and
- S2A server.

These components are interconnected, as shown in Figure 7.1. Detailed information on the system components is provided below.

7.1 Amazon Web Services

AWS is a bundle of remote computing services that provides a cloud-computing platform offered over the Internet. Both the L13 report and L13A team's assessments indicated that the cloud-based service is a viable solution for hosting the Archive. From the project team's point of view, the architecture proposed in the L13 report (See Section 3.1.10 earlier in this report) was slightly outdated. To that end, the team modified the proposed architecture and leveraged the extensive cloud-based services Amazon provides to the public. The team deployed the Archive system on a bundle consisting of the following components:

- *Amazon Elastic Compute Cloud (EC2).* EC2 provides virtual servers and is delivered on the CentOS operating system. EC2 manages the data and information via Elastic Block Storage (EBS). EBS provides volume-based storage that has a separate life span and can be attached to any instance. EBS module size is 200 GB and can be resized. For now the team has used the medium M3 instance for the EC2 module. It should be noted that in the design, the team has not implemented a hot standby instance as a backup for cases in which the operation of the EC2 module fails. Amazon guarantees uptime of more than 99%. In case of any potential failure, the administration team can set up another instance in a couple of hours.
- *Amazon Relational Database Service (RDS).* Database administration (e.g., configuration, backup, monitoring resource consumption) can be an expensive and error-prone task. The purpose of this module is to provide a relational database service via Amazon cloud that helps users save money and avoid errors. RDS supports three popular relational databases: MySQL, SQL Server, and Oracle. The Archive uses MySQL for managing its database system. As of April 2014, the size of the database was 500 GB. The service is elastic. Therefore scaling up the storage is easy.
- *Simple Storage Service (S3)/Glacier.* This service is used to back up the database and the file system. The Archive backs up the contents of the EBS daily and the RDS biweekly on S3. S3 keeps the data for 1 month and then moves them to Glacier, a cost-efficient archival storage system with very high availability and very low failure rate. It should be noted that sending and retrieving data to and from Glacier can take time. The size of the S3 storage service is 2 TB (as of April 2014).

7.2 WordPress

WordPress is one of the most popular open source content management and blogging systems available. WordPress was selected as the core CMS of the Archive after a thorough assessment of various COTS CMSs (see Section 3.5.4 for more information).

WordPress requires a web server with PHP support, a URL rewriting facility, and an instance of MySQL. The Archive system uses Apache as the HTTP server. Apache is a preferred option that developers normally implement with WordPress because it provides PHP interpretation and URL rewriting.

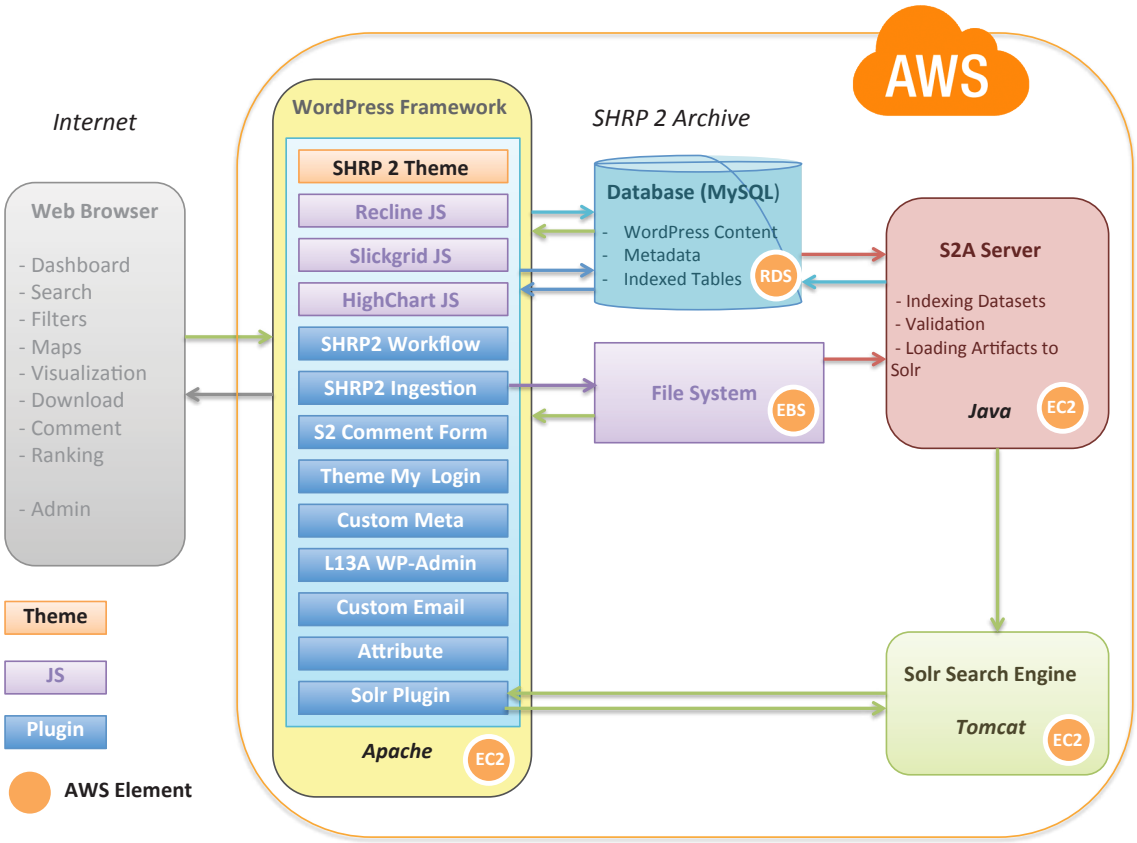


Figure 7.1. Components of SHRP 2 Archive.

7.2.1 Themes

The WordPress theme is the face and graphical aspect of the website which encompasses the entire user experience. Therefore, the appearance of user interface is built on the basis of a theme. A theme is a bundle of template files (PHP files to provide logic and structure), CSS files (to keep the style), images, and JavaScripts.

There are many WordPress theme resources available that can be used directly or customized. The SHRP 2 Archive theme is a child theme of WordPress's Twenty Eleven general theme. The SHRP 2 Archive theme was customized for the Archive user interface.

7.2.1.1 Key Open Source JavaScript Libraries Used in the Archive

JavaScript works within WordPress. It can be used within WordPress template files in WordPress themes or child themes. As recommended by the L13 report, the project team effort was to use open source libraries as much as possible. Table 7.1 summarizes the list of open source JavaScript libraries used to deliver some of the core functionalities of the Archive system.

7.2.2 Plugins

In WordPress, a plugin is a PHP file that provides specific functionality to a website. It allows the theme to achieve a certain objective and help users tailor the website for their specific needs. Table 7.2 shows the list of plugins used for the Archive.

Table 7.1. Open Source JavaScript Libraries

JavaScript	Description
Recline	Library to build data applications. It can be integrated with Leaflet, Slickgrid, and Highcharts. This library was used as a platform that delivers the visualization functionalities on the Data tab located on top of the data set pages.
Slickgrid	Grid/spreadsheet view of the data sets
Highcharts	Data set plots and graphs
Leaflet	Interactive maps features (i.e., markers, overlapping marker spiderfier)
Cloudmade	Map tiles based on OpenStreetMap. At the time of writing this report, Cloudmade stopped providing the free service. The team is looking into finding alternatives, such as Google or Nokia.

Table 7.2. Plugins Used in Archive

Plugin	Description
Attributes plugin	Used to handle inappropriate content, ratings, and such. This feature was implemented into the system but is not being used.
Custom e-mail	Sends custom e-mail from SHRP 2 Archive plugins and adds a custom registration e-mail.
L13A ingestion	Implements the custom file ingestion process for the L13A Reliability data archive.
L13A WordPress-Admin Restriction Mod	Hides the WordPress admin banner on top of the site.
Meteor Slides	Easily creates responsive slideshows with WordPress that are mobile friendly and simple to customize. In the SHRP 2 Archive system, the administrator has the ability to insert a slideshow at the homepage.
S2 comment form	A plugin to add custom fields to the comment form.
SHRP 2 Custom Meta	This plugin defines and enables custom metadata fields.
SHRP 2 Workflow	Enables administrators to manage artifacts in the SHRP 2 Archive.
Solr for WordPress	Indexes, removes, and updates documents in the Solr search engine.
Theme My Login	Themes the WordPress log in, registration, and forgot password pages according to your theme.

7.3 MySQL Database

The only database that is supported by WordPress is MySQL version 5.0.15 or greater (the version number may change later). For most applications WordPress normally deals with the database by itself. So the developer does not need to worry about the structure and the design of the database. However, for this project, the development team has customized the database. The customization was implemented in two forms: modifying existing WordPress tables and adding new tables. Section 7.3.1 and Section 7.3.2 review the native WordPress tables and the SHRP 2–specific tables in more detail.

Note that the Archive stores data sets in two formats:

- Flat file, which is the original .csv format kept in WordPress's file system; and
- Database table, which is used for visualizing and filtering the data sets. The system generates these tables automatically by converting .csv files to database tables during the back-end processing (see Figure 5.3).

7.3.1 Database Diagram

Figure 7.2 provides a visual overview of the SHRP 2 Archive database and the relations between the tables required to operate the Archive. Tables starting with "s2_dset_" are converted from original data set files in .csv format. The general naming convention for a data set table is "s2_dset_ArtifactID"; *ArtifactID* is a unique identifier that is assigned to each file (artifact) by WordPress. Note that the s2_dset_1001 table is only an example of a data set table.

7.3.2 Overview of Database Tables

Table 7.3 lists database tables for the Archive.

Table 7.4 to Table 7.8 show fields in tables created or modified for the Archive. Table names starting with "s2" represent relations specifically created for the Archive system.

7.4 Solr Search Engine Server

Solr is an open source enterprise search engine that performs keyword search on the Archive. Solr is written in Java and runs as a standalone full-text search server within a servlet container such as Jetty. Solr uses the Apache Lucene Java search library at its core for full-text indexing and search, and has REST-like HTTP/XML and JavaScript Object Notation (JSON) APIs that make it easy to use from virtually any programming language (Apache Lucene 2014). The Archive's Solr engine has been installed on Apache Tomcat. Solr indexes any artifact and metadata being uploaded into the Archive before they become available on the Archive.

7.5 S2A Server

S2A server is a back-end module, written in Java, to manage each artifact's workflow and processing states in the Archive. Depending on the type, an artifact goes through different back-end processes. The workflow controls various processing paths that an artifact goes through, from the time it is uploaded into the Archive until the moment it becomes available in (or gets deleted from) the Archive. S2A core functionalities are listed in Subsection 5.2.4. (Step 4. Back-End Processing).

There are three state variables by which the status of an artifact in the Archive is defined. These variables are

(text continues on page 85)

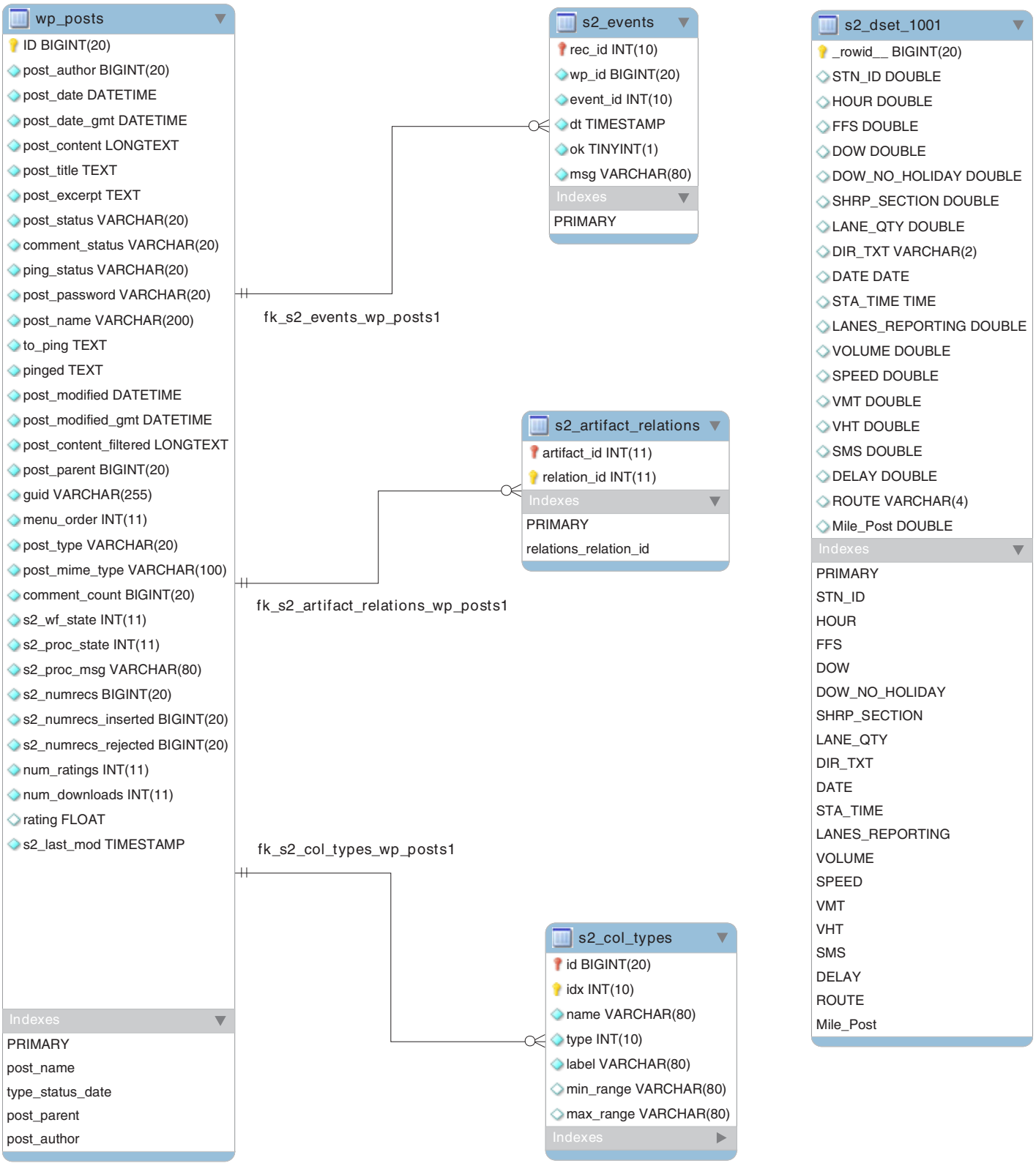


Figure 7.2. SHRP 2 Archive database diagram.

Table 7.3. SHRP 2 Archive Database Tables

Table Name	Description	Created By
wp_commentmeta	Each comment features information called the metadata, and it is stored in the wp_commentmeta table.	WordPress[a]
wp_comments	The comments within WordPress are stored in the wp_comments table.	WordPress
wp_links	The wp_links table holds information related to the links entered into the Links feature of WordPress. (This feature has been deprecated but can be reenabled with the Links Manager plugin.)	WordPress
wp_options	The options set under the Administration > Settings panel are stored in the wp_options table. See Option Reference for option_name and default values.	WordPress
wp_postmeta	Each post features information called the metadata, and it is stored in the wp_postmeta. Some plugins may add their own information to this table.	WordPress
wp_posts	The core of the WordPress data is the posts. They are stored in the wp_posts table. Also pages and navigation menu items are stored in this table. This table is customized for the Archive and includes information on workflow state, record inserted, number of ratings, average rating, number of downloads, and last time the artifact was modified.	WordPress (This table is customized for the Archive.)
wp_terms	The categories for both posts and links and the tags for posts are found within the wp_terms table.	WordPress
wp_term_relationships	Posts are associated with categories and tags from the wp_terms table, and this association is maintained in the wp_term_relationships table. The associations of links to their respective categories are also kept in this table.	WordPress
wp_term_taxonomy	This table describes the taxonomy (category, link, or tag) for the entries in the wp_terms table.	WordPress
wp_usermeta	Each user features information called the metadata, and it is stored in wp_usermeta.	WordPress
wp_users	The list of users is maintained in table wp_users.	WordPress (This table is customized for the Archive.)
s2_artifact_relations	The table stores the relationships among artifacts.	L13A team
s2_col_types	The column types of each data set are stored in this table.	L13A team
s2_dset_*ArtifactID*	*Artifact_ID* represents the artifact ID number of a data set (automatically generated by WordPress). This table stores the content of a data set and is used for visualizing and filtering.	L13A team
s2_events	This table stores the ingestion state of all the artifacts.	L13A team

[a] For more information on WordPress database descriptions, visit http://codex.wordpress.org/Database_Description.

Table 7.4. S2_artifact_relations Table Fields

Field	Type	Null	Key	Default	Extra
artifact_id	int (11)	NO	PRI		
relation_id	int (11)	NO	PRI		

Table 7.5. S2_col_types Table Fields

Field	Type	Null	Key	Default	Extra
id	bigint (20) unsigned	NO	PRI		
idx	int (10) unsigned	NO	PRI		
name	varchar (80)	NO			
type	int (10) unsigned	NO			
label	varchar (80)	NO			
min_range	varchar (80)	YES			
max_range	varchar (80)	YES			

Table 7.6. S2_dset_ArtifactID Table Fields

Field	Type	Null	Key	Default	Extra
_rowid__	bigint (20)	NO	PRI		auto_increment
Data set column[a]	*Column type*	*Depends*	*MUL*		

[a] This table stores data set columns. The field and type vary depending on the data set.

Table 7.7. Wp_posts Table Fields

Field	Type	Null	Key	Default	Extra
ID	bigint (20) unsigned	NO	PRI		auto_increment
post_author	bigint (20) unsigned	NO	MUL	0	
post_date	datetime	NO		0000-00-00 00:00:00	
post_date_gmt	datetime	NO		0000-00-00 00:00:00	
post_content	longtext	NO			
post_title	text	NO			
post_excerpt	text	NO			
post_status	varchar (20)	NO		publish	
comment_status	varchar (20)	NO		open	
ping_status	varchar (20)	NO		open	
post_password	varchar (20)	NO			
post_name	varchar (200)	NO	MUL		
to_ping	text	NO			
pinged	text	NO			
post_modified	datetime	NO		0000-00-00 00:00:00	
post_modified_gmt	datetime	NO		0000-00-00 00:00:00	
post_content_filtered	longtext	NO			
post_parent	bigint (20) unsigned	NO	MUL	0	
guid	varchar (255)	NO			
menu_order	int (11)	NO		0	
post_type	varchar (20)	NO	MUL	post	
post_mime_type	varchar (100)	NO			
comment_count	bigint (20)	NO		0	
s2_wf_state	int (11)	NO		0	
s2_proc_state	int (11)	NO		0	
s2_proc_msg	varchar (80)	NO			
s2_numrecs	bigint (20) unsigned	NO		0	
s2_numrecs_inserted	bigint (20) unsigned	NO		0	
s2_numrecs_rejected	bigint (20) unsigned	NO		0	
num_ratings	int (11)	NO		0	
num_downloads	int (11)	NO		0	
s2_last_mod	timestamp	NO		CURRENT_TIMESTAMP	on update CURRENT_TIMESTAMP

Table 7.8. S2_events Table Fields

Field	Type	Null	Key	Default	Extra
rec_id	int (10) unsigned	No	PRI	NULL	Auto_increment
wp_id	bigint (20) unsigned	No		NULL	
event_id	int (10) unsigned	No		0	
dt	timestamp	No		CURRENT_TIMESTAMP	
ok	tinyint (1)	No		1	
msg	varchar (80)	No			

(continued from page 81)
stored in the wp_posts table. Table 7.9 summarizes the state variables.

- *s2_wf_state* shows an artifact's workflow state. Figure 7.3 depicts the various workflow states.
- *s2_proc_state* indicates the back-end processing status of an artifact. See Subsection 5.2.4 (Step 4. Back-End Processing) for more information.
- *s2_proc_msg* provides processing outcomes in a message for the creator. The message is displayed on the My Artifact list located on the My Profile page.

Table 7.9. State Variables

State Variable	Description	Values
s2_wf_state	Workflow approval state	0 = Ingest, the artifact is in the ingestion process but not yet submitted. This is the default state for a new artifact. 3 = Unprocessed • Triggered by: Submit button clicked in Step 4 of the ingestion process 2 = Processing • Triggered by: administrator reviews and approves 1 = Published, available for public use • Triggered by: S2A server completes processing –1 = Pretrash • Triggered by: administrator sends artifact to Bin state –3 = Trash • Triggered by: S2A server moves artifact from Gulag to Bin state –4 = Processing error (validation, loading, or indexing) • Triggered by: S2A server (see s2_proc_state and s2_proc_msg for details)
s2_proc_state	Processing state	–1 = Error 0 = Unprocessed (default) 1 = Validating 2 = Validation failed (see s2_proc_msg) 3 = Loading 4 = Load failed (see s2_proc_msg) 5 = Indexing 6 = Indexing failed (see s2_proc_msg) 10 = Processing success
s2_proc_msg	Message for users from artifact processing	"Data set ingestion finished." "Could not parse XXX fields. Optimizing table." "Calculating column extents." "Internal error: unknown column type." "Failed to load."

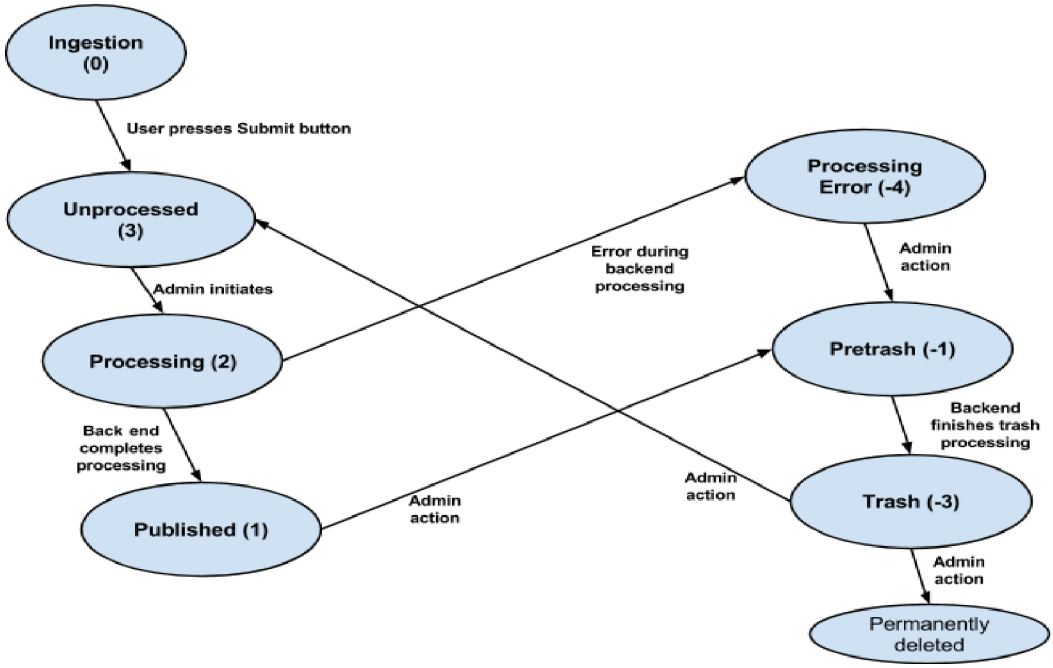


Figure 7.3. Archive processing finite state diagram.

CHAPTER 8

Test Plan

The objective of the testing discipline is to verify that the SHRP 2 Archive functions as designed. The areas tested and their objectives are as follows:

- Functionality—verify that the functional requirements are satisfied;
- Performance—verify that the performance requirements are satisfied;
- Availability—verify that the procedures designed to guarantee the desired availability function correctly;
- Scalability—verify that the scalability requirements are satisfied;
- Maintainability—verify that the code that supplies the SHRP 2 Archive functionality is supported by a comprehensive unit test suite that gives confidence that system functionality is not broken by code changes;
- Integration—verify that the SHRP 2 Archive functions in the required target environments; and
- Security—verify that necessary software security updates have been applied.

Note that the test plan provided in this chapter was based on the latest version (Version 0.3) of the document at the time of developing the final report.

8.1 Development Testing Strategy

Every story/feature that is developed must be tested. The team will run two instances of the Archive system on two servers: (1) production server, and (2) development server. The development server will be used to test the newly implemented stories/features based on the test plan provided below. Once the development server passes the tests, the production server will be updated with the latest changes. The team will use the Bugzilla platform to keep track of the bugs and enhancement stories.

8.1.1 Unit Testing

8.1.1.1 Objective

Produce an application that is maintainable. This objective is met by giving confidence to development engineers that unanticipated side effects of code changes will be detected before a code update is released into production. Unit testing is an industry standard way of doing precisely this.

Unit testing is used to test individual units of the Archive source code. Because the team is following the agile approach, the programmers are mandated to

1. Write unit tests for every class/module written, and
2. Verify that updated code does not break any preexisting unit test.

Note that the written tests should not cross the unit/class boundaries. For example, the unit test code should not try to interact with the database. In this case, a mock object has to be created and used. The details of the test depend on the module/class that is being tested.

8.1.1.2 Items to Be Tested

All code modules must have unit tests. Below are some additional guidelines that are useful to construct a complete unit test suite for a class.

- Test any SHRP 2 unit that sends a request to the database. Make sure the requests and the output at the client side are displayed correctly. Errors, if any, must be caught by the corresponding plugin and logged or shown to the administrator only, not to the end user.
- Test if any errors are shown while executing database queries.

- Test data integrity while creating, updating, or deleting data in the database.
- Check that incorrect values entered in the form fields are handled gracefully.
- Check that invalid latitude/longitude values are handled gracefully.
- Check that ranking an artifact results in the corresponding value adjustment in the database.
- Check that an attempt to plot nonnumeric values is handled gracefully.
- Check that every graph type in the visualization displays as expected.
- Test PHP code that retrieves data for visualization.
- Test Java code that parses a CSV string.
- Test Java code that validates column types.

8.2 System Tests

8.2.1 Functionality Tests

8.2.1.1 Objective

Verify that the functional requirements are satisfied. They include

1. System access control—only valid users are permitted access, determined by their roles;
2. Searching the Archive based on the spatial and functional (i.e., metadata) areas;
3. Uploading an artifact and specification of relevant metadata;
4. Visualization of previously loaded artifacts; and
5. Downloading previously loaded artifacts.

8.2.1.2 Test Cases

8.2.1.2.1 Userid And Password

1. Verify that the userid, L13ATestUser, and password, L13AsECURITYiSaWESOME!, can log in.
2. Verify that the userid, L13ATestUser, and password, L13AsECURITYiSaWESOME!, can log in. Note that userids are not case-sensitive.
3. Verify that the userid, L13ATestUser, and password, L13AsECURITYiSaWESOME!, cannot log in.
4. Verify that the userid, L13ATestUser, and password, L13aSecurityIsNotAwesome, cannot log in.
5. Verify that a user's login credentials are cleared when the browser is restarted. Steps
   a. Log into the SHRP 2 Archive.
   b. Close the browser used to log in.
   c. Restart the browser.
   d. Go to the SHRP 2 Archive URL.
   e. Verify that the user must log in again.

8.2.1.2.2 Project L02

Search criteria (values subject to change), by project: Project L02

Returns:
Orange 5 over Atlanta, GA
Orange 7 over Washington, DC
Orange 9 over San Diego, CA
Orange 10 over Philadelphia, PA
Orange 12 over San Francisco, CA
Blue report over Lake Tahoe, CA

8.2.1.2.3 Project L03

Search criteria (values subject to change), by project: Project L03

Returns:
Orange 3 over Los Angeles, CA
Orange 3 over San Diego, CA
Orange 5 over Houston, TX
Orange 5 over Jacksonville, FL
Orange 9 over San Francisco, CA
Orange 20 over Minneapolis, MN
Blue report over state of Georgia

8.2.1.2.4 Project L05

Search criteria (values subject to change), by project: Project L05

Returns:
Orange 2 over Knoxville, TN
Blue report over Detroit, MI
Blue report over state of Washington

8.2.1.2.5 Final Reports

Search criteria (values subject to change), non–data set types: final reports

Map returns:
Blue report over Lake Tahoe, CA
Blue report over San Diego County, CA
Blue report over Atlanta, GA
Blue report over Knoxville, TN
Blue report over state of Virginia
Blue report over New York City, NY

List returns (values subject to change):
(26) Artifacts, first three are
San Diego case study
New York City case study
Task 3: Technical Memorandum on User Engagement . . .

8.2.1.2.6 Artifact

Upload Artifact A1 and assign the following metadata. Steps:

1. Go to the Upload a New Artifact wizard in the My Profile dialog.
2. Select File→click Browse, and use chooser to locate and load the artifact.
3. Select Save and Continue.
4. Complete the Set Metadata dialog:
   a. Set Description to "Test Artifact."
   b. Set Project to "Project L15—Innovative IDEA Projects."
   c. Set Artifact Type to "Final Report."
   d. Set Locations to "Sacramento, California."
   e. Click Save and Continue.
5. When the Publish Artifact dialog appears, click Submit.

8.2.2 Performance Tests

8.2.2.1 Objective

Verify that the performance requirements are satisfied:

1. UI responsiveness (UIR) as follows:
   a. Basic features return in less than 3 s for 90% of requests;
   b. Search results return in less than 5 s for 90% of requests;
   c. Queries on the Data tab return in less than 30 s per GB of content, 90% of the time;
   d. Visualization returns in less than 5 s per GB of content; and
   e. Ingestion completes in less than 30 s per GB of content.
      i. Assuming an upload feed of 3 MB/s or more.
2. Newly ingested artifacts appear in search results within 4 h of upload.
3. Deleted artifacts no longer appear in search results within 4 h of deletion.

8.2.2.2 Test Cases

8.2.2.2.1 Time Interval

1. Time interval from SHRP 2 Login page to SHRP 2 Landing page meets UIR objective.
2. Time interval from SHRP 2 Landing page to Artifact page meets UIR objective.
3. Time interval from Artifact page, Metadata tab, to Discussion tab meets UIR objective.
4. Time interval from Artifact page, Discussion tab, to Home tab meets UIR objective.
5. Time interval from SHRP 2 Landing page, Home tab, to Search Archive tab meets UIR objective.
6. Time interval from SHRP 2 Landing page, Search Archive tab, to Explore by Focus Area/Project tab meets UIR objective.
7. Time interval from SHRP 2 Landing page, Explore by Focus Area/Project tab; navigate from Reliability Focus Area to Capacity Focus Area meets UIR objective.
8. Time interval from SHRP 2 Landing page, Explore by Focus Area/Project tab; navigate from Capacity Focus Area to Renewal Focus Area meets UIR objective.
9. Time interval from SHRP 2 Landing page, Explore by Focus Area/Project tab; navigate from Renewal Focus Area to Safety Focus Area meets UIR objective.
10. Time interval from SHRP 2 Landing page, Explore by Focus Area/Project tab; navigate from Safety Focus Area to Other meets UIR objective.
11. Time interval from SHRP 2 Landing page, Explore by Focus Area/Project tab; navigate from Safety Focus Area to Home meets UIR objective.

8.2.2.2.2 Specific Location

1. Starting from SHRP 2 Landing page, Search tab, select the orange circle #11 above San Francisco, CA; 11 bubbles appear within the time interval UIR b.
2. Starting from SHRP 2 Landing page, Search tab, with orange circle #11 above San Francisco, CA, selected and 11 bubbles visible, click the bubble closest to the orange circle; a bubble containing 101NB Palo Alto to SR92 Jan 5-31, 2009, appears within the time interval UIR 1.b.

8.2.2.2.3 Artifact Page

1. Starting from the artifact page located at http://shrp2archive.org/?attachment_id=848, go to the Data tab and select the date range from 07/01/08 to 07/02/08; the query result shows up within the time interval UIR 1.c.
2. Verify that ingested artifacts appear in search results within UIR 2 time frame. The steps are as follows:
   a. Ingest Artifact A2 into the SHRP 2 Archive.
   b. As the administrator, process the artifact.
   c. Start a timer and wait for the time interval UIR 2 to pass.
   d. Complete a SHRP 2 search, and verify that Artifact A2 appears in the result set.
3. Verify that deleted artifacts do not appear in search results within the UIR 3 time frame. The steps are as follows:
   a. Delete Artifact A2 from the SHRP 2 Archive.
   b. Start a timer and wait for the time interval UIR 3 to pass.
   c. Complete a SHRP 2 search, and verify that Artifact A2 does not appear in the result set.

8.2.3 Availability Tests

8.2.3.1 Objective

Verify that the procedures designed to guarantee 99% of availability function correctly:

1. Time required to detect that login is not functional is limited to no more than 5 min.

2. Time required to detect that database is no longer communicating is limited to no more than 5 min.
3. Time required to fail over from primary server to backup server completes in less than 4 h.
4. Time required to rebuild a failed database is less than 2 h.
5. Data loss is limited to no more than the artifacts loaded in the last 24 h.

8.2.3.2 Test Cases

8.2.3.2.1 Connection Issues

1. Verify that the login monitoring process detects the inability to log in within the time frame listed in the objectives above. The steps are as follows:
   a. Stop the WordPress Service process (which processes login requests).
   b. Start a timer.
   c. Verify that a notification is received which informs the administrators that the login function has failed within the time frame described in the Objective section above.
2. Verify that the database connection monitoring process detects the inability to connect to the database. The steps are as follows:
   a. Stop the database process.
   b. Start a timer.
   c. Verify that a notification is received which informs the administrators that the database is no longer supporting connections.

8.2.3.2.2 Server Failure

3. Verify that the server failover process completes within the time frame described in the objectives above. The steps are as follows:
   a. Stop the primary server.
   b. Start a timer.
   c. Verify that the backup server takes over within the time frame described in the Objective section. A backup server takeover is successful if a user can log in and perform Test Case 8.2.1.2.
4. Verify that the database reconstruction process can complete a database rebuild within the time frame described in the objectives above. The steps are as follows:
   a. Perform the database reconstruction process and connect it to a trial SHRP 2 Archive server.
   b. Successfully perform all the functional tests described in Subsection 8.2.1, Functionality Tests.
5. Verify that an artifact loaded more than 24 h ago is included in a backup server's Archive inventory. The steps are as follows:
   a. Load Artifact A3 into the primary SHRP 2 server.
   b. Wait 24 h.
   c. Verify that Artifact A3 is present in the artifact list in the backup server.

8.2.4 Scalability Tests

8.2.4.1 Objective

Verify that the system's scalability requirements are satisfied:

1. Five concurrent users observe the performance targets described in the Performance Testing section above.
2. Artifacts up to 2.5 GB can be ingested into the system in the time frame described in the Performance Testing section. Artifacts larger than 2.5 GB are rejected.

8.2.4.2 Test Cases

1. Run five instances in parallel of the tests described in Subsection 8.2.2, Performance Tests. Verify that the performance requirements are satisfied.
2. Verify that Artifact A4, which is 2.5 GB in size, can be ingested into the Archive.
3. Verify that Artifact A5, which is greater than 2.5 GB in size, is rejected within the ingestion process.

8.2.5 Maintainability Tests

8.2.5.1 Objective

Verify that the code that supplies the SHRP 2 Archive functionality is supported by a comprehensive unit test suite that gives confidence that system functionality is not broken by code changes. This is accomplished by inspecting each class to verify that unit tests exist that

1. Verify that invalid values of each incoming parameter are detected and further processing is prevented;
2. Verify positive operation of the class with at least one test case; and
3. Verify invocation of each "catch" block in at least one test case.

8.2.5.2 Test Procedure

Verify unit test existence. The steps are as follows:

1. Procure the source code for the project along with the test cases.
2. For each class, inspect the unit tests and verify
   a. There are test(s) that verify validity of code processing the incoming parameters.
   b. There is at least one test that verifies positive operation.
   c. There is at least one test case for each "catch" block, assuming that the block can be reached with a

combination of input parameter values and/or mock object behavior.

8.2.6 Integration Tests

8.2.6.1 Objective

Verify that the SHRP 2 Archive functions in the required target environments. As this is a web application, the supported integration environments are

1. Internet Explorer (IE) 9 on Windows 7,
2. Safari 5.1 on Mountain Lion (10.8),
3. Firefox 25.0 on Windows 7, and
4. Chrome 30.0 on Windows 7.

Test Procedure

1. Run the tests listed in Subsections 8.2.1, Functionality Tests, and 8.2.2, Performance Tests, using IE9 on Windows 7.
2. Run the tests listed in Subsections 8.2.1, Functionality Tests, and 8.2.2, Performance Tests, using Safari 5.1 on Mountain Lion (10.8).
3. Run the tests listed in Subsections 8.2.1, Functionality Tests, and 8.2.2, Performance Tests, using Firefox 25.0 on Windows 7.
4. Run the tests listed in Subsections 8.2.1, Functionality Tests, and 8.2.2, Performance Tests, using Chrome 30.0 on Windows 7.

8.2.7 Security Tests

8.2.7.1 Objective

Verify that necessary software security updates have been applied. Vulnerabilities described in the National Vulnerability Database must be addressed within 90 days of a fix being produced. The following environments must be updated:

1. CentOS,
2. Java—OpenJDK,
3. Apache Tomcat,
4. WordPress,
5. MySQL, and
6. Lucene and Solr.

Test Procedure

For each of the software modules listed in the Objective section,

1. Identify recently discovered vulnerabilities by searching http://web.nvd.nist.gov/view/vuln/search.
2. Apply remediation within the time frame described in the Objective section.

8.2.8 Visual GUI Testing

8.2.8.1 Objective

Verify the visual appeal of all graphical user interface elements of the Archive.

8.2.8.2 Test Procedure

Check all the pages and GUI elements (e.g., containers, menus, buttons) for size, position, width, length, and acceptance of characters or numbers. GUI elements that have to be checked are as follows:

1. Homepage
   a. Slider on the home page
   b. Latest artifacts on the Recent Artifacts list. Long titles and descriptions should be truncated.
2. Top menus
3. Metadata page
   a. Metadata page and position of the text when the artifact metadata contains a lot of information
   b. Metadata map
   c. Leaflet credentials. These have to be viewable on the map.
   d. Metadata information. Check the metadata information to make sure it is consistent.
   e. Data tab's left and right containers
   f. Text on the filter drop-down menus. This text has to be readable.
4. Grid view
   a. Scroll bar on the grid page
   b. Scroll bar on the filter page when the user enters too many filtering criteria
5. Graph view
6. Map view
7. Discussion tab
   a. Location of the rating stars
8. Ingestion page
9. User profile page
10. Search Archive page
    a. Leaflet credentials. These have to be viewable on the map.
    b. Filter check boxes
    c. Search results on the map
    d. List results

8.2.9 Testing Automation

Functionality and performance test cases were automated using the Selenium framework. The testing code was written

in Python. After each modification (according to the stories submitted through the Bugzilla platform), the testing team ran the code to make sure other elements of the system were not affected. The code is written in a way that it can be run against different browsers (i.e., Mozilla Firefox, Google Chrome, and Internet Explorer).

8.3 List of Artifacts Needed to Run the Test Plan

Table 8.1 lists the artifacts needed to run the test plan.

Table 8.1. Artifacts Needed to Run Test Plan

Artifact Number	Test	Relevant Characteristics
A1	8.2.1.2.6	Used to demonstrate artifact upload function
A2	8.2.2.2.3	Used to demonstrate artifact upload time Used to demonstrate artifact deletion
A3	8.2.3.2.2	Used to demonstrate that artifact upload propagates to backup server
A4	8.2.4.2	2.5 GB in size
A5	8.2.4.2	Greater than 2.5 GB in size

CHAPTER 9

Notes on Operations and Maintenance of the Archive

The design and operation of the Archive system depends not only on the requirements driven by the users but also on financial, technical, and policy-related constraints. Although the L13 report attempted to shed light on those issues (e.g., life-cycle costs, archiving approaches), key strategic questions needed to be revisited and discussed for the L13A project given that the requirements had evolved since the inception of the project.

In the L13A project, the team addressed various issues crucial to design, operations, and maintenance of the Archive system by developing white papers; these were put into discussion among the members of TETG and the SHRP 2 team. The papers were structured in a manner that would provide solution alternatives and were intended to obtain stakeholders' feedback. They reflected only the project team's perspectives at the time of development and were designed to trigger internal discussions at the management level.

The white papers were the basis for some of the key conclusions on the design and operation of the Archive. However, some of the final decisions made on the basis of the papers did not exactly follow the suggestions provided in the papers because of the evolving nature of the project.

This chapter summarizes the project team's assessment of various issues and final conclusions made on the basis of the papers. It also includes concerns that were raised in the white papers to draw the SHRP 2 management team's attention to the Archive's key operations and maintenance risks. The topics that the team investigated in the white papers are as follows:

1. Inclusion of user-submitted data in the Archive;
2. Operations and maintenance; and
3. Data ownership and personally identifiable information.

9.1 Inclusion of User-Submitted Data

One of the outcomes of the June 4, 2012, L13A workshop was the subject matter expert (SME) panel's suggestion to add a feature that allows Archive users to submit their processed/transformed data sets and objects, derived from the original archived data, back into the Archive. SHRP 2 staff believed that feeding back user-generated products was aligned with the SHRP 2 strategic goals, so they were very interested in this idea. The project team investigated the implications of implementing this feature in the Archive system. The results are provided below.

9.1.1 What Are the Submission Scenarios?

The team proposes three user-submission scenarios. Note that any artifact submitted via any scenario is grouped as a user-submitted artifact.

9.1.1.1 Scenario 1

All users can upload flat files only:

- Users could submit only flat files (file size restriction would apply). The system would treat the submitted object as a binary large object (BLOB).
- The metadata requirements would be minimal. As a result, the submission process would be quick and short.
- The administrator would need to validate the submitted file to make sure that it was not corrupted or infected, but no preprocessing step would be required.
- Users would be able to submit their objects under the community pages.

9.1.1.2 Scenario 2

All users can upload any files with no file type limit:

- The ingestion process would be similar to the one for submitting the SHRP 2 Reliability digital objects. Like any other Archive objects, user-submitted objects would need to be validated and preprocessed by the administrator and/or the user.

- Users would be able to submit any digital object that is accepted by the Archive system.
- Users would be able to submit their digital objects from the project pages, the data set pages, and the community pages.
- If the submitted object is a sensor data object, the user would be able to submit two types of data sets: the original file (in .csv format) that includes data extracted from various sensors/segments or a set of sensor-level/segment-level data sets (in .csv or .xls format) in which each set represents data collected from a single sensor/segment.

9.1.1.3 Scenario 3

Trusted users can upload any files with no file type limit:

- In terms of file upload constrains, this scenario is similar to Scenario 2. The only difference is that only a trusted group of users (in addition to PIs) could upload artifacts. At the time of writing the white paper, this scenario was not discussed as an option. It was added later after in-depth discussion with the SHRP 2 and FHWA teams.

9.1.2 Comparison

Table 9.1 compares the three scenarios in terms of major functionality provided by the Archive system. This functionality includes list search, map search, full download, subset download, visualization, and collaboration. Based on the table, Scenarios 2 and 3 would be able to support all of the functionality that is envisioned for SHRP 2 Reliability data objects.

Table 9.2 compares the three scenarios based on various elements that are important to the development and operation of the system. These factors are categorized under five groups: strategic alignment, cost, technology, administration, and risk to project and system.

Table 9.1. Supported Functionality for User-Submission Scenarios

Feature	Scenario 1 (Users may submit flat files only.)	Scenario 2 (All users may upload any files with no file type limit.)	Scenario 3 (Selected users may upload any files with no file type limit.)
List search	●	●	●
Map search	●[a]	●[b]	●[b]
Full download	●	●	●
Subset download		●[b]	●[b]
Visualization		●[b]	●[b]
Collaboration	●	●	●

[a] No sensor location.

[b] Preprocessing of the data set is needed to leverage the feature.

9.1.3 Conclusion

In general, the project team concluded that adding the user-submitted data feature was technically feasible. From the project team's point of view, Scenarios 2 and 3 were more appealing because

- They provide all of the envisioned functionality for the SHRP 2 archived data (see Table 9.1).
- They use the submission system/procedure that the PIs use to submit SHRP 2 objects. Therefore coding efforts would be minimal.

As a result, the team added a new artifact category, "user-submitted," to the system. A feature was also implemented to enable users to report artifacts as "inappropriate." The goal was to help the administrator identify irrelevant artifacts.

The issue of PII was the biggest hurdle, which hindered availability of Scenario 2 (see Section 9.3 for more information). At the moment, the cost of employing a thorough monitoring process to prevent users from submitting PII data is too high for SHRP 2. Therefore, per the SHRP 2 team's request, the project team implemented only Scenario 3, in which only a trusted group of users, namely SHRP 2 contractors, can upload artifacts for the time being.

Lastly, the team believes the adverse implications of needing excessive storage space to host user-submitted data are not significant enough, when compared with the benefits, as long as users submit valuable artifacts to the Archive. As a result, the team proposes an interim solution in which the operating entity creates a small group of trusted members. This group can leverage the already-developed ingestion functionality to submit external Reliability-related artifacts into the Archive.

9.2 Key Issues Associated with Operations and Maintenance

The core objective of this section is to review the various alternatives, as well as their implications, for the operations and maintenance (O&M) of the Archive system. This section tries to discuss the following questions:

- Who is going to operate and maintain the Archive?
- What are system O&M requirements?
- What are hosting options for the SHRP 2 L13A system O&M phase?
- What are the O&M costs?

Table 9.2. Effect of User-Submission Scenarios on Development and Operations

Category	Type	Scenario 1	Scenario 2	Scenario 3	Note
SHRP 2 strategic alignment	Alignment with SHRP 2 strategic goals	Medium	High	Medium	Some features are not supported in Scenario 1.
Cost	Direct cost of hardware	na	na	na	Team will use cloud-computing model.
	Cost of software development	Low	Low	Low	Project team will leverage the existing data ingestion feature for Scenarios 2 and 3.
	Cost of Internet and web services (i.e., cloud)	Medium	High	Medium	Scenario 2 cost is higher because it requires more storage space.
	Recurrent/operation and maintenance costs	Medium	High	Medium	Scenario 2 requires ample administration time to review the artifacts submitted by the regular users.
	Charges for provision of backup services and equipment	Medium	High	Medium	Scenario 2 requires larger storage for backed up files.
Technology	Back-end coding effort	Low	Low	Low	See "Cost of software development."
	Alignment with user requirements	Medium	High	High	
	UI development/implementation effort	Low	Low	Low	
	Metadata entry effort	Low	High	High	
	Required database size	Medium	High	Medium	
Administration	Administration staff effort/cost for processing and validation of artifacts	Low	High	Medium	Scenario 2 requires more administration time to review the artifacts submitted by the users.
	Risk to the project success (budget/schedule risks)	Low	Low	Low	
	Adverse effect of technology evolution on the system operation	Low	Low	Low	

Note: na = not applicable.

9.2.1 Who Is Going to Operate and Maintain the Archive?

Answering this question is beyond the scope of the L13A project, which is only concerned with archiving data from Reliability-related research and development projects. Future operation and maintenance of the Archive is an implementation issue for others to determine, a topic that has already received substantial discussion.

9.2.2 What Are the System and Operations Requirements?

The Archive is currently designed for 99% availability, using routine backup and recovery systems and processes. This guarantees the availability of web pages. In case of disaster recovery, accessing the artifacts (especially data sets) may take longer. All options for the continued operations and maintenance of the system assume the same availability requirements and an operational methodology that sustains the system over the O&M term. The operational methodology includes quarterly updates to the application software and the supporting database software running the Archive, as well as bug fixes for the existing functionality, if issues are found. The following requirements have been used to design the options described below for the L13A O&M phase.

9.2.2.1 Availability and Outage Tolerance Requirements

- Annual availability
  - 99%
- Outage tolerance
  - Application outages are acceptable, but data need to be recoverable, and the annual availability needs to be met.
  - No outage will be greater than 72 h. (This would only occur with a major system failure; the new system would need to reindex the database.)

9.2.2.2 Backup, Storage, and Maintenance Requirements

- Backup strategy
  - Hot backup—required;
  - Recovery testing—reconstruction of a working archive from backup artifact (annual);
  - Backup location—data center or cloud; and
  - Backup frequency—daily.
- Software maintenance
  - Server patches and updates
    - CentOS, MySQL database (quarterly);
  - Software patches and updates
    - WordPress, PHP, Highcharts (quarterly);
  - Application bugs
    - Defects which impede archive functionality—ingest, search, or download (within 4 weeks);
    - Diagnose, patch, and release;
- Disk space
  - Data storage up to 2 TB.

9.2.3 What Are the Hosting Options for the Archive?

There are four options for the long-term O&M of the SHRP 2 L13A Archive system and its artifacts:

- Option 1—server-based with server backup (using existing data center);
- Option 2—server-based with server backup (hosted at a highly available data center);
- Option 3—server-based with cloud backup (hybrid); and
- Option 4—cloud only (Amazon EC2).

Each hosting option provides an effective strategy, but each option balances risks and costs differently. In the server-only–based model (Option 1), the costs are lower; the risks are a single point of failure and the timeliness of recovery in the event of a catastrophic issue. Those risks can be mitigated by transitioning the server-based system to a highly available data center (Option 2). In the server/cloud–based hybrid (Option 3) or cloud-based model (Option 4), the costs are marginally higher, but the data and applications risks are mitigated through cloud-based server models. Also, the potential risks regarding meeting the IT requirements of the system operator can be avoided in the cloud-based approach. Each option is described in more detail below. Option 4 is the recommended option.

For all options, training is required as part of the transition. Training would include 5 days of training material preparation, 2 days of inside training for system administrative staff, and travel to support training activities. The cost for transition training support is $2,500.

9.2.3.1 Option 1: Server-Based with Server Backup (Existing Data Center)

The current server would continue to be the primary server running the SHRP 2 Archive. Additional details are provided below.

- Changes to the current design
  - Purchasing a second server to support backup and hot recovery;
- Benefits
  - Uses equipment already paid for, with only a marginal cost for a secondary server;
  - Redeployment is unnecessary;
- Limitations
  - Power and network are a single point of failure;
  - Same facility supports delivery and backup of application and data;
  - Equipment will need to be replaced every 3 years;
- Cost
  - New equipment
    - Additional disk space for existing server,
    - Backup server,
    - Backup system for backup server,
    - Installation,
    - Total: $6,500;
  - Staff support for two server-based systems (8 h/week): $75,000/year
    - Review and maintenance
      - Weekly deployment review,
      - Patches,
      - Functionality bug fixes;
    - Backup and recovery
      - Annual recovery verification,
      - Full backup (every 3 months) (1 to 2 TB),
      - Incremental backups
        - User-related tables (daily) (500 GB),
        - New artifacts (on upload) (10 GB)—triggered by administration process.

9.2.3.2 Option 2: Server-Based with Server Backup (Hosted at Highly Available Data Center)

The current server would continue to be the primary server running the SHRP 2 Archive. Additional details are provided below.

- Changes to the current design
  - Purchasing a second server to support backup and hot recovery;
  - Moving server location to highly available (HA) data center;

- Benefits
  - Uses equipment already paid for, with only a marginal cost for a secondary server;
  - Redundant power and network;
  - Space is available for servers at HA data center;
  - Could add warranties to extend the expected lifetime of server up to 7 years (approximately $140/year);
  - Spare systems are available on site;
  - The systems are configured with operating system (OS) installations on a dedicated redundant array of independent disks (RAID)–1 pair and data storage on a separate RAID-10 array;
  - Support staff are available 24/7;
- Limitations
  - Equipment will need to be replaced every 3 years (unless warranty extension is used);
  - One-time additional cost to install system in the data center;
- Cost
  - New equipment:
    - Additional disk space for existing server,
    - Backup server,
    - Backup system for backup server,
    - Installation,
    - Total: $6,500;
  - Transition to HA data center
    - Cost to move and reinstall: $2,000;
  - Staff support for two server-based systems (8 h/week): $75,000/year
    - Review and maintenance
      - Weekly deployment review,
      - Patches,
      - Functionality bug fixes;
    - Backup and recovery
      - Annual recovery verification,
      - Full backup (every 3 months) (1 to 2 TB),
      - Incremental backups
        - User-related tables (daily) (500 GB),
        - New artifacts (on upload) (10 GB)—triggered by administration process.

9.2.3.3 Option 3: Server-Based with Cloud Backup (Hybrid)

The current server would continue to be the primary server running the SHRP 2 Archive. Additional details are provided below.

- Changes to the current design
  - Using Amazon S3 for backup and hot recovery;
- Benefits
  - Uses existing equipment for primary server functions;
  - Uses on-demand Amazon service to back up data once a day;
  - Provides off-site risk mitigation with data stored in a secondary location;
  - Minimizes cost by limiting the number of backups a day;
  - Only pay for the hours that the backup runs and the cloud instance needs to function as the primary server during recovery;
- Limitations
  - In the event of a catastrophic failure the maximum amount of data loss is 24-h of data;
  - Marginal cost of an Amazon S3 backup/secondary server is more than a secondary server;
  - Single server point of failure—same facility supporting delivery of data, only data backup in the cloud, not application or server;
- Cost
  - On-demand, large instance (Amazon S3): $8,000/year;
  - Transition to cloud database backup
    - Cost to move and reinstall: $1,000;
  - Staff support for two systems (8 h/week): $75,000/year
    - Review and maintenance
      - Weekly deployment review,
      - Patches,
      - Functionality bug fixes;
    - Backup and recovery
      - Annual recovery verification,
      - Full backup (every 3 months) (1 to 2 TB),
      - Incremental backups
        - User-related tables (daily) (500 GB),
        - New artifacts (on upload) (10 GB)—triggered by administration process.

9.2.3.4 Option 4: Cloud Only (Amazon EC2)

The current server would be decommissioned and Amazon EC2/S3 would be the primary server running the SHRP 2 Archive. Additional details are provided below.

- Changes to the current design
  - Using Amazon EC2 for primary and backup application server and data;
- Benefits
  - No equipment to support or replace;
  - Data and applications are stored in a redundant system;
- Limitations
  - Cost is higher than server version and slightly higher than the cloud data backup version;
- Cost
  - For heavy reserve, large instance (Amazon EC2/S3): $8,500/year;
  - Transition to cloud hosting and database backup
    - Cost to move and reinstall: $2,000;

- Staff support for two cloud-based systems (8 h/week): $75,000/year
  - Review and maintenance
    - Weekly deployment review,
    - Patches,
    - Functionality bug fixes;
  - Backup and recovery
    - Annual recovery verification,
    - Full backup (every 3 months) (1 to 2 TB),
    - Incremental backups
      - User-related tables (daily) (500 GB),
      - New artifacts (on upload) (10 GB)—triggered by administration process.

9.2.3.5 Hosting Options Summary

Each system design described in this section provides an O&M solution for the Archive. The options include a range of physical and cloud-based machines with different configurations for the servers and the supporting database infrastructure. Embedded in each option's technical details are variations of risk for system availability (uptime) and recovery strategy (see Figure 9.1). Given the program's need for a large scalable Archive system, the current uncertainty of which agency will support the Archive in the long term, and the desire for redundancy of data and uptime support, having a flexible and scalable system is important. Therefore, the recommended option is Option 4. Option 4 provides the highest flexibility of server maintenance and data transfer and risk management. Using cloud services through a scalable system like Amazon lowers the O&M risks to the L13A system and provides redundant safety for the Archive. Amazon maintains the physical equipment and supporting infrastructure, and the contract selected with Amazon can be adjusted if the service needs to be supported in a different way in later years.

The L13 report also suggested a hosting approach similar to Option 4.

9.2.4 What Are Operations and Maintenance Costs?

9.2.4.1 Annual Costs

A summary of the four options is provided in Table 9.3 and a summary of their O&M costs is provided in Table 9.4. The cost elements are described in more detail below.

9.2.4.2 Cost Elements

To run the Archive during the O&M phase requires both one-time and ongoing costs. The one-time costs include equipment (servers, hard drive disk space), installation of equipment, and the management of the system transition (program management, transition costs for installation/transfer of equipment, and training). The following definitions describe the one-time and annual costs.

9.2.4.2.1 One-Time Costs

- Disk space—cost of external hard drive to back up the code, artifacts, and other files;
- Backup server—cost of the mirror server that is used when the original server fails to operate;
- Backup system—cost for the equipment used to run backups of the applications server and database files for the backup server;
- Installation—costs to install and configure new supporting physical computer equipment;
- Server transition—costs to transition existing L13A Archive to redundant facilities, whether the facilities are at a physical location or provided by a cloud service like Amazon;

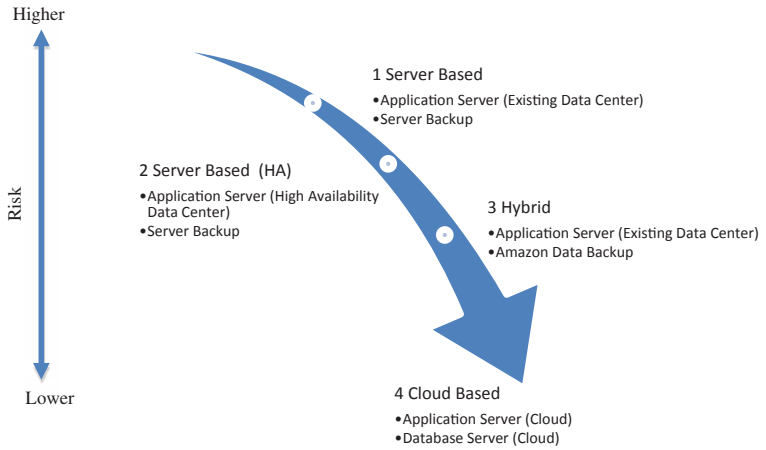


Figure 9.1. L13A options uptime risks.

Table 9.3. Archive System Options Summary

Option	Type	Primary System	Backup	Location
1	Server-based	Server	Server	Existing data center
2	Server-based	Server	Server	High availability data center
3	Hybrid	Server	Cloud (data only)	Existing/HA center/Amazon
4	Cloud-based	Cloud	Cloud	Amazon

- Training transition—costs to train the administrator(s)/operator(s) of the L13A Archive on revised design for O&M and teaching the process of administering the Archive during the O&M phase; and
- Project management—costs to manage the transition to the O&M phase.

9.2.4.2.2 Annual Support Costs

- Server warranty—cost to purchase a warranty for physical servers that guarantees availability of parts and timely service by the equipment manufacturer;
- Annual cloud, on demand—cost to provide on-demand cloud server and database computing units;
- Annual cloud, heavy reserved—cost to provide reserved cloud server and database computing units; and
- Support years—number of years used to calculate annual cost values (Amazon hosting costs).

9.3 Managing Issues with Non-SHRP 2 Data

It was stated earlier that data in the SHRP 2 Archive comes just from SHRP 2 Reliability-related projects. The Archive is currently configured to house a static data set; in the future, though, it could be easily and quickly reconfigured for use by others to add new Reliability-related data. A major concern is that the SHRP 2 Archive not contain data that can be used to personally identify individuals. PII data is simply not allowed in the Archive, and numerous steps have been taken to enforce this:

1. Nearly all the travel time data comes from loop detectors. Travel time from loop data is calculated from data pertaining to many vehicles that pass over a loop in a time slice such as a 5-min period. Thus it is not possible

Table 9.4. Archive System Operations and Maintenance Costs Summary

Item	Options			
	1	2	3	4
Disk space	$500	$500	$0	$0
Backup server	$3,000	$3,000	$0	$0
Backup system	$2,000	$2,000	$0	$0
Installation costs	$1,000	$1,000	$0	$0
Server transition costs	$0	$2,000	$1,000	$2,000
Training Transition	$2,500	$2,500	$2,500	$2,500
One-time costs	**$9,000**	**$11,000**	**$3,500**	**$4,500**
Annual staff support	$75,000	$75,000	$75,000	$75,000
Server warranty	$0	$140	$0	$0
Annual cloud, on demand	$0	$0	$8,000	$0
Annual cloud, heavy reserved	$0	$0	$0	$8,500
Support years	1	1	1	1
Annual costs	**$75,000**	**$75,140**	**$83,000**	**$83,500**
Total costs	**$84,000**	**$86,140**	**$86,500**	**$88,000**

to identify individual vehicles from the loop data in the Archive.
2. For traffic detection technology that can be used to identify origins and destinations, in accordance with standard practices, derived trip lengths have been truncated at both ends so origins and destinations cannot be identified.
3. Standard practices have been employed so that no personal identifiers are associated with the data in the Archive (e.g., personal or machine identifiers have been removed from the record for an individual driver).

In addition, it is important to bear in mind that all the data were generated under contracts of the National Academy of Sciences. Under the contracts, all subject data—including the wide range of types in the SHRP 2 Archive—are owned by the National Academy, and the Academy may authorize others to publish any of the data. SHRP 2 contractors furnishing data to the Archive are fully cognizant of these provisions and, to the best that can be determined, removed all PII and proprietary data from their deliverables so as not to inhibit compliance with the Academies' contract provisions.

To further assure the absence of PII data in the SHRP 2 Archive, a national laboratory has been conducting an independent investigation of the data in the SHRP 2 Archive to make sure there is no PII data in preparation for the Archive's implementation.

This section provides a review of industry practices for data rights protections, author attribution, and options for protecting PII that might be inadvertently added by users of the Archive in the event that the Archive is opened up to non-SHRP contractors in the future. To enable such users to submit Reliability-related artifacts, the project team offers options to manage data licenses and address PII data protections (in case the user artifact upload feature becomes available). The proposed options in this section have not been implemented in the Archive and are raised only to help those concerned with these issues make informed decisions on the topic should a decision be made to turn the SHRP 2 Archive into a dynamic repository in the future.

9.3.1 Open Data

SHRP 2 recognized the benefits of an open system by requiring an open data structure for the outputs of SHRP 2. Specifically, the intent was to design the Archive to be a collective and open data source to foster future research. While open systems encourage collaboration and access, the contribution of user-generated data, and therefore open data sets, adds complexity. Uncontrolled data require guiding principles for data ownership rights, data use requirements, and protection of private data. These principles will need to be effectively managed as a set of requirements that are followed by any contributors to, and users of, the data after the Archive has been fully populated with SHRP 2 Reliability-related data, as originally intended. These issues were raised during the January 2013 stakeholder meeting at the Transportation Research Board annual meeting and are further analyzed here.

SHRP 2 is not the first to provide an open archive to researchers. Fortunately, the benefits of later adoption are significant, as SHRP 2 can learn from the existing open models and their implementation of data rights management. In the last 5 years, many new open data sites have been created by the public sector. As data rights provisions are legal terms, the team examined data rights protections used in open data implementations that follow the same or similar legal structure as would apply to SHRP 2 data.

9.3.2 Open Data Licensing Options

The project team proposes two licensing options.

9.3.2.1 Option 1: Creative Commons License

One of the common content licensing tools used by a large number of sites is Creative Commons (CC). CC licensing provides a structure that is simplified, describes legal terms in plain language, and offers machine-readable licenses that can tell automated programs, including search engines, the license terms. Individuals can then include or exclude data with specific license types from their queries.

9.3.2.2 Types of Creative Commons Licenses

There are seven versions of CC licenses. The first is for "no known copyright" works, called a *public domain license.* The license graphic that accompanies a public domain artifact is shown in Figure 9.2. If the artifact is not in the public domain, six other licenses are available—with four variables that can be selected. The licenses shown in Figure 9.3 are from the CC website at http://creativecommons.org/licenses/.

The four CC license variables are the following:

- *Attribution* ensures that authors of the artifact are mentioned appropriately in derivative works for commercial or noncommercial use.
- *Share-alike* requires users of the artifact to license any derivative works under the same license terms.
- *No-derivatives* allows use of the artifact as is, but does not allow derivatives of the work to be created.

Figure 9.2. Creative Commons public domain license.

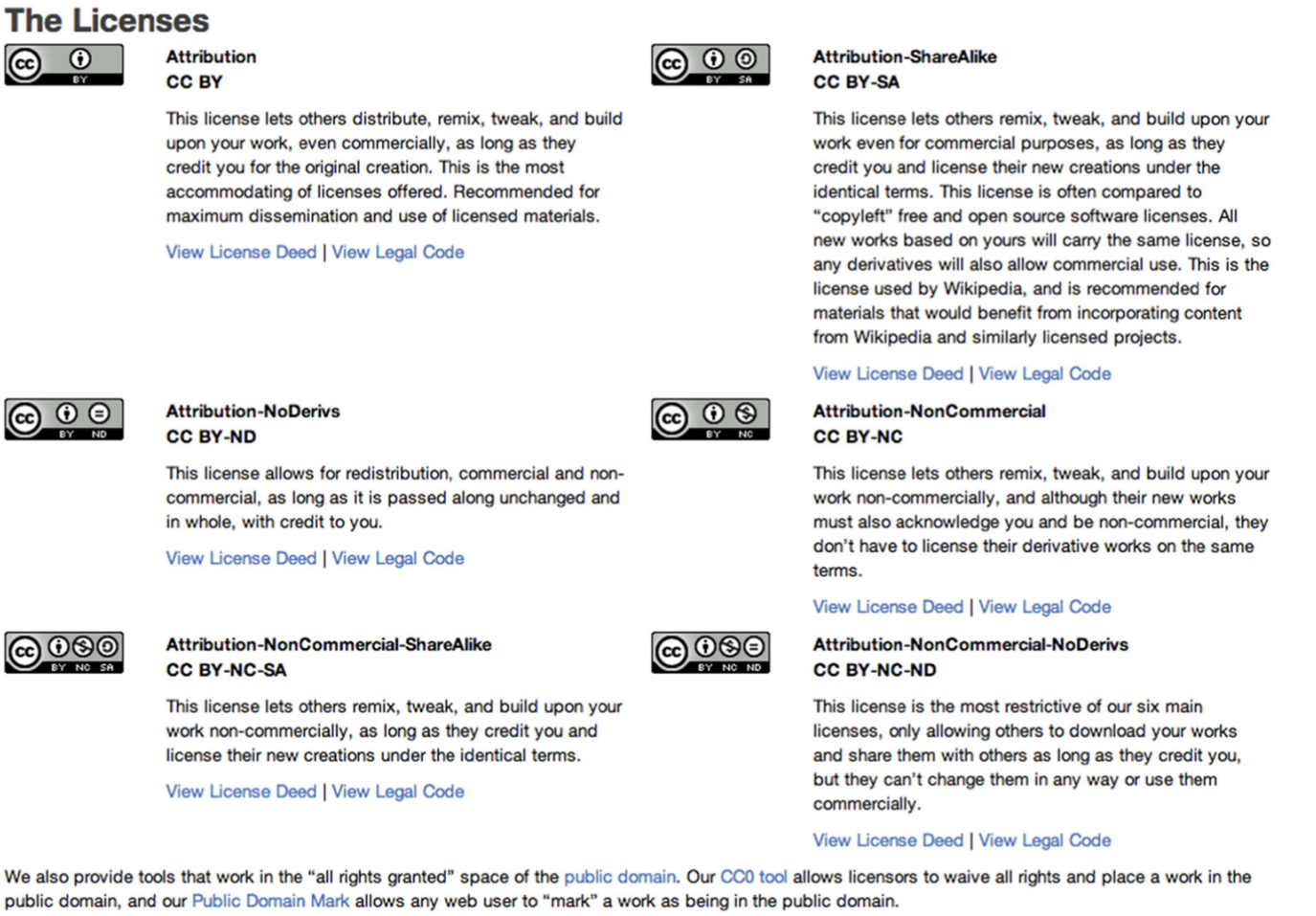


Figure 9.3. Creative Commons license types.

- *Noncommercial* allows noncommercial use of the artifact, but does not allow commercial use.

The main advantages of CC licensing are clarity of license terms, ease of use, and machine-enabled rights tracking. CC offers a clear path to users in the license selection process and tools to see the specific legal terms for more sophisticated legal reviews (see http://wiki.creativecommons.org/Before_Licensing).

9.3.2.3 Option 2: Open Knowledge Foundation—Open Databases

Another data rights management system, specifically designed for databases and the content of databases, is supported by the Open Knowledge Foundation (OKF) Project (http://opendatacommons.org). The OKF is a nonprofit organization based in Great Britain that supports open data projects around the globe. It discourages limitations on data, as its mission is to foster transparency and openness through the opening of data. To support the licensing of databases and data, the OKF provides three license types (http://opendatacommons.org/licenses/):

- Public Domain Dedication and License (PDDL)—public domain for data/databases;
- Attribution license—attribution for data/databases; and
- Open database license—attribution share-alike for data/databases.

Unlike the CC licenses, OKF licenses do not provide any limitations for commercial use or nonderivatives. The foundation provides a narrative that illustrates the differences in database versus data license needs for databases that might be controlled by the author and data that might be controlled under different license terms. The narrative is located at http://opendatacommons.org/faq/licenses/#db-versus-contents.

The narrative describes how to treat the different databases in terms of homogenous databases and nonhomogenous databases (see http://opendatacommons.org/faq/licenses/#db-versus-contents, where the license descriptions below were obtained). When the user controls the database and its content,

the OKF calls the database *homogenous* and uses the following rights permissions:

- Share-alike. Use Open Data Commons Open Database License (ODbL) plus Database Contents License (DbCL) or some other suitable contents license of your choosing.
- Public domain. Use PDDL (it covers both the database and contents).

When the owner of the database and the content of the database are different, the OKF calls the database *nonhomogenous* and uses the following rights:

- Share-alike. Use ODbL for database qua database, plus whatever license you wish/can for contents.
- Public domain. Use PDDL for database qua database, plus whatever license you wish/can for contents.

Note that the CC licenses could be used in conjunction with the OKF licenses in the latter cases to appropriately license the content.

9.3.2.4 Managing Data Rights

Whether Creative Commons, Open Knowledge Foundation, or an alternative licensing form is used, to ensure appropriate treatment of databases and contents of databases and the appropriate digital rights management, a business process should be in place to request that the submitting individual supply the data rights requirements of submitted databases as well as database contents. This can be done through user-based license selection and business processes written into the Archive that capture the input and the proposed license in an administrative review before posting the data, database, or other artifact to the system.

To manage data rights of an open archive, many sites use open data portal software back ends that provide mechanisms for titling and licensing data sets. To ensure appropriate data rights attribution, the business process for upload and data management can be managed to ensure that users select the license to submit; an administrator is able to review the submission before the data are available to the public. This forms-based process ensures that the resulting metadata contains the license terms.

9.3.3 Personally Identifiable Information

As stated above, user-contributed data sets are not permitted now but potentially can be after completion of this project. The goal of allowing users to contribute artifacts and data sets to the SHRP 2 Archive is to expand the amount of data available to researchers for future innovations and discovery. With user-contributed data, there is a risk that users could upload data that contain PII. The goal here is to raise awareness regarding this issue and its potential solutions.

The *Recommendations for Standardized Implementation of Digital Privacy Controls* (U.S. Federal Chief Information Officers Council 2012) expands on a strategy document, *Digital Government: Building a 21st Century Platform to Better Serve the American People* (White House 2012). The two documents refer to the public-sector role in data protection in the following way: "as good stewards of data security and privacy, the federal government must ensure that there are safeguards to prevent the improper collection, retention, use or disclosure of sensitive data such as PII." A more formal definition of PII is provided in the April 2010 special publication by the National Institute of Standards and Technology called the *Guide to Protecting the Confidentiality of Personally Identifiable Information* (McCallister et al. 2010). The two solutions to the PII issue are as follows:

9.3.3.1 Option 1: Reviewing and Managing PII Risk

The strategy and recommendations white papers also defined steps for handling and mitigating risk with PII. How to review PII in the SHRP 2 Archive and whether these are the appropriate procedures to follow to review and manage PII risks should be considered.

1. Define PII and minimize the retention of PII (U.S. Department of Justice 2010):
   a. Complete a Privacy Threshold Analysis (PTA) (U.S. Department of Homeland Security 2012);
   b. Define PII for the SHRP 2 Archive (McCallister et al. 2010);
   c. Determine if data are linked or can be linked ("linkable") to a specific individual;
   d. Use an existing System of Records Notice (SORN) or draft a new SORN, if required (U.S. DOT 2014);
   e. Determine the role of PII in the inventory; and
   f. Determine PII elements that are permitted.
2. Inventory and manage:
   a. Inventory PII in existing files, called an Initial Privacy Assessment (IPA);
   b. Manage PII for existing data by protecting, removing, or making the data not linkable; and
   c. Manage the process of new data sources for PII.
3. Review:
   a. Run periodic reviews of artifacts to determine if PII policies are being enforced.

If it is determined in the PTA that a PIA (Privacy Impact Assessment) is necessary, the National Institute of Standard

and Technology (NIST) recommends asking the following questions in the PIA review (McCallister et al. 2010):

- What information is to be collected?
- Why is the information being collected?
- What is the intended use of the information?
- With whom will the information be shared?
- How will the information be secured?
- What choices has the agency made regarding an IT system or collection of information as a result of performing the PIA?

The NIST guide recommends several other methods that could be used to check whether PII exists in the SHRP 2 Archive and whether files that are added by users contain PII, including "reviewing system documentation, conducting interviews, conducting data calls, using data loss prevention technologies (e.g., automated PII network monitoring tools), or checking with system and data owners" (McCallister et al. 2010). Furthermore, the scope of determining how to manage PII is contingent on the risk associated with the PII data. The NIST guide also reviews how to measure risk for PII, including impact-level definitions for low, moderate, and high risk PII; factors for determining PII confidentiality impact-level procedures of the type recommended by NIST should be followed during the PIA. Following the procedures above would align the SHRP 2 Archive with other federal data sets following the latest guidance and requirements for data protection.

9.3.3.2 Option 2: Defining a Formal PII Process

It will be particularly important to define a process for users to follow and terms to agree to when, in the future, they are allowed upload data sets to the SHRP 2 Archive. When users upload files and data, it is customary to post an agreement to legal terms. The user needs to accept the terms to proceed. The language should indicate that the data being uploaded are free of PII and that the data have become unlinked or anonymous. Additionally, before becoming accessible to users of the Archive, the data should be posted to an administrative area for a review of the data set to determine if the data contain any PII.

Once data are in the Archive, the administrator must serve as a data steward who performs an initial review of all the data. For example, the administrator could use a set of automated tools or manual processes to review the file(s) that will be uploaded. Automated processes to check and ensure anonymous data could be applied for known PII patterns, such as Mac addresses from Bluetooth readers. These processes do not provide a foolproof mechanism, but each step reduces the risks and assigns traceability to the appropriate parties. (Realistically, no one is going to have a reason to go to the time or trouble to download Bluetooth data from the SHRP 2 Archive and try to infer personal information from a trip record for which an equipment identification number has been expunged and then try to go to the next step to link a trip to an individual.)

While the Archive is actively managed, it is desirable to conduct periodic reviews or audits to check the data files for PII, and best practices/policies should be followed.

References

Apache Lucene. 2014. *Apache Solr.* http://lucene.apache.org/solr/. Accessed May 29, 2014.

ASTM International. 2011. *ASTM E2259-03a, Standard Guide for Archiving and Retrieving ITS-Generated Data.* West Conshohocken, Pa.

Bertini, R. 2007. Lessons from Developing an Archived Data User Service in Portland, Oregon: Who Is Using It? Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C.

Bertini, R. L., S. Matthews, S. Hansen, A. Delcambre, and A. Rodriguez. 2005. ITS Archived Data User Service in Portland, Oregon: Now and Into the Future. *Proc., 8th International IEEE Conference on Intelligent Transportation Systems,* Vienna, Austria.

Choe, T., A. Skabardonis, and P. Varaiya. 2002. Freeway Performance Measurement System (PeMS): An Operational Analysis Tool. Presented at 81st Annual Meeting of the Transportation Research Board, Washington, D.C.

Committee for the Strategic Highway Research Program 2: Implementation. 2009. *Special Report 296: Implementing the Results of the Second Strategic Highway Research Program: Saving Lives, Reducing Congestion, Improving Quality of Life.* Transportation Research Board of the National Academies, Washington, D.C.

Courage, K. G., and S. Lee. 2009. *Development of a Central Data Warehouse for Statewide ITS and Transportation Data in Florida, Phase III Final Report.* Transportation Research Center, University of Florida, Gainesville.

Courage, K. G., and S. Lee. 2008. *Development of a Central Data Warehouse for Statewide ITS and Transportation Data in Florida.* Transportation Research Center, University of Florida, Gainesville.

Dailey, D. 2003. *ITS Backbone Infrastructure.* ITS Research Program, Washington State Transportation Center (TRAC), University of Washington, Seattle.

Dailey, D. J., D. Meyers, L. Pond, and K. Guiverson. 2002. *Traffic Data Acquisition and Distribution (TDAD).* ITS Research Program, Washington State Transportation Center (TRAC), University of Washington, Seattle.

Federal Highway Administration. 1998. *Archived Data User Service (ADUS): An Addendum to the* ITS Program Plan. FHWA, U.S. Department of Transportation. September.

Haskins, C. (ed.). 2007. *Systems Engineering Handbook, version 3.1.* INCOSE, Seattle, Wash.

Houston TranStar Consortium. 2010. *Houston TranStar 2010 Annual Report.* http://www.houstontranstar.org/about_transtar/docs/Annual_2010_TranStar.pdf. Accessed April 5, 2012.

Kwon, T. M. 2004. *TMC Traffic Data Automation for Mn/DOT's Traffic Monitoring Program.* University of Minnesota, Duluth.

Lu, C., A. P. Boedihardjo, and J. Zheng. 2006. Towards an Advanced Spatio-Temporal Visualization System for the Metropolitan Washington, D.C. Presented at 2006 TRB International Visualization in Transportation Symposium and Workshop, Denver, Colo.

Margiotta, R. 1998. *ITS as a Data Resource: Preliminary Requirements for a User Service,* Report EDL #3875. Office of Highway Policy Information, FHWA, U.S. Department of Transportation.

McCallister, E., T. Grance, and K. Scarfone. 2010. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).* National Institute of Standard and Technology, Gaithersburg, Md.

Minnesota Department of Transportation. 2012. *Regional Transportation Management Center.* St. Paul, Minn. http://www.dot.state.mn.us/rtmc/overview.html. Accessed April 6.

Petty, K., and T. Barkley. 2011. *Arterial Performance Measurement in the Transportation Performance Measurement System (PeMS).* Report for Metropolitan Transportation Commission, Arterial Operations. Oakland, Calif.

Regiolab-Delft Project. 2012. http://www.regiolab-delft.nl/?q=node/41. Accessed July 30.

Tao, A., J. Spotts, and E. Hess. 2011. *SHRP 2 Report S2-L13-RW-1: Requirements and Feasibility of a System for Archiving and Disseminating Data from SHRP 2 Reliability and Related Studies.* Transportation Research Board of the National Academies. http://onlinepubs.trb.org/onlinepubs/shrp2/SHRP2_S2-L13-RW-1.pdf.

Traffic Data Clearinghouse. 2012. http://trafficdata.iis.u-tokyo.ac.jp/regiolab2.htm. Accessed July 30.

Turner, S. 2007. *Quality Control Procedures for Archived Operations Traffic Data: Synthesis of Practice and Recommendations.* Texas Transportation Institute, The Texas A&M University System, College Station.

Turner, S. 2001. *Guidelines for Developing ITS Data Archiving Systems.* Texas Transportation Institute, The Texas A&M University System, College Station.

University of Maryland. 2012. CATT Laboratory. College Park. http://www.cattlab.umd.edu/index.php?page=about. Accessed April 6.

University of Washington. 2012. ITS Research Program. ITS Research Group. Seattle. http://www.its.washington.edu/its_ws.html#tms. Accessed April 6.

U.S. Department of Homeland Security. 2012. *Privacy Threshold Analysis (PTA).* Washington, D.C. http://www.dhs.gov/xlibrary/assets/privacy/privacy_pta_template.pdf. Accessed May 29, 2014.

U.S. Department of Justice. 2010. *Initial Privacy Assessment (IPA) Instruction and Template.* http://www.usdoj.gov/opcl/initial-privacy-assessment.pdf. Accessed May 29, 2014.

U.S. DOT. 2014. *Privacy Act System of Records Notice.* http://www.dot.gov/individuals/privacy/privacy-act-system-records-notices. Accessed May 29, 2014.

U.S. DOT. 2003. ITS Standards Program. Archived Data User Service (ADUS).

U.S. Federal Chief Information Officers Council. 2012. *Recommendations for Standardized Implementation of Digital Privacy Controls.* https://cio.gov/wp-content/uploads/downloads/2012/12/Standardized_Digital_Privacy_Controls.pdf. Accessed May 29, 2014

White House. 2012. *Digital Government: Building a 21st Century Platform to Better Serve the American People.* Washington, D.C. http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government-strategy.pdf. Accessed on May 29, 2014.

APPENDIX A

Data Dictionary Template

Data Dictionary

Artifact Title(s): *Enter the artifact title.*

A data dictionary is a companion document that describes the data stored in the data set. It is a user guide about the data set file.

Template Instructions

This is a template that principal investigators should use to make their data dictionaries. The following points provide some instructions for completing this template:

- *The italicized text shows the data dictionary instruction text. Please update or delete the italicized text before uploading.*
- *You may upload the same data dictionary to more than one data set if appropriate.*
- *Please add or remove headings as you wish. You could add other headings that briefly explain interesting observations in the data set that corresponds to this data dictionary.*

Background

This data set was *collected/processed* for the SHRP 2 Project XX.

Data Collection

This section is only for detector data (e.g., loop data, weather data, Bluetooth data, cell phone data). Summarize the major points about the data collection. These might include

- *Detector type (Bluetooth, loop, etc.);*
- *Road/road authority that owns the road (e.g., state);*
- *Number of stations;*
- *Date/duration of data collected;*
- *Other relevant information (e.g., whether it was in a construction site, poor weather).*

Processing Techniques

Quickly summarize any processing techniques used, in one to two sentences. Any users who want more background about processing techniques can read the final report.

Column Descriptions

Using Table A.1, include a description of each column, the units of measurement, and any other relevant information. Add as many rows as necessary.

Acknowledgments

Acknowledge those who

- *Supplied the detector data (i.e., road authorities or other organizations);*
- *Primarily processed the data; or*
- *Contributed in other ways (but no personal acknowledgments).*

Table A.1. Column Descriptions

CSV Column Header Title	Ingested Data Set Column Header Title	Column Description	Units of Measurement
SPEED	*Example—speed*	*Average speed of vehicles passing the detector station*	*Miles/hour*
OCC	*Occupancy*	*Average occupancy of the detector station*	*Percentage*

APPENDIX B

Federal System Security Guidelines

- Categorize systems and data:
  - FIPS 199: Standards for Security Categorization of Federal Information and Information Systems;
  - NIST SP 800-60: Volume 1: Guide for Mapping Types of Information and Information Systems to Security Categories (a second volume provides more detail).
- Select security controls:
  - FIPS 200: Minimum Security Requirements for Federal Information and Information Systems;
  - NIST SP 800-53: Recommended Security Controls for Federal Information Systems and Organizations (appendices are available with more detail).
- Implement security controls:
  - NIST SP 800-70: Security Configuration Checklists Program for ITS Products—Guidance for Checklists Users and Developers.
- Assess security controls:
  - NIST SP 800-53A: Guide for Assessing the Security Controls in Federal Information Systems.
- Authorize and monitor security state:
  - NIST 800-37: Guide for Applying the Risk Management Framework to Federal Information Systems.

TRB OVERSIGHT COMMITTEE FOR THE STRATEGIC HIGHWAY RESEARCH PROGRAM 2*

CHAIR: **Kirk T. Steudle,** *Director, Michigan Department of Transportation*

MEMBERS

H. Norman Abramson, *Executive Vice President (retired), Southwest Research Institute*
Alan C. Clark, *MPO Director, Houston–Galveston Area Council*
Frank L. Danchetz, *Vice President, ARCADIS-US, Inc. (deceased January 2015)*
Malcolm Dougherty, *Director, California Department of Transportation*
Stanley Gee, *Executive Deputy Commissioner, New York State Department of Transportation*
Mary L. Klein, *President and CEO, NatureServe*
Michael P. Lewis, *Director, Rhode Island Department of Transportation*
John R. Njord, *Executive Director (retired), Utah Department of Transportation*
Charles F. Potts, *Chief Executive Officer, Heritage Construction and Materials*
Ananth K. Prasad, *Secretary, Florida Department of Transportation*
Gerald M. Ross, *Chief Engineer (retired), Georgia Department of Transportation*
George E. Schoener, *Executive Director, I-95 Corridor Coalition*
Kumares C. Sinha, *Olson Distinguished Professor of Civil Engineering, Purdue University*
Paul Trombino III, *Director, Iowa Department of Transportation*

EX OFFICIO MEMBERS

Victor M. Mendez, *Administrator, Federal Highway Administration*
David L. Strickland, *Administrator, National Highway Transportation Safety Administration*
Frederick "Bud" Wright, *Executive Director, American Association of State Highway and Transportation Officials*

LIAISONS

Ken Jacoby, *Communications and Outreach Team Director, Office of Corporate Research, Technology, and Innovation Management, Federal Highway Administration*
Tony Kane, *Director, Engineering and Technical Services, American Association of State Highway and Transportation Officials*
Jeffrey F. Paniati, *Executive Director, Federal Highway Administration*
John Pearson, *Program Director, Council of Deputy Ministers Responsible for Transportation and Highway Safety, Canada*
Michael F. Trentacoste, *Associate Administrator, Research, Development, and Technology, Federal Highway Administration*

* Membership as of January 2015.

RELIABILITY TECHNICAL COORDINATING COMMITTEE*

CHAIR: **Carlos Braceras,** *Deputy Director and Chief Engineer, Utah Department of Transportation*
VICE CHAIR: **John Corbin,** *Director, Bureau of Traffic Operations, Wisconsin Department of Transportation*
VICE CHAIR: **Mark F. Muriello,** *Assistant Director, Tunnels, Bridges, and Terminals, The Port Authority of New York and New Jersey*

MEMBERS

Malcolm E. Baird, *Consultant*
Mike Bousliman, *Chief Information Officer, Information Services Division, Montana Department of Transportation*
Kevin W. Burch, *President, Jet Express, Inc.*
Leslie S. Fowler, *ITS Program Manager, Intelligent Transportation Systems, Bureau of Transportation Safety and Technology, Kansas Department of Transportation*
Steven Gayle, *Consultant, Gayle Consult, LLC*
Bruce R. Hellinga, *Professor, Department of Civil and Environmental Engineering, University of Waterloo, Ontario, Canada*
Sarath C. Joshua, *ITS and Safety Program Manager, Maricopa Association of Governments*
Sandra Q. Larson, *Systems Operations Bureau Director, Iowa Department of Transportation*
Dennis Motiani, *Executive Director, Transportation Systems Management, New Jersey Department of Transportation*
Richard J. Nelson, *Nevada Department of Transportation*
Richard Phillips, *Director (retired), Administrative Services, Washington State Department of Transportation*
Mark Plass, *District Traffic Operations Engineer, Florida Department of Transportation*
Constance S. Sorrell, *Chief of Systems Operations, Virginia Department of Transportation*
William Steffens, *Vice President and Regional Manager, McMahon Associates*
Jan van der Waard, *Program Manager, Mobility and Accessibility, Netherlands Institute for Transport Policy Analysis*
John P. Wolf, *Assistant Division Chief, Traffic Operations, California Department of Transportation (Caltrans)*

FHWA LIAISONS

Robert Arnold, *Director, Transportation Management, Office of Operations, Federal Highway Administration*
Joe Conway, *SHRP 2 Implementation Director, National Highway Institute*
Jeffrey A. Lindley, *Associate Administrator for Operations, Federal Highway Administration*

U.S. DEPARTMENT OF TRANSPORTATION LIAISON

Patricia S. Hu, *Director, Bureau of Transportation Statistics, U.S. Department of Transportation*

AASHTO LIAISON

Gummada Murthy, *Associate Program Director, Operations*

CANADA LIAISON

Andrew Beal, *Manager, Traffic Office, Highway Standards Branch, Ontario Ministry of Transportation*

* Membership as of July 2014.