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

PREPUBLICATION DRAFT • NOT EDITED
ACKNOWLEDGMENT
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.

NOTICE
The project that is the subject of this document 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 to review this document were chosen for their special competencies and with regard for appropriate balance. The document 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 document are those of the researchers who performed the research. They are not necessarily those of the second Strategic Highway Research Program, the Transportation Research Board, the National Research Council, or the program sponsors.

The information contained in this document was taken directly from the submission of the authors. This document has not been edited by the Transportation Research Board.

Authors herein are responsible for the authenticity of their materials and for obtaining written permissions from publishers or persons who own the copyright to any previously published or copyrighted material used herein.

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.
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, upon 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 Project L13A
Designing the Archive for SHRP 2 Reliability and Reliability-Related Data

Final Report

Prepared for:
second Strategic Highway Research Program
(SHRP 2)

June 2014

Itesis, Inc. – Berkeley California
Rob Hranac, (PI.)

Kittelson & Associates – Oakland, California
Jorge Barrios
CONTENTS

List of Figures ........................................................................................................................................ viii
List of Tables ........................................................................................................................................... xi
Author Acknowledgments ......................................................................................................................... xiii
Abstract .................................................................................................................................................... xiv
Executive Summary ................................................................................................................................... 1
Chapter 1: Background ............................................................................................................................ 4
  1.1 Objective ........................................................................................................................................ 4
  1.2 Target Audience ............................................................................................................................. 4
  1.3 Benefits ......................................................................................................................................... 4
  1.4 Summary of Archived Projects and Artifacts .............................................................................. 5
  1.5 Document Organization .............................................................................................................. 7
Chapter 2: Approach ................................................................................................................................ 9
  2.1 General Approach ......................................................................................................................... 9
  2.2 Software Development Methodology ............................................................................................ 11
    2.2.1 Agile Approach ....................................................................................................................... 11
  2.3 User Engagement .......................................................................................................................... 12
    2.3.1 Objective of the L13A Subject Matter Experts Group ............................................................ 12
    2.3.2 Mode of Engagement ............................................................................................................. 12
    2.3.3 Group Structure ...................................................................................................................... 14
Chapter 3: Preparatory Analysis .............................................................................................................. 17
  3.1 Review of the L13 Report ................................................................................................................ 17
    3.1.1 Summary ................................................................................................................................ 17
    3.1.2 Findings .................................................................................................................................. 18
    3.1.3 System Requirements Proposed by L13 Project ................................................................. 19
    3.1.4 User Interfaces ....................................................................................................................... 19
    3.1.5 Data Integrity and Quality ..................................................................................................... 20
    3.1.6 Data Rights ............................................................................................................................ 20
    3.1.7 Institutional Framework and Governance ........................................................................... 21
    3.1.8 Technical Issues ..................................................................................................................... 21
    3.1.9 Establishing Solution Alternatives ....................................................................................... 22
    3.1.10 Solution Components and Implementation Approaches ................................................. 22
    3.1.11 Life-Cycle Costs Analysis ................................................................................................. 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.5</td>
<td>Explore by Focus Area / Project</td>
<td>97</td>
</tr>
<tr>
<td>6.3.6</td>
<td>View Latest Artifacts</td>
<td>98</td>
</tr>
<tr>
<td>6.3.7</td>
<td>View Your Own Artifacts</td>
<td>98</td>
</tr>
<tr>
<td>6.4</td>
<td>Working With Artifacts</td>
<td>99</td>
</tr>
<tr>
<td>6.4.1</td>
<td>View Metadata</td>
<td>99</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Visualize Data</td>
<td>99</td>
</tr>
<tr>
<td>6.5</td>
<td>Download an Artifact</td>
<td>108</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Full Download</td>
<td>108</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Download Filtered Artifact</td>
<td>108</td>
</tr>
<tr>
<td>6.6</td>
<td>Discussion</td>
<td>108</td>
</tr>
<tr>
<td>6.6.1</td>
<td>Comment</td>
<td>109</td>
</tr>
<tr>
<td>6.6.2</td>
<td>Rating Artifacts</td>
<td>109</td>
</tr>
<tr>
<td>6.7</td>
<td>Uploading Artifacts</td>
<td>109</td>
</tr>
<tr>
<td>6.7.1</td>
<td>Who Can Upload Artifacts?</td>
<td>109</td>
</tr>
<tr>
<td>6.7.2</td>
<td>Artifact Upload Wizard</td>
<td>109</td>
</tr>
<tr>
<td>6.7.3</td>
<td>What Happens After the User Submits an Artifact?</td>
<td>113</td>
</tr>
<tr>
<td>6.7.4</td>
<td>Unsuccessful Artifact Processing</td>
<td>113</td>
</tr>
<tr>
<td>6.7.5</td>
<td>Edit Artifact Metadata</td>
<td>113</td>
</tr>
<tr>
<td>6.7.6</td>
<td>Delete Artifacts</td>
<td>114</td>
</tr>
<tr>
<td>6.8</td>
<td>Automatically Generated Email Notifications</td>
<td>114</td>
</tr>
<tr>
<td>6.8.1</td>
<td>Validate the Identity of New User</td>
<td>114</td>
</tr>
<tr>
<td>6.8.2</td>
<td>Artifact is Processing</td>
<td>115</td>
</tr>
<tr>
<td>6.8.3</td>
<td>Artifact has Finished Processing</td>
<td>115</td>
</tr>
<tr>
<td>6.8.4</td>
<td>Artifact Processing was Unsuccessful</td>
<td>116</td>
</tr>
<tr>
<td>6.8.5</td>
<td>Artifact has been Removed</td>
<td>116</td>
</tr>
<tr>
<td>6.8.6</td>
<td>Someone has Commented on a Principal Investigator’s or Creator’s Artifact</td>
<td>117</td>
</tr>
</tbody>
</table>

Chapter 7: System High-Level Architecture | 118 |

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Amazon Web Services (AWS)</td>
<td>118</td>
</tr>
<tr>
<td>7.2</td>
<td>WordPress</td>
<td>119</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Themes</td>
<td>119</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Plugins</td>
<td>120</td>
</tr>
<tr>
<td>7.3</td>
<td>MySQL Database</td>
<td>121</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Database Diagram</td>
<td>121</td>
</tr>
</tbody>
</table>
7.3.2 Overview of Database Tables ................................................................. 122
7.4 Solr Search Engine Server ......................................................................... 125
7.5 S2A Server ................................................................................................. 125

Chapter 8: Test Plan .......................................................................................... 128
8.1 Development Testing Strategy ..................................................................... 128
  8.1.1 Unit Testing ......................................................................................... 128
8.2 System Tests ............................................................................................... 129
  8.2.1 Functionality Tests .............................................................................. 129
  8.2.2 Performance Tests ............................................................................... 131
  8.2.3 Availability Tests ................................................................................ 133
  8.2.4 Scalability Tests .................................................................................. 134
  8.2.5 Maintainability Tests .......................................................................... 134
  8.2.6 Integration Tests .................................................................................. 135
  8.2.7 Security Tests ...................................................................................... 135
  8.2.8 Visual GUI Testing ............................................................................. 136
  8.2.9 Testing Automation ............................................................................. 137
8.3 List of Artifacts Needed to Run the Test Plan ............................................ 138

Chapter 9: Notes on Operation and Maintenance of the Archive ..................... 139
9.1 Inclusion of User-Submitted Data in the Archive ....................................... 139
  9.1.1 What are the Submission Scenarios? .................................................. 139
  9.1.2 Comparison ....................................................................................... 140
  9.1.3 Conclusion ......................................................................................... 140
9.2 Key Issues Associated with Operation and Maintenance of the Archive ...... 143
  9.2.1 Who is Going to Operate and Maintain the Archive? ......................... 143
  9.2.2 What are System and Operations Requirements? .............................. 143
  9.2.3 What are Hosting Options for the Archive? ....................................... 144
  9.2.4 What are Operation & Maintenance Costs? ...................................... 147
9.3 Managing Data Rights and Personally Identifiable Information ............... 150
  9.3.1 Open Data ......................................................................................... 150
  9.3.2 Open Data Licensing Options .......................................................... 151
  9.3.3 Personally Identifiable Information ................................................... 153

References ....................................................................................................... 156

Appendix A Data Dictionary Template ........................................................... A-1
Figure 6-15: Graph View .......................................................... 102
Figure 6-16: Displaying Data Series.............................................. 103
Figure 6-17: Types of Graphs ....................................................... 103
Figure 6-18: Printing and Downloading Options ......................... 104
Figure 6-19: Map View ................................................................ 105
Figure 6-20: Cluster Detector Stations ........................................ 105
Figure 6-21: Defining Filter Criteria ............................................. 106
Figure 6-22: Slider Filter ............................................................... 107
Figure 6-23: Number Filter ........................................................... 107
Figure 6-24: Text Filter ................................................................. 107
Figure 6-25: Example of Search with AND & OR Operators .......... 108
Figure 6-26: Rating Artifacts ......................................................... 109
Figure 6-27: Uploading an Artifact ............................................... 110
Figure 6-28: Artifact Upload Wizard – Step 2 ............................ 111
Figure 6-29: Example of Related Data sets ................................. 112
Figure 6-30: Editing an Artifact’s Metadata ............................... 113
Figure 6-31: Deleting an Artifact .................................................. 114
Figure 7-1: Components of SHRP 2 Archive.............................. 118
Figure 7-2: SHRP 2 Database Diagram ...................................... 122
Figure 7-3: The Archive Processing Finite State Diagram ............ 127
Figure 9-1: L13A Options Uptime Risks ...................................... 149
Figure 9-2: Creative Commons Public Domain License ............... 151
Figure 9-3: Creative Commons License Types ............................. 152
Figure 10-1: SHRP 2 Reliability Logo ........................................... Error! Bookmark not defined.
Figure 10-2: SHRP2Archive.org Flyer Concept ........................... Error! Bookmark not defined.
LIST OF TABLES

Table 1-1: List of Archived/To Be Archived Projects ....................................................... 5
Table 1-2: Summary of Artifacts ........................................................................................ 6
Table 2-1: Technical Expert Task Group Members ......................................................... 15
Table 2-2: External Advisory Panel Members ................................................................. 16
Table 3-1: Guiding Principles for ADMS Development .................................................. 27
Table 3-2: Needs of ADUS Stakeholders ......................................................................... 28
Table 3-3: Archived Data Levels ..................................................................................... 44
Table 3-4: Sample of Existing Online Data Archives Focused on Research ................. 47
Table 3-5: Data Archival Technologies ........................................................................... 50
Table 4-1: User roles ........................................................................................................ 60
Table 5-1 - Site Metadata ................................................................................................. 80
Table 5-2 - Focus Area Metadata ..................................................................................... 80
Table 5-3 - Project Metadata ............................................................................................ 81
Table 5-4 - User Metadata ................................................................................................ 82
Table 5-5 – General Metadata Element Set ..................................................................... 84
Table 5-6 – Data Set Metadata Element Set .................................................................. 85
Table 5-7 – Data Set Preparation Checklist ..................................................................... 86
Table 6-1: Metadata information required during the Upload Wizard ........................... 111
Table 7-1: List of open source JavaScript Libraries ....................................................... 120
Table 7-2: List of plugins used in the Archive ............................................................... 120
Table 7-3: List of the SHRP 2 Database Tables ............................................................. 122
Table 7-4: s2_artifact_relations table fields .............................................................. 123
Table 7-5: S2_col_types table fields .............................................................................. 123
Table 7-6: S2_dset_ArtifactID table fields .................................................................... 124
Table 7-7: wp_posts table fields ..................................................................................... 124
AUTHOR 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 and Associates. Robert Hranac, Ali Mortazavi, and Karl Petty of Iteris, Inc., were the Principal Investigators (PIs). 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, Dr. Alex Skabardonis Dr. Senanu Ashiabor, Dr. Rick Dowling, and Daphne Dethier of Kittelson and Associates.

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.
ABSTRACT

This Strategic Highway Research Program (SHRP 2) L13A report, titled “Designing the Archive for SHRP 2 Reliability and Reliability-Related Data” documents and presents the results of a project to develop, test and deploy the SHRP 2 Reliability Archive system.

This Archive is a repository that stores the rich data and information from SHRP 2 Reliability projects and Reliability-related projects. These projects produce various resources, such as raw datasets, analysis results, tools and models, and documentation. The SHRP 2 Reliability Archive makes these valuable artifacts accessible to researchers and practitioners via the Internet. In this way, the research findings, raw data, models, tools, and other archive-worthy artifacts can be leveraged again and again by others, multiplying the value of the research conducted in this program’s focus area.

The SHRP 2 Reliability Archive enables uploading of artifacts and searching for artifacts. The system is a user-friendly toolset that facilitates visualizations of user-selected data and collaborations between multiple researchers.

This report reviews the SHRP 2 Reliability Archive system features and how users can work with the system. It provides high-level technical information about the system and discusses the operation and maintenance challenges.
EXECUTIVE SUMMARY

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

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, thereby increasing the value of the data itself 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 Project L13 report and past work on data archiving systems and technologies. This analysis resulted in the selection of WordPress as the core CMS (Content Management System), on which the Archive system was built upon. 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, where needs and requirements evolved 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 (T-ETG) 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 operation and maintenance of the Archive.
• 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 five hundred artifacts collected from more than 30 Reliability-related projects. The approximate size of the SHRP 2 Archive is anticipated to reach about one terabyte. A detailed user guide 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 system is more than a data repository. It is a system that enables uploading of artifacts, searching with results arranged by list or map as well as 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 SHRP2 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, etc.). Artifacts submitted under each category need to meet certain requirements. Therefore, the producer of an artifact has to pre-process it before submitting it to the system.

![Figure ES-1: SHRP 2 Reliability Archive System](image-url)
• **Search and Download:** The Archive provides faceted and text search tools to help the 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 the users to explore the Archive’s repository. The users can filter the search results by selecting various related criteria. The system provides the search results on a map and in a list. One can also search by projects within a focus area.

• **Map Visualization:** The system has the capability to map the geo-location information of the traffic detectors provided in a data set. Thus the geo-filter and map view capabilities can help users explore sensor locations.

• **Data Visualization:** The Archive 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) 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, the users can save and print the plots for future use.

• **Collaboration:** The 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, the users have to be registered.

• **Administration and User Profile:** The Archive administrator is capable of granting access to the 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.

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 (hereinafter 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 (hereinafter 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;
- 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 (non-sensitive information only) for SHRP 2 Reliability research. 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. 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 your needs and for downloading only the information that the user desires;
Quick registration process. 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. “Non-sensitive” data is provided so that it 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 over 35 Reliability and Reliability-related projects. At the time of writing this report, a total of 526 Reliability-related artifacts have been stored or have been 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.

Table 1-1: List of Archived/To Be Archived Projects

<table>
<thead>
<tr>
<th>Project Name</th>
<th>Reliability Project Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>L01</td>
<td>Integrating Business Processes to Improve Reliability</td>
</tr>
<tr>
<td>L02</td>
<td>Establishing Monitoring Programs for Mobility and Travel Time Reliability</td>
</tr>
<tr>
<td>L03</td>
<td>Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies</td>
</tr>
<tr>
<td>L04</td>
<td>Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools</td>
</tr>
<tr>
<td>L05</td>
<td>Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes</td>
</tr>
<tr>
<td>L06</td>
<td>Institutional Architectures to Advance Operational Strategies</td>
</tr>
<tr>
<td>L07</td>
<td>Evaluation of Cost-Effectiveness of Highway Design Features</td>
</tr>
<tr>
<td>L08</td>
<td>Incorporation of Travel Time Reliability into the Highway Capacity Manual</td>
</tr>
<tr>
<td>L09</td>
<td>Incorporation of Non-recurrent Congestion Factors into the AASHTO Policy on Geometric Design</td>
</tr>
<tr>
<td>L10</td>
<td>Feasibility of Using In-Vehicle Video Data to Explore How to Modify Driver Behavior that Causes Non-Recurring Congestion</td>
</tr>
<tr>
<td>L11</td>
<td>Evaluating Alternative Operations Strategies to Improve Travel Time Reliability</td>
</tr>
<tr>
<td>L12</td>
<td>Improving Traffic Incident Scene Management</td>
</tr>
<tr>
<td>L13</td>
<td>Archive for Reliability and Related Data</td>
</tr>
<tr>
<td>L13A</td>
<td>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</td>
</tr>
<tr>
<td>L14</td>
<td>Traveler Information and Travel Time Reliability</td>
</tr>
<tr>
<td>L15</td>
<td>Innovative IDEA Projects</td>
</tr>
<tr>
<td>L16</td>
<td>Assistance to Contractors to Archive Their Data for Reliability and Related Projects (Combined with L13A)</td>
</tr>
<tr>
<td>L17</td>
<td>A Framework for Improving Travel Time Reliability</td>
</tr>
<tr>
<td>L32B</td>
<td>e-Learning for Training Traffic Incident Responders and Managers</td>
</tr>
<tr>
<td>L32C</td>
<td>Post-Course Assessment and Reporting Tool for Trainers and TIM Responders Using the SHRP 2 Interdisciplinary Traffic Incident Management Curriculum</td>
</tr>
<tr>
<td>L33</td>
<td>Validation of Urban Freeway Models</td>
</tr>
</tbody>
</table>
**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

### Project Name | Capacity Project Title
---|---
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
C10B | Partnership to Develop an Integrated Advanced Travel Demand Model with Mode Choice Capability and Fine-Grained, Time-Sensitive Networks
C11 | Development of Improved Economic Analysis Tools Based on Recommendations from project C03 (2)

### Project Name | Renewal Project Title
---|---
R11 | Strategic Approaches at the Corridor and Network Levels to Minimize Disruption from the Renewal Process

### Project Name | Safety Project Title
---|---
S04A | Roadway Information Database Development and Technical Coordination and Quality Assurance of the Mobile Data Collection Project (S04B)

### Table 1-2: Summary of Artifacts

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

* Data is not available at this time

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 develop the Archive;
- **Chapter 3: Preparatory Analysis** presents the results of the project team’s literature review and preparatory analysis;
- **Chapter 4: System and User Needs and Requirements** discusses the general user needs and system requirements upon which the system was built;
• Chapter 5: Artifact Upload explains processes to prepare and upload artifacts into the Archive;
• Chapter 6: Working With The Archive – User Guide 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;
• Chapter 9: Notes on Operation and Maintenance of the Archive discusses operation and maintenance-related challenges;
• Error! Reference source not found. Error! Reference source not found. 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 that users may refer to in order to provide information regarding data sets.
• Appendix B: Federal System Security Guidelines summarizes the federal requirements for system security.
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 right 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 they are developing are worth the value they deliver.

To meet the aforementioned characteristics critical to the success of the L13A project, Iteris’ 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 the users. This method is based on incremental development, where 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.

Our plan was to develop the Archive system and transfer the final product to SHRP 2 through six phases. Figure 2-1 shows our proposed approach. The proposed phases were:

**Phase I – 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.

**Phase II – Prototyping and Acceptance Testing:**

The team developed an archive prototype in Phase II based on the user stories defined in Phase I. 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 (T-ETG) to answer key questions regarding the operation and maintenance of the Archive system. This collaborative effort was extremely helpful for the development team to make key decisions regarding the Archive system features and specifications. At the end of Phase II, the completed prototype was demonstrated to the stakeholders (Decision Gate 1) to gain their approval for starting Phase III (Archive Development).
Phase III – Archive Development:

Phase III focused on delivering an operating archive system and finalizing user interface and backend 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 T-ETG members’ approval to release the system for operation (Decision Gate 2).

Phase IV – Outreach and Training:

The objective of this phase was to perform outreach and training activities after release of the Archive; to promote the site through social media, academic conferences or any transportation-related venue that discusses highway systems management and operation. Due to significant workload associated with the artifact upload task, this phase was scaled down to save budget for the upload activities.

Phase V – Operation and User Engagement:

Phase V 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.
Phase VI – 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 together, to develop and deploy the Archive. The general development process was based on prototyping methodology where 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. On the other hand, at the software coding and system design level, the development was performed based on the agile approach.

2.2.1 Agile Approach

Our team utilized 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.

Agile process is iterative and relies on user/stakeholder continuous feedback 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 2-1). It should be noted that system and user requirements were not identified at the inception of the project. This fact 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.
2.3 USER ENGAGEMENT

As part of Phase I, Phase II, and Phase III, the L13A project team established a collaborative effort in which a group (so called SME Group) consisting of the T-ETG 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

In order 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;
- Utilize outside expertise and advice on system design and architecture;
- Obtain user feedback on the system requirements and interface;
- 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 utilized 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;
- 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;
- Archive system design review;
- Prototype;
- Final system acceptance test.

Figure 2-3: Stakeholder Engagement: Dates and Objectives
2.3.2.2 **Stakeholder Website**

This is the SHRP 2 L13A project-restricted website that is available only to the SHRP 2 clients, the contractor team, and the T-ETG. The website can be accessed by those with proper permissions at [http://sites.kittelson.com/SHRP2_ReliabilityDataArchive](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.

![Figure 2-4: Stakeholder Website](image)

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 the users that could help the project team develop a user-centric system successfully. The L13A SME group comprised of (see Figure 2-5):

- T-ETG group (see Table 2-1).
- External advisory panel:
  - Senior researchers from major academic transportation centers.
  - Practitioners: Departments of Transportation (DOTs) (state and federal).
  - Private sector researchers.
- Iteris and Kittelson project management team.

Figure 2-5: L13A Subject Matter Expert Group Structure

Table 2-1: Technical Expert Task Group Members

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Richard T. Goeltz</td>
<td>Oak Ridge National Laboratory</td>
</tr>
<tr>
<td>Michael L. Pack</td>
<td>CATT Lab, University of Maryland</td>
</tr>
<tr>
<td>William H. Schneider</td>
<td>The University of Akron</td>
</tr>
<tr>
<td>Dustin Sell</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Theodore J. Trepanier</td>
<td>INRIX</td>
</tr>
<tr>
<td>Kristin A. Tuft</td>
<td>Portland State University</td>
</tr>
<tr>
<td>Marcus Ramsay Wigan</td>
<td>Oxford Systematics</td>
</tr>
<tr>
<td>Mike Bousliman</td>
<td>Montana Department of Transportation</td>
</tr>
</tbody>
</table>

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;
• Being potential users from different areas.

Table 2-2: External Advisory Panel Members

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian Hoef</td>
<td>Las Vegas RTC FAST</td>
</tr>
<tr>
<td>Dale Thompson</td>
<td>FHWA</td>
</tr>
<tr>
<td>Daniela Bremmer</td>
<td>Washington DOT</td>
</tr>
<tr>
<td>Gene McHale</td>
<td>FHWA</td>
</tr>
<tr>
<td>Heng Wei</td>
<td>University of Cincinnati</td>
</tr>
<tr>
<td>James Hall</td>
<td>University of Illinois</td>
</tr>
<tr>
<td>John Shaw</td>
<td>Wisconsin DOT</td>
</tr>
<tr>
<td>Ken Courage</td>
<td>University of Florida</td>
</tr>
<tr>
<td>Mark Hallenbeck</td>
<td>University of Washington</td>
</tr>
<tr>
<td>Mei Chen</td>
<td>University of Kentucky</td>
</tr>
<tr>
<td>Nazy Sobhi</td>
<td>FHWA</td>
</tr>
<tr>
<td>Paul Pisano</td>
<td>FHWA</td>
</tr>
<tr>
<td>Saman Moshafi</td>
<td>IndraSoft Inc.</td>
</tr>
<tr>
<td>Steven Beningo</td>
<td>FHWA</td>
</tr>
<tr>
<td>Steven Parker</td>
<td>University of Wisconsin</td>
</tr>
<tr>
<td>Walter During</td>
<td>FHWA</td>
</tr>
</tbody>
</table>
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 utilized 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.
2. Identify an existing Commercial Off-the-Shelf (COTS) content management system on which the Archive system can be built (Section 3.5.3).

3.1 REVIEW OF THE L13 REPORT

3.1.1 Summary


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

- 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 the criteria above, 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.

Out 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. Based on this finding the recommended prototype was a cloud storage system, which uses a pay-as-you-go, web-based access model.

The research team found that the in-house alternatives, on the other hand, 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 and effectiveness of archiving digital resources had 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 as a reference model. 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 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;
- 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 cannot be thought of as a database, where the structure is known up front. The team proposed that the conceptual design for the archival system follows 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 towards the end of each project, would need to commence early with 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 pre-accessioning work has been performed, the six core archive functions of ingest, 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. It was also expected that consumers would 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. Hence, 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;
- Non-transportation professionals;
- Researchers/analysts.

Managers of transportation agencies are interested in business processes, strategies, institutional structures, and performance measures. They need to quickly find conclusions of each project, executive summaries and presentations. The technical staff of transportation agencies is interested in various Reliability products, such as data sets, tools, and reports. They need to quickly find end products of the projects, which may be organized and grouped by categories such as planning, design, and operations. Non-transportation 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 three general categories of Planning, Design, and Operations;
- 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;
- 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 its quality control standards, methods and assessment. The L13 report suggested Reliability project L13A would be responsible for reviewing the data quality assessment and 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 authors 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. They 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 access protection to data 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 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 RFP. The issues were:

- Data normalization and denormalization;
- Online analytical processing (OLAP) and user-defined functions;
- Service-oriented architectures (SOA);
- 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 does not have any application in the proposed Archive in terms of the post-research part of the process of preparing it for preservation.

#### 3.1.8.2 Online Analytical Processing (OLAP) 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 (SOA) 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. It was found to be 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 the Proposed Archival System](image)

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;
- 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 vs. proprietary hardware, Open Source Software (OSS), in-house developed software, hosting, and software and storage 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 where the need is short-term. Based on the L13 report, community-supported OSS should also only be considered under similar circumstances since 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. It was recommended that cloud storage be considered since the cost of acquiring and managing storage is likely the single largest cost of the system’s lifetime.

The visioning and filtering process that the L13 research team went through led to the conceptual solution framework as shown in Figure 3-2.

![Figure 3-2: Solution Framework](image)

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 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 like 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, semi-automatic and manual workflows resulting in editing, deleting, and approving the content prior to 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 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 XLS 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 alternative is shown in Figure 3-5.

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).

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, maintenance as well as periodic upgrades to accommodate technology advances and obsolescence.

3.1.12 Life-Cycle Costs of the Alternatives

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 one and a half years.

3.2 ARCHIVED DATA USER SERVICES

The United States Department of Transportation included Archived Data User Services (ADUS) in the National Intelligent Transportation System (ITS) Architecture in 1999, envisioning “the unambiguous interchange and reuse of data and information throughout all functional areas.” 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 U.S.

3.2.1 Introduction to Federal Highway Administration ADUS Guidelines

The Federal Highway Administration (FHWA) funds and monitors many state ADUS programs. In the past ten 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 (USDOT 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;
- 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 importantly, 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 American Society for Testing and Materials (ASTM) to devise national ADUS standards (ASTM 2011). The ASTM report is focused on the technical considerations of implementing an ADUS system, which they refer to as an Archived Data Management System (ADMS). ASTM developed ten guiding
principles, which it grouped based on 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

<table>
<thead>
<tr>
<th>Acquiring Data</th>
<th>Managing the ADMS (ADUS)</th>
<th>Retrieving Data and Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manage the archive to account for data quality</td>
<td>Provide security for the ADMS</td>
<td>Process user requests for data</td>
</tr>
<tr>
<td>Get archived data from other centers</td>
<td>Specify and maintain metadata to support the ADMS</td>
<td>Support analysis of the archived data</td>
</tr>
<tr>
<td>Manage the interfaces of the Archive Data Administrator</td>
<td>Manage the interfaces of the Archive Data Administrator</td>
<td>Prepare data for government reporting systems</td>
</tr>
<tr>
<td>Integrate selected other transportation data including roadside data collection</td>
<td>Interact with other archives and monitor other standards</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from (ASTM 2011)

A large portion of Rich Margiotta’s April 1998 report to the FHWA’s Office of Highway Information is dedicated to ADUS’ “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 is 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 practice and to equivalent solutions available from ADUS systems. The table was compiled by Margiotta and published in (Margiotta 1998).
### Table 3-2: Needs of ADUS Stakeholders

<table>
<thead>
<tr>
<th>Stakeholder Group</th>
<th>Application</th>
<th>Method or Function</th>
<th>Current Data</th>
<th>ITS-Generated Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan Planning Organization (MPO) and State Transportation Planners</td>
<td>Congestion Management Systems</td>
<td>Congestion Monitoring</td>
<td>Travel times collected by &quot;floating cars:&quot; usually only a few runs (small samples) on selected routes. Speeds and travel times synthesized with analytic methods (e.g., <em>Highway Capacity Manual</em>, simulation) using limited traffic data (short counts). Effect of incidents missed completely with synthetic methods and minimally covered by floating cars.</td>
<td>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 &quot;floating cars&quot; 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.</td>
</tr>
<tr>
<td>Long-Range Plan Development</td>
<td>Travel Demand Forecasting (TDF) Models</td>
<td>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.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Corridor Analysis</td>
<td>Traffic Simulation Models</td>
<td>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).</td>
<td>Most input data can be collected automatically and models can be directly calibrated to actual conditions.</td>
<td></td>
</tr>
<tr>
<td>Traffic Management Operators</td>
<td>ITS Technology</td>
<td>Program and Technology Evaluations</td>
<td>Extremely limited; special data collection efforts required.</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Pre-Determined Control Strategies</td>
<td>Short-duration traffic counts and &quot;floating car&quot; travel time runs. A limited set of pre-determined control plans is usually developed mostly due to the lack of data.</td>
<td>Continuous roadway surveillance data makes it possible to develop any number of pre-determined control strategies.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predictive Traffic Flow Algorithms</td>
<td>Extremely limited.</td>
<td>Analysis of historical data form the basis of predictive algorithms: &quot;What will traffic conditions be in the next 15 minutes?&quot; (Bayesian approach).</td>
<td></td>
</tr>
<tr>
<td>Transit Operators</td>
<td>Operations Planning</td>
<td>Routing and Scheduling</td>
<td>Manual travel demand and ridership surveys; special studies.</td>
<td>Electronic Fare Payment System and Automatic Passenger Counters allow continuous boardings to be collected.</td>
</tr>
<tr>
<td>Role/Agency</td>
<td>Task/Analysis</td>
<td>Current Methods/Technologies</td>
<td>Practical Source/Innovation</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Computer-aided dispatch systems</strong></td>
<td>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.</td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air Quality Analysts</td>
<td>Conformity Determinations</td>
<td>Analysis with the MOBILE Model Areawide speed data taken from TDFs. Vehicle Miles Traveled (VMT) and vehicle classifications derived from short counts.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>MPO/State Freight and Intermodal Planners</td>
<td>Port and Intermodal Facilities Planning</td>
<td>Freight Demand Models Data collected through rare special surveys or implied from national data (e.g., Commodity Flow Survey).</td>
<td>Electronic credentialing and AVI allows 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.</td>
<td></td>
</tr>
<tr>
<td>Safety Planners and Administrators</td>
<td>Safety Management Systems</td>
<td>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.</td>
<td>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).</td>
<td></td>
</tr>
<tr>
<td>Maintenance Personnel</td>
<td>Pavement and Bridge Management</td>
<td>Historical and Forecasted Loadings Volumes, vehicle classifications, and vehicle weights derived from short-duration counts (limited number of continuously operating sites).</td>
<td>Roadway surveillance data provide continuous volume counts, vehicle classifications, and vehicle weights, making more accurate loading data and growth forecasts available.</td>
<td></td>
</tr>
<tr>
<td>Commercial vehicle enforcement personnel</td>
<td>Enforcement of Commercial Vehicle Regulations</td>
<td>Hazardous Material Inspections and Emergency Response Extremely limited.</td>
<td>Electronic credentialing and AVI allows tracking of hazardous material flows, allowing better deployment of inspection and response personnel.</td>
<td></td>
</tr>
<tr>
<td>Emergency Management Services (local police, fire, and emergency medical)</td>
<td>Incident Management</td>
<td>Emergency Response Extremely limited.</td>
<td>Electronic credentialing and AVI allows tracking of truck flows and high incident locations, allowing better deployment of response personnel.</td>
<td></td>
</tr>
<tr>
<td>Transportation Researchers</td>
<td>Travel Behavior Models</td>
<td>Mostly rely on infrequent and costly surveys: stated preference and some travel diary efforts (revealed preference).</td>
<td>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 imbedded in these technologies as well.</td>
<td></td>
</tr>
<tr>
<td>Traffic Flow Models</td>
<td>Detailed traffic data for model development must be collected through special efforts.</td>
<td>Roadway surveillance data provide continuous volume counts, densities, truck percents, and speeds at very small time increments. GPS-instrumented vehicles can provide second-by-second performance characteristics for microscopic model development and validation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private Sector</td>
<td>Truck Routing and Dispatching</td>
<td>Congestion Monitoring Current information on real-time or near real-time congestion is extremely limited.</td>
<td>Roadway surveillance data and probe vehicles can identify existing congestion and can be used to show historical patterns of congestion.</td>
<td></td>
</tr>
</tbody>
</table>
3.2.2 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. On the other hand, 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 US 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 setup to process 30-second loop 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, 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;
- 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/](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 is streamed
to the server located at PSU and are then archived in a RDBMS. This system is known as the Portland Transportation Archive Listing (PORTAL).

The PORTAL system is focused mainly on freeway data. One of the design goals of the PORTAL system has been to adhere to the National ITS Architecture. In line with this, the PORTAL system includes a detailed metadata repository and maintains meta-schema for all data entering the system, including information generated in the field at the controller and in the traffic management center.

The PORTAL 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 covers the Portland-Vancouver metropolitan region. The current PORTAL 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 Portal-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:

- **Systems** – 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. These data can be obtained by selecting date and time ranges. The resolution of these data is available in 5-min, 15-min, 1-hour, monthly, and yearly basis.
- **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 include freeway transit and arterial data for the I-205 corridor in Portland, Oregon. The selected corridor is approximately 10 miles 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, configuration errors, etc.
3.2.3.3 Center for Advanced Transportation Technology Laboratory, Maryland

The University of Maryland Center for Advanced Transportation Technology Laboratory (CATT Lab) builds, operates, and maintains the transportation data archive for Washington Metropolitan Area and other states (University of Maryland 2012). Their system is called the Regional Integrated Transportation Information System (RITIS). The data include volume, speed, incidents, weather, system delays, etc., and 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.

![Figure 3-7: Screenshot Showing Data Selection Options in RITIS](image)

3.2.3.4 Center for Transportation Studies, Virginia

The “ADMS Virginia” project is hosted at the Smart Travel Laboratory, a joint facility of Virginia Department of Transportation and University of Virginia. The ADMS Virginia is a development
effort to archive ITS data for transportation applications. The web-based system uses historic 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 Virginia ADMS users need to have an account. The account can be requested online via email at the project website.

3.2.3.5 Advanced Interactive Traffic Visualization Systems, Virginia

Virginia Tech’s Spatial Data Management Lab has developed the Advanced Interactive Traffic Visualization System (AIVS) that provides real-time highway monitoring capabilities via comprehensive visualization components. AIVS provides a rich set of multidimensional visual components for real-time and historical traffic data analyses (Lu et al. 2006).

The AIVS 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 vs. time plot, highway stations vs. day of the week plot, and time vs. day of the week plot (Lu et al. 2006). The speed profile, volume, and occupancy plot can be obtained by selecting pairs of stations as shown in Figure 3-9.
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 Automated Vehicle Identification (AVI) information. In addition, Closed-Circuit Television (CCTV) cameras cover 335 freeway centerline miles. TranStar has been archiving 15-minute 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 since May 2002.

Houston TranStar provides information for multi-agency 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.

![Figure 3-9: Sample Plots of Volume, Speed, and Occupancy from AITVS](image)
3.2.3.7 Traffic Data Acquisition and Distribution, Washington

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). The Washington TDAD makes its historical data available online.

The Washington TDAD obtains its data from loop detectors across the state, which report volume and occupancy at 20-second intervals. TDAD depends upon the Washington ITS “Backbone Project” to obtain the 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 Itersis, Wavetronix, HERE (formerly NAVTEQ), and AT&T have developed applications to continuously download, process, and re-use the WSDOT data. Unfortunately, according to the University of Washington’s ITS website, the funding for the data feed has not been renewed and thus the ADUS is unavailable at the moment. This is an example of what can happen if adequate funding is not set aside for operation and maintenance when an ADUS system is initially designed.

3.2.3.8 Minnesota DOT Regional Transportation Management Center

MnDOT built the original transportation management center in 1972 to manage the freeway system in the Twin Cities metro 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 miles 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 of 30-second interval loop detector count and other traffic data from the site to the server located at the RTMC. The data is continuously archived and more than six years is available for download. Lane-by-lane traffic data including volume, speed, occupancy, headway, and density are collected from the permanent loop detectors. The data is 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 the 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.

*The arrow lines indicate Internet data connections, and the sequence of data flow

Figure 3-11: System Level Concept of Data Automation (MnDOT)

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, 15, and 60 minutes. STEWARD can be accessed at http://ece-trc-cdwserv.ce.ufl.edu/steward/.
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 Florida Database](image)

**Figure 3-12: STEWARD Florida Database**

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 (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;
- Provide data for research and special studies.

The existing STEWARD database contains Traffic Sensor Subsystem data from all TMC stations over a 24-hour 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 consists mainly of minute data from inductive loop detectors and variable message signs on the national highways in the province of South Holland. Traffic data is collected from detectors on approximately every 500-meter interval on motorways. In addition to the loop detectors, local data from traffic control systems and camera's in the municipality of Delft is 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 is available for download to registered researchers from the Regiolab website. The website provides a Matlab Toolbox (program 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. The project website displays 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 Delft RegioLab project. The aim is to attract more partners and researchers to share their data sets in order 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](http://trafficdata.iis.u-tokyo.ac.jp/index.php). A map of Regiolab in 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, is 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 message signs. The purpose of this service is to help the motoring public to 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 is 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

In order 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) 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.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Example Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Raw data, including raw traffic data such as volumes and speeds.</td>
<td>Raw digital data and imagery</td>
</tr>
<tr>
<td>Level 1</td>
<td>Geo-referenced data, such as speed associated with a specific route and direction.</td>
<td>Individual records, processed images</td>
</tr>
<tr>
<td>Level 2</td>
<td>Derived variables at the same resolution and location as the Level 1 source data from which the variables are derived.</td>
<td>Individual records, processed images</td>
</tr>
</tbody>
</table>
### Level Description Example Formats

#### 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

### 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. CMS 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 to support individual user requirements implies that data is organized, catalogued and stored such that it can be accessed according to what data is 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:


Early examples of transportation document/data archives were relatively simple web sites 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 transportation archive systems were 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 the weather and social science archives are more focused 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 non-transportation 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/](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 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 is 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.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>National Environmental Satellite, Data and Information Service (NESDIS)</td>
<td>Climate and Weather</td>
<td>300 TB (digital)</td>
<td>80 TB/year</td>
<td>1-4</td>
<td>Some data is available NRT. NRT varies from minutes to weeks, depending on the data</td>
<td>Maps with configurable layers</td>
<td>No</td>
<td>Queries entire site content</td>
<td>Privately hosted data centers, including digital and non-digital media</td>
</tr>
<tr>
<td><a href="http://lwf.ncdc.noaa.gov/oa/climate/climatedata.html">http://lwf.ncdc.noaa.gov/oa/climate/climatedata.html</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarus System</td>
<td>Weather</td>
<td>400 GB</td>
<td>80 GB/yr</td>
<td>0-1</td>
<td>Hourly files</td>
<td>Map interface linking to data and quality flags; no visualization</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://www.its.dot.gov/clarus/">http://www.its.dot.gov/clarus/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earth Observing System Data and Information System (EOSDIS)</td>
<td>Climate</td>
<td>4.8 PB</td>
<td>600 TB/yr</td>
<td>0-4</td>
<td>Many data feeds available in NRT (minutes, hours)</td>
<td>Varies by research team; map interfaces and layer visualization</td>
<td>Projects, Standards and Working groups</td>
<td>Queries entire site content; separate facilities for searching archives</td>
<td>Privately hosted, distributed data archival and distribution facilities</td>
</tr>
<tr>
<td><a href="http://earthdata.nasa.gov">http://earthdata.nasa.gov</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data.gov</td>
<td>Public data across a wide variety of domains</td>
<td>50 GB</td>
<td>20 GB/yr</td>
<td>0-5</td>
<td>None</td>
<td>Depends on the data set, but much of it is viewable in a visualization tool</td>
<td>Yes, forums, blogs, various RSS feeds</td>
<td>Yes, across the entire site or subsections</td>
<td>Uses Socrata</td>
</tr>
<tr>
<td><a href="http://www.data.gov/">http://www.data.gov/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Size based on current storage of roughly 250,000 data sets, and each set averaging 200KB in size. Rate of increase based on establishment in 2009.</td>
</tr>
<tr>
<td>Simple Online Data</td>
<td>Social Sciences</td>
<td>&gt; 500 GB</td>
<td>*</td>
<td>0-4</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Queries</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>------</td>
<td>----------</td>
<td>-------------</td>
<td>--------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Archive for Population Studies (SodaPop)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>entire site content; separate facilities for searching archives</td>
</tr>
<tr>
<td><a href="http://sodapop.pop.psu.edu">http://sodapop.pop.psu.edu</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCLA Social Science Data Archive</td>
<td>Social Sciences</td>
<td>&gt; 500 GB</td>
<td>*</td>
<td>0-5</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Search for data only</td>
<td>Heavily hyperlinked between multiple universities.</td>
</tr>
<tr>
<td><a href="http://www.sscnet.ucla.edu/issr/da/">http://www.sscnet.ucla.edu/issr/da/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US Census Bureau</td>
<td>Social Sciences</td>
<td>&gt; 250 GB</td>
<td>*</td>
<td>4-5</td>
<td>None</td>
<td>Many data sets can be displayed on a map.</td>
<td>Feedback only</td>
<td>Very detailed and powerful search engines, global site search as well as detailed data search.</td>
<td>Endeca (Oracle) powered search</td>
</tr>
<tr>
<td><a href="http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml">http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bureau of Transportation Statistics</td>
<td>Transportation</td>
<td>*</td>
<td>*</td>
<td>3-5</td>
<td>None</td>
<td>Some data sets have pre-drawn visual summaries</td>
<td>None</td>
<td>Global site search</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.bts.gov/">http://www.bts.gov/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next Generation Simulation Community</td>
<td>Transportation</td>
<td>70 GB</td>
<td>*</td>
<td>0-5</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>User Info and Forums</td>
<td>Global site search</td>
</tr>
<tr>
<td><a href="http://ngsim-community.org/">http://ngsim-community.org/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------------------</td>
<td>------------</td>
<td>----------</td>
<td>-------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>---------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>PORTAL ITS data archive</td>
<td>Transportation</td>
<td>&gt; 60 GB</td>
<td>~10 GB/yr</td>
<td>0-4</td>
<td>Current traffic data is real time, all available through visualization. No external feeds.</td>
<td>Extensive map and performance measure-based plots</td>
<td>News, Facebook integration</td>
<td>Neither global nor data search. All data is accessed through a variety of intuitive interfaces</td>
<td>Notes: None</td>
</tr>
<tr>
<td><a href="http://portal.its.pdx.edu">http://portal.its.pdx.edu</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Transportation Library (NTL)</td>
<td>Transportation</td>
<td>*</td>
<td>*</td>
<td>5</td>
<td>None</td>
<td>None</td>
<td>Interaction with librarian only</td>
<td>Search documents</td>
<td>Notes: None</td>
</tr>
<tr>
<td><a href="http://ntl.bts.gov/">http://ntl.bts.gov/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caltrans Performance Measurement System (PeMS)</td>
<td>Transportation</td>
<td>11 TB</td>
<td>1 TB/yr</td>
<td>0-4</td>
<td>Real time data is included in the archive, but not distributed</td>
<td>Map-based</td>
<td>Map-based presentation of traditional traffic measures and incidents</td>
<td>Global site search</td>
<td>Notes: None</td>
</tr>
<tr>
<td><a href="http://pems.dot.ca.gov/">http://pems.dot.ca.gov/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connected Vehicle Research Data Exchange (RDE)</td>
<td>Transportation</td>
<td>2 TB</td>
<td>Projected 500 GB/yr</td>
<td>0-4</td>
<td>As available from external providers, will distribute real-time feeds</td>
<td>None</td>
<td>Forums, feedback to operators</td>
<td>Global site, real time and archive data by metadata</td>
<td>Planning to use Alfresco or Nuxeo technologies. Prototype uses Drupal On-going project</td>
</tr>
<tr>
<td><a href="https://www.its-rde.net/home">https://www.its-rde.net/home</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Undetermined
3.4 COMMERCIALY 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 prior to starting the development phase (Phase III).

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

- Would provide the core functionality of the Archive;
- 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 3-5: Data Archival Technologies

<table>
<thead>
<tr>
<th>Tool</th>
<th>Application</th>
<th>Appropriateness (scale of 1-10, 10 being highest)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordPress</td>
<td>Content Management</td>
<td>10</td>
<td>WordPress provides a flexible environment for the developers to easily modify the UI.</td>
</tr>
<tr>
<td>Alfresco</td>
<td>Enterprise Content Management (ECM)</td>
<td>8</td>
<td>Alfresco and Nuxeo are premier and affordable. Drupal is capable but smaller scale. A detailed analysis of these tools should be performed to select one.</td>
</tr>
<tr>
<td>Nuxeo</td>
<td>Enterprise Content Management (ECM)</td>
<td>8</td>
<td>Provides the total package, but not really focused on archival of large data sets.</td>
</tr>
<tr>
<td>Socrata</td>
<td>Service</td>
<td>6</td>
<td>Document management, not content management. Could be used but would require additional work.</td>
</tr>
<tr>
<td>OpenKM</td>
<td>Document Management</td>
<td>5</td>
<td>Would need additional work to manage metadata.</td>
</tr>
<tr>
<td>Drupal</td>
<td>Content Management</td>
<td>7</td>
<td>Drupal and CKAN would have to be used together.</td>
</tr>
<tr>
<td>CKAN</td>
<td>Portal</td>
<td>7</td>
<td>Nesstar is a system for data publishing and visualization. Nesstar does not have built-in collaboration. It is not clear how to integrate third-party tools.</td>
</tr>
<tr>
<td>Cyan.In</td>
<td>Content Management</td>
<td>5</td>
<td>Nesstar is a system for data publishing and visualization. Nesstar does not have built-in collaboration. It is not clear how to integrate third-party tools.</td>
</tr>
<tr>
<td>S4PA</td>
<td>File Management</td>
<td>3</td>
<td>Nesstar is a system for data publishing and visualization. Nesstar does not have built-in collaboration. It is not clear how to integrate third-party tools.</td>
</tr>
<tr>
<td>OpenDocMan</td>
<td>Document Management</td>
<td>2</td>
<td>Nesstar is a system for data publishing and visualization. Nesstar does not have built-in collaboration. It is not clear how to integrate third-party tools.</td>
</tr>
<tr>
<td>KnowledgeTree</td>
<td>Document Management</td>
<td>1</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Fedora-Commons</td>
<td>Data Repository</td>
<td>1</td>
<td>Insufficient</td>
</tr>
<tr>
<td>EPrints</td>
<td>Electronic Publishing</td>
<td>1</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Nesstar</td>
<td>Data Cataloging System</td>
<td>6</td>
<td>Insufficient</td>
</tr>
<tr>
<td>DSpace</td>
<td>Data Archive Management</td>
<td>9</td>
<td>Insufficient</td>
</tr>
<tr>
<td>Tool</td>
<td>Application</td>
<td>Appropriateness (scale of 1-10, 10 being highest)</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>--------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Alfresco (<a href="http://www.alfresco.com">http://www.alfresco.com</a>)</td>
<td>Custom metadata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuxeo (<a href="http://www.nuxeo.com">http://www.nuxeo.com</a>)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socrata (<a href="http://www.socrata.com">http://www.socrata.com</a>)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.4.1 Overview of Applicable Supporting Technologies

The team reviewed the technologies below. They provide some components of content management, document management, and web portal functionality.

3.4.1.1 WordPress (http://www.wordpress.org)

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 makes UI building tasks very easy. For more information on WordPress see Section 3.5.4 and Section 7.2.

3.4.1.2 Alfresco (http://www.alfresco.com)

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


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.

The team had problem with customizing the front-end.

3.4.1.3 Nuxeo (http://www.nuxeo.com)

Nuxeo is a free ECM system written in Python that includes a host of related functionality, including document management, social collaboration, case management and digital asset management. Nuxeo is similar in scope, scale and cost to Alfresco, and was considered a viable alternative. Like Alfresco, front-end customization was the biggest challenge.

3.4.1.4 Socrata (http://www.socrata.com)

Socrata is a cloud-based data publication and collaboration service. Socrata is not a component that one uses 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 would require some negotiation with the service’s sales staff.
3.4.1.5  **OpenKM (www.openkm.com)**

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

3.4.1.6  **Drupal (http://www.drupal.org)**

Drupal is an open source content management system. It provides database cataloging and storing data sets, web front-end and API. It is distributed under GNU GPL v.3. It is smaller and less capable 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).

3.4.1.7  **CKAN (http://www.ckan.org)**

CKAN is an open source data portal system. It provides database cataloging and storing data sets, web front-end and 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.

3.4.1.8  **Cyn.in (http://www.cynapse.com)**

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.

3.4.1.9  **S4PA (http://disc.sci.gsfc.nasa.gov/additional/techlab/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 ingest, and a data distribution module that provides data files to users. Data is managed in a tightly organized UNIX file structure. Data storage and distribution is 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.

3.4.1.10  **OpenDocMan (http://www.opendocman.com)**

OpenDocMan is a free, open source web-based PHP DMS distributed under GPL license. 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 was inadequate for L13A.

3.4.1.11 KnowledgeTree (http://www.knowledgetree.com)

KnowledgeTree provides a cloud-based service for document management and workflow. Its Representational State Transfer (REST) and Simple Object Access Protocol (SOAP) Application Programming Interfaces (APIs) allow integration into third-party web sites. This solution did not seem adequate for L13A as it was not highly customizable, and does not handle user metadata.

3.4.1.12 Fedora-Commons (http://www.fedora-commons.org)

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 quite an extensive software development on top of it to be able to run even 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.

3.4.1.13 EPrints (http://www.eprints.org)

EPrints is open source software under GPLv.3 and LGPLv.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 clear instructions about integration with third party tools.

3.4.1.14 Nesstar (http://www.nesstar.com)

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 significant limitations in UI customization. All Nesstar catalogues 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, allowing search, browsing and visualization. Nesstar does not have built-in collaboration. It was not clear how to integrate third-party tools and thus data upload capability for collaborating users.

3.4.1.15 DSpace (http://www.dspace.org)

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. Alfresco has more content management functionality, allowing more flexible processing of the uploaded content, which is important for special treatment of data sets that are to be visualized. DSpace focuses on the approval of content rather than wider workflow customization.
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 in which this archive system would be built upon. 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 proposes a web-based solution, which utilized 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 recommends COTS over open-source and in-house development.

The Iteris team generally agreed with the analysis performed by the L13 research team except for their de-emphasis of high-level data visualization and the 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 as well as a standardized way for visualizing and aggregating objects.

Our 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 Services

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.
- Incorporating training and outreach in the project. The key to successful outreach will be to show that ADUS systems help to perform common tasks faster, easier and more accurately.

3.5.2.2 Summary of ADUS Systems

Other than the Washington 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 had 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 their ADUS. This process helped them 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

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

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 is similarly obfuscated to its 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, some visualization is desirable to include.
4. Provision of system performance characteristics, so that contributors can see how their data is being used and thus quantify the benefits of sharing their data.
5. Collaboration tools with feedback mechanisms, whereby 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. Similar to the previous point, but including a notion of quality will 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

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

1. WordPress
2. Socrata
3. Alfresco

The team ruled out the possibility of using Socrata due its high annual service fee. A thorough comparison between WordPress and Alfresco systems was performed through building small prototypes. The team tested functionalities and features provided by each platform to check which fit the Archive needs well. One feature that the development team looked into was the flexibility of each platform to front-end and back-end customization. At the end, the team decided to utilize WordPress. The following is the summary of the comparison that led to selection of WordPress as the core CMS:

**Alfresco Limitations:**

- **Complexity of content management standards** – Alfresco and some other ECM systems are mostly built around two content management standards: Content Management Interoperability Service (CMIS) and Java Content Repository (JCR). Both of these standards are complex. These standards are relatively low-level. Writing an application based on one of these standards would take a significant amount of time.

- **Inflexibility:**
  - **Customization** – The combination of the above factor with the need to customize the UI (e.g. visualization and ingestion) and back-end functionality (e.g. support for very large data sets) means that using Alfresco’s web based UI or any other ECM UI would require a great deal of customization. Developing an application on top of the existing UI infrastructure (e.g., Alfresco Share) was possible; however, the complexity of the software made the customization a big challenge.
  - **XML content modeling** – The available format limited different types of content model relationships. For instance, creating hierarchical relationships within constraints was not possible.

- **Steep learning curve** – Deep knowledge of how the ECM system and its UI operate was required to customize the software for the Archive. For instance, Alfresco documentation was poorly written, which made the learning of this complex platform difficult.

- **Direct Database access issue** – This is a problem with some of the ECM systems, specifically Alfresco, in which the developers cannot access the database directly.

- **High risk** – The above factors increased the level of risk in the project, which adversely affected the chance of success.

- **Cost** – ECM software is normally open-source. However, for some systems, such as Alfresco, the users need to pay a fee for the enterprise version to benefit from customer/software support services. The enterprise versions are more reliable and error-prone. As an example, the subscription fee for the Enterprise version of Alfresco – the version that includes the support package – was $50k+/year.

**WordPress Benefits:**

- **Simplicity** – UI implementation and content creation is simple in WordPress.
- **Shallow learning curve** – Lots of easy-to-read formal and user-based informal documentation was available. The operation of the software was transparent and easy to figure out.
- **Extensibility** – Thousands of plugins and themes were available to create a variety of functionality for the Archive.
• **Market-leading platform** - Tens of millions of sites are using WordPress. The community also includes thousands of active developers.

• **Flexibility:**
  - **Customization** - Customizing the UI in WordPress was extremely simple. WordPress uses common, simple programming languages, including PHP and MySQL. Tomcat and stand-alone Java applications can be easily integrated with WordPress. Flexibility to integrate Java-based applications was particularly helpful for implementing the visualization and ingestion tools.
  - **Content modeling** – Custom content modeling was widely documented and could be implemented easily. WordPress provides base and Custom Taxonomy support along with standard and custom metadata field support.

• **Low risk** - Due to simplicity and flexibility of the platform, the risk of failure was very low.
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, which were captured as user stories, were developed based on two workshops conducted as part of the L13A project, aiming to collect SME feedback on essential features of the Archive. This chapter begins with the results of the feasibility study conducted under L13. It should be noted, while the feasibility report did include a set of requirements, these 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 (INCOSE 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 this 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 will form the basis for system verification. For L13A, the needs were defined in a series of user stories. A 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 System Archive 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 is submitted by users and held separately until an administrator can examine and validate its format and contents, at which time that data may be moved into the data Archive where it is available to the web server and thus other users. Figure 4-1 illustrates this very high level view of the System Archive.
4.1.1 Types of Artifacts

The reference to “artifact” in this document covers any stored digital objects (i.e., document, data set, computer code, etc.) in the Archive system that will be viewed and used by the researchers. This document will refer to two types of artifacts: Primary SHRP 2 Artifacts and 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 the 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

There are four major user roles envisioned for artifact ingestion: administrator, registered user, guest and Principal Investigator (PI). In order 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.
Table 4-1: User roles

<table>
<thead>
<tr>
<th></th>
<th>Guest</th>
<th>Registered User</th>
<th>Principal Investigators</th>
<th>Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download artifacts</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Search and visualize artifacts</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Participate in the discussion</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Upload artifacts from the</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>webpage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delete artifacts</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Permanently delete artifacts</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>from Archive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 & 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 IT resources to address areas of concern, providing technical support to users on an as-needed basis, and creating regular preservation planning reports (~ every 3-5 years).

- **Preservation of PII:** Administrators have to make sure no PII data would be shared to the public via the Archive.

4.2.2 Registered User

Registered Users create a user account, which contains a minimum of: username, password and email address. The email address of the user is verified during the registration process. The user may choose to store additional information into 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 only edit metadata on artifacts that they have uploaded.

4.2.3 Principal Investigator

Principal Investigator users are Registered Users who will be able to upload Primary SHRP 2 Artifacts. A PI takes on the role only for those artifacts that he/she uploads to the
Archive system. There can only be one PI per artifact in the Archive. The PI is responsible for pre-processing 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 both 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 pre-processing and metadata information collection tasks on behalf of the PI.

4.2.4 Guest

In order to create the most open system possible, 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 in order to meet operational goals.
- User needs describe what the system needs to do in order to satisfy users.

The needs were developed based on the comments gleaned from the June 2012 stakeholder workshop and project revised work plan. They were written in a language compatible with IEEE-1362. While this document does not follow that standard completely (it is not a fully fledged 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, etc.). 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 is no longer needed. This provides the essence of a long term data archive.

3) **Distribute data:** The system needs to provide the ability for users to download data that 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 datasets, and in turn enables multiple dimensionality of association between data sets, access rights to data sets, and also enables search functionality.

5) **Search for Data:** The system needs to provide a mechanism for users to search through available data. This mechanism will allow the user to input criteria, and the system will then provide a list of data that is relevant to that 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 web site. This allows users to browse available data and discussions. Essentially, this means to develop 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 datasets and files. This may be restricted based on user permissions. This allows enhancement to and correction of metadata. This also provides for the possibility of expanded definitions of metadata, which is relevant for those datasets 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 only be available 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 manager 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 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 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 datasets 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 data it stores. This should be available from search or other navigation means, and also should present the user with images of geographically-referenced data.

### 4.3.2 Assumptions and Constraints

**A1) Assumption:** Automated ingest will be provided by an API.

**C1) Ingestion Constraints:** User-data ingest must support HTTP.

**C3) Metadata Standards:** Metadata must be developed according to a documented standard.

**C4) System Backups:** System backup formats must follow a documented industry standard.

**C5) Administration Security:** Administration functionality must be provided over a secure connection except for those instances when such a connection is unavailable.

**C6) Interface Documentation:** All interfaces and backdoors must be documented.

**C7) Parallelism:** The system must allow multiple files to be accessed simultaneously.

**C8) 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, etc.). Artifacts submitted under each category need to meet certain requirements. Therefore, the producer of an artifact has to pre-process it before submitting into the system.

4.4.2 Search and Download

The Archive provides faceted and text search tools to help the users look for artifacts. The text search feature enables the users to conduct content search within artifacts and metadata. In addition, the faceted search tool allows the users to explore the Archive’s repository. The 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, the users can download the desired objects along with their metadata and metadata documents (if available). The 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 the users. The users will benefit from two types of visualization tools:

Map Visualization: The system has the capability to map the geo-location information provided in a data set. The geo-filter and map view capabilities can help users explore sensor locations (Figure 4-4).

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) 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

The 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, the users have to be registered.

4.4.5 Administration and User Profile Interface

The Archive administrator is capable of granting access to the 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, the 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. In order to understand the perspective of the system from a user, users are defined as the following:

**U1) Archive User:** The Archive User wants services from the Archive system. He/she may provide data to be shared, may download data others have provided, and may comment on his own or others projects and data sets.

**U2) Archive Administrator (administrator access control level):** The Archive Administrator operates the Archive system. He 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 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 used with Agile software development methodologies as the basis for defining the functions a business system must provide, and to facilitate requirements management. It 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, and to 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 SHRP 2 Reliability focused 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 thorough help document to guide the Users.

#### 4.5.1.2 Ingestion

**US2.** 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 "Ingestion Guideline." 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 “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 Users and logs both successful and unsuccessful submissions.
US2.6. The Archive System notifies the Administrator once an artifact is submitted.

US3. **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. Upon successful submission of the artifact and approval of 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. **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/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. **Archive User registers with the Archive System**
US5.1. The Archive User acknowledges a terms of use agreement, provides a username, 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. **Archive Administrator delete/updates artifacts**

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

US7. **Archive PI cannot permanently delete SHRP 2 Primary artifacts**
US7.1. Under no circumstances can an Archive User permanently delete SHRP 2 Primary artifacts. The PI can only request for permanent deletion.

US8. **Archive PI updates only artifacts that he/she submitted**

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

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

US9. **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. **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. **Archive System replicates the artifacts**

US11.1. The Archive Manager 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. **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 backup based on various parameters (e.g., timescale, names, area, etc.). The Archive System must provide the ability to save system configuration and system metadata in order to reconstitute system recovery and reconstruction.

US13. **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. The Users can sign up to the system by providing their name and a valid email address.

US15. **Archive Administrator manages user accounts**

US15.1. The Archive Administrator is able to create, modify and delete Archive User accounts.

US16. **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. **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. The abovementioned activities include:

- Starting and stopping Archive System services.
- Updating and patching application and system software.
- Content moderation.
- Customer service.
- Access control.
- Verification.

US18. **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. In order 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. **Archive System can be easily expandable without service interruption**

US19.1. The Archive System’s storage must be expandable while maintaining services to Archive Users and the Archive Administrator.

US20. **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. **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.
- Location associated with artifacts within their metadata.

4.5.1.5 **Download**

US22. **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. **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. **Archive User comments on project**

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. The registered Archive User can rate a project or an artifact and provide feedback.

US25. **Archive Administrator deletes comment on a project**

US25.1. An Archive Administrator has the capability to delete comments on a project.

US26. **Archive User comments on an artifact within the 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. The 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. **Archive Administrator deletes comment on an artifact**

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

**4.5.1.7 Visualization**

US28. **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 utilized by a third party front-end visualization application.

US29. **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.

71
**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 geo-information is provided on the metadata, the Users can search for artifacts on the map.

US29.11. If geo-data 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:

Data sets

- Examples include traffic engineering data such as travel time, flow, and occupancy. An example data set is shown in Figure 5-1;
- Must be in .csv format;
- Can be used for visualization;
- Can be queried;
- 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 was collected. Otherwise, the data is less usable;
- Require special and general metadata;
- Require a data dictionary.

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

Figure 5-1: Data Set Example

Non-Data Sets

- Examples include documents, computer codes, simulation models, spreadsheets, presentations, etc. An example of a non-data set is shown in Figure 5-2;
- Documents must be in .pdf format;
- Require only general metadata;
- Not for visualization;
- Excel spreadsheets are in this category;
- Data dictionaries required 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.

**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.
**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 requirements. The general requirements are common for all types of artifacts. For instance, file size should not exceed 1GB. 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.

**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 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 the User Guidebook.

![Figure 5-4 Upload Wizard](image-url)
Step 4. Backend Processing

Upon completion of the upload task, the Administrator and User receive an email 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 post-upload 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 content. The post-upload workflow process consists of the following steps:

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.

**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;
- Acknowledgement to the people who contributed to creating the data set, such as the road authority who owns the vehicle detector or individuals/organizations who helped to 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 was collected (data sets only) and much more.

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

- Overall site;
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, artifact and user – 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 team determined the controlled vocabulary, cardinality (1:1, 1:many, many:1, or many:many), generator (system vs. 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, whereas 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), whereas 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 of state names to select from)
and encoding schemes (e.g. YYYY-MM-DD format) prevent unintended metadata entry errors and help to 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 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 final changes were made to the site.

5.4.2 Site Metadata

Table 5-1 provides a brief summary of the Site Metadata elements.
### Table 5-1 - Site Metadata

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Type</th>
<th>Mandatory?</th>
<th>Editable?</th>
<th>Multiple?</th>
<th>Generator</th>
<th>Format</th>
<th>Controlled Vocabulary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>none</td>
</tr>
<tr>
<td>Description</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>none</td>
</tr>
<tr>
<td>URL</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Administrator</td>
<td>URL</td>
<td>none</td>
</tr>
<tr>
<td>Child Focus Areas</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Administrator</td>
<td>focus area</td>
<td></td>
</tr>
<tr>
<td>Admin(s)</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Creator</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Registered User PI</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Registered User</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Guest User</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
</tbody>
</table>

### 5.4.3 Focus Area Metadata

Table 5-2 provides a brief summary of the Focus Area Metadata elements.

### Table 5-2 - Focus Area Metadata

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Type</th>
<th>Mandatory?</th>
<th>Editable?</th>
<th>Multiple?</th>
<th>Generator</th>
<th>Format</th>
<th>Controlled Vocabulary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Description</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Date Created</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>date-time</td>
<td>Yes</td>
</tr>
<tr>
<td>Date Modified</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>date-time</td>
<td>Yes</td>
</tr>
<tr>
<td>Child Project(s)</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>project</td>
<td></td>
</tr>
<tr>
<td>Admin(s)</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Creator</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Registered User PI</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Registered User</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
</tbody>
</table>
5.4.4 Project Metadata

Table 5-3 provides a brief summary of the Project Metadata elements.

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Type</th>
<th>Mandatory?</th>
<th>Editable?</th>
<th>Multiple?</th>
<th>Generator</th>
<th>Format</th>
<th>Controlled Vocabulary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Title</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Background</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Objectives</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Research Agency</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Primary Investigator</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Keywords/Tags</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Date Created</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>date-time</td>
<td>Yes</td>
</tr>
<tr>
<td>Date Modified</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>date-time</td>
<td>Yes</td>
</tr>
<tr>
<td>Parent Site</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>site</td>
<td></td>
</tr>
<tr>
<td>Parent Focus Area</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>focus area</td>
<td></td>
</tr>
<tr>
<td>Child Artifact(s)</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Administrator</td>
<td>artifact</td>
<td></td>
</tr>
<tr>
<td>Admin(s)</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Creator</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Registered User</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Administrator</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Registered User</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>Viewer(s) - Guest User</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>user</td>
<td></td>
</tr>
</tbody>
</table>
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 Username, Password, Email, and Display Name. All other elements of a user profile are optional or set by the site administrator (e.g. Roles). Table 5-4 provides a brief summary of the User Metadata elements.

Table 5-4 - User Metadata

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Type</th>
<th>Mandatory?</th>
<th>Editable?</th>
<th>Multiple?</th>
<th>Generator</th>
<th>Format</th>
<th>Controlled Vocabulary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Username</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Registered User</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Password</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Registered User</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Display Name</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Registered User</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Email address</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Registered User</td>
<td>email</td>
<td>Yes</td>
</tr>
<tr>
<td>First Name</td>
<td>Attribute</td>
<td>No</td>
<td>Yes**</td>
<td>No</td>
<td>Registered User</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Last Name</td>
<td>Attribute</td>
<td>No</td>
<td>Yes**</td>
<td>No</td>
<td>Registered User</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Bio Info</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Registered User</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Site Role</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Administrator</td>
<td>role</td>
<td>Yes</td>
</tr>
<tr>
<td>Submitted Artifact(s)</td>
<td>Relationship</td>
<td>Yes*</td>
<td>No</td>
<td>Yes</td>
<td>System-Generated</td>
<td>artifact</td>
<td>No</td>
</tr>
<tr>
<td>User Comment(s)</td>
<td>Relationship</td>
<td>Yes*</td>
<td>Yes</td>
<td>Yes</td>
<td>System-Generated</td>
<td>text</td>
<td>No</td>
</tr>
</tbody>
</table>

*If Applicable
** Editable only by Administrator
5.4.6 Artifact Metadata

Artifact metadata is fundamental to the Archive’s search function and is the supporting information users rely upon 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 tradeoff 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 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, is dependent upon 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: 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 and Table 5-6 summarizes the additional metadata requirements for data sets.
Table 5-5 – General Metadata Element Set

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Type</th>
<th>Mandatory?</th>
<th>Editable?</th>
<th>Multiple?</th>
<th>Generator</th>
<th>Format</th>
<th>Controlled Vocabulary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Title</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Abstract/Description</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>ID</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>System-Generated</td>
<td>URL</td>
<td>No</td>
</tr>
<tr>
<td>Primary SHRP 2 Artifact?</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Focus Area</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>focus area</td>
<td>Yes</td>
</tr>
<tr>
<td>Project</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>project</td>
<td>Yes</td>
</tr>
<tr>
<td>Artifact Type</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Artifact Relation</td>
<td>Relationship</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Location(s)</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes (up to 10)</td>
<td>Creator</td>
<td>City, State</td>
<td>Yes</td>
</tr>
<tr>
<td>Latitude, Longitude</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes (up to 10)</td>
<td>System-Generated</td>
<td>Decimal Degrees</td>
<td>Yes</td>
</tr>
<tr>
<td>Year Range</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>YYYY - YYYY</td>
<td>Yes</td>
</tr>
<tr>
<td>Date Uploaded</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>date-time</td>
<td>Yes</td>
</tr>
<tr>
<td>Date Last Modified</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>date-time</td>
<td>Yes</td>
</tr>
<tr>
<td>File Format</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>File Size</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>number</td>
<td>No</td>
</tr>
<tr>
<td>Number of Downloads</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>number</td>
<td>No</td>
</tr>
<tr>
<td>Related Comments</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>comment</td>
<td>Yes</td>
</tr>
<tr>
<td>Review Status</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>System-Generated</td>
<td>number</td>
<td>Yes</td>
</tr>
<tr>
<td>Workflow Status</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>number</td>
<td>Yes</td>
</tr>
<tr>
<td>Validation Status</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>number</td>
<td>Yes</td>
</tr>
<tr>
<td>Indexing Status</td>
<td>Attribute</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>number</td>
<td>Yes</td>
</tr>
<tr>
<td>Admin(s)</td>
<td>Relationship</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>user</td>
<td>Yes</td>
</tr>
<tr>
<td>Creator</td>
<td>Relationship</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>System-Generated</td>
<td>user</td>
<td>Yes</td>
</tr>
<tr>
<td>Viewer(s) - Registered User PI</td>
<td>Relationship</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>user</td>
<td>Yes</td>
</tr>
<tr>
<td>Element Name</td>
<td>Type</td>
<td>Mandatory?</td>
<td>Editable?</td>
<td>Multiple?</td>
<td>Generator</td>
<td>Format</td>
<td>Controlled Vocabulary?</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>--------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Column Name</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Column Index</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>number</td>
<td>Yes</td>
</tr>
<tr>
<td>Column Type</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Column Label</td>
<td>Attribute</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Latitude 1 Column</td>
<td>Attribute</td>
<td>Yes*</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>column</td>
<td>Yes</td>
</tr>
<tr>
<td>Longitude 1 Column</td>
<td>Attribute</td>
<td>Yes*</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>column</td>
<td>Yes</td>
</tr>
<tr>
<td>Latitude 2 Column</td>
<td>Attribute</td>
<td>Yes*</td>
<td>No</td>
<td>No</td>
<td>Creator</td>
<td>column</td>
<td>Yes</td>
</tr>
<tr>
<td>Longitude 2 Column</td>
<td>Attribute</td>
<td>Yes*</td>
<td>No</td>
<td>No</td>
<td>Creator</td>
<td>column</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Set Dictionary</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Creator</td>
<td>URL</td>
<td>No</td>
</tr>
<tr>
<td>Data Source(s)</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Data Type(s)</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Corridor(s)</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>text</td>
<td>No</td>
</tr>
<tr>
<td>Collection Technology(ies)</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Collection Frequency</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Days of the Week</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
<tr>
<td>Holidays Included?</td>
<td>Attribute</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Creator</td>
<td>text</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*If Applicable

Table 5-6 – Data Set Metadata Element Set

This data is required in addition to the metadata requirements from Table 5-5.
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 B (i.e., there is a link to artifact B in artifact A’s page), artifact B also contains a related artifact 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 being 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 Why is Preparation Needed for Data Sets?

A standardized format for data sets increases the usability of the user’s research data by future users, and maximizes the distribution and impact of the user’s 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 can 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 only available on data sets.

5.6.1.2 Data Set Preparation Checklist

Table 5-7 shows the checklist to prepare data sets.

<table>
<thead>
<tr>
<th>Data Set Size</th>
<th>The size of the data set is: (please tick)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐ Less than 1GB – proceed to Step 2</td>
</tr>
<tr>
<td></td>
<td>☐ Greater than 1GB – contact the Administrator</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set Format</th>
<th>Check that the data set meets the following conditions:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐ CSV format – see extra information</td>
</tr>
<tr>
<td></td>
<td>☐ Has at least 1 column of data</td>
</tr>
<tr>
<td></td>
<td>☐ Has less than 60 columns of data</td>
</tr>
<tr>
<td></td>
<td>☐ Each row of data has the same number of columns (i.e. same number of commas)</td>
</tr>
<tr>
<td></td>
<td>☐ The first row contains header names</td>
</tr>
<tr>
<td></td>
<td>☐ Each column has at least one non-null field</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column Headings</th>
<th>Check that each heading name:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐ Is between 1 and 80 characters long</td>
</tr>
<tr>
<td></td>
<td>☐ Is unique</td>
</tr>
</tbody>
</table>
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

Data Set Size Restrictions

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

Data Set Format

Data set file 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 to a .csv file (Figure 5-7). Larger files may require manipulation using a programming language.

Example spreadsheet file

<table>
<thead>
<tr>
<th>Date_time</th>
<th>Volume</th>
<th>Travel time</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/25/2011 17:00</td>
<td>7</td>
<td>248.6</td>
<td>39.311977</td>
<td>-120.495774</td>
</tr>
<tr>
<td>4/25/2011 17:01</td>
<td>57</td>
<td>37.5</td>
<td>39.311977</td>
<td>-120.495774</td>
</tr>
<tr>
<td>4/25/2011 17:02</td>
<td>5</td>
<td>204</td>
<td>39.311977</td>
<td>-120.392934</td>
</tr>
<tr>
<td>4/25/2011 17:03</td>
<td>71</td>
<td>4.1</td>
<td>39.329142</td>
<td>-120.392934</td>
</tr>
<tr>
<td>4/25/2011 17:04</td>
<td>9</td>
<td>261.4</td>
<td>39.329142</td>
<td>-120.303034</td>
</tr>
</tbody>
</table>

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,-120.392934
4/25/2011 17:00,71,4.1,39.329142,-120.392934
4/25/2011 17:00,9,261.4,39.329142,-120.303034

Figure 5-7: Examples of File Formats

Column Headings

Headings must be between 1 and 80 characters long, unique, and only contain permitted characters including a-z, A-Z, 0-9, dashes, spaces and underscores. Example column headings may include:
• Time;
• Date;
• Volume;
• Speed;
• Travel Time;
• Station ID;
• Latitude;
• 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.

**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;
• Latitude and Longitude – see below.

**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.

---

**Date only columns**
- yyyy/MM/dd
- yyyy-MM-dd
- MM/dd/yyyy
- MM-dd-yyyy
- yyyyMMdd

**Time only columns**
- HH:mm:ss
- HH:mm

**Date & time columns**
- MM/dd/yy HH:mm
- MM-dd-yy HH:mm
- MM/dd/yyyy HH:mm:ss
- MM-dd-yyyy HH:mm
- MM/dd/yyyy HH:mm:ss
- yyyy-MM-dd HH:mm:ss
- yyyy/MM/dd HH:mm:ss

---

*Figure 5-8: Date and Time Formats*

**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, where north is positive and south is negative. The longitude values range from -180 to 180 where 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, 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:

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.

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

5.6.2.2 File Format

Reports and documents should be in pdf format. Many other file types are accepted and the complete list is shown below (in alphabetical order).

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 sub-consultants, is providing consistency in the experience of the Archive’s users. For that purpose, a series of documents has been drafted by
Iteris and SHRP 2 to help PIs complete their task and to foster an organized and internally consistent Archive.

- **Data Dictionary Template.** Iteris produced the Data Dictionary Template (Appendix A) to help PIs get started on documenting their data set artifacts. This template helps to 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 Iteris 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 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: \textit{SHRP 2_SoftwareName_LXXA_Contractor_Vn.n_dd/mm/yyyy}. In addition, a label that is visible and easily read upon 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 (for example, “Loop Count Station #”), is there a Related Artifact that provides the geographical location...
of stations (i.e., a stations configuration file) or latitude / longitude coordinates in an adjacent column?

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 sub tab 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 sub tab 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, and then downloading and open the document or presentation from the Archive, does the file open OK?

5.8.2.2 Spreadsheets and Computer Code

In terms of complexity in uploading and processing, spreadsheets and computer code are between documents/presentations and data sets. Although considered non-data sets, spreadsheets and computer code typically benefit from documentation in the form of a metadata, user guide, or 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 to upload data dictionaries for spreadsheets and computer code is different than that for data sets. While the data set specifically requires a data dictionary upon upload, 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 also be provided in a similar manner.
CHAPTER 6: WORKING WITH THE ARCHIVE – USER GUIDE

This chapter will review the Archive functionalities and explain how a user can utilize the Archive. In another word, this chapter serves as a user guide, describing system features and providing a step-by-step guide to use the system. The web address to access the Archive is http://www.shrp2archive.org. The latest version of the user guide can be found 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 three to eight minutes.

6.1.1 How to Create an Account?

An account can be created by finding the Log In or Register link at the top left of the Archive (Figure 6-1) and completing the following steps:

1. The user clicks on 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 email will be sent to the user to validate his/her identity. This email contains a temporary password and a link to the log in page.
5. The user enters his/her email address and temporary password into the Log In screen. The user can change the temporary password if he/she desires.

Figure 6-1: Log In or Register
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/her details on the My Profile tab.
3. The user clicks Update Profile to confirm the changes. If the user changes an email address, the user will be automatically logged out. To resume activities, the user needs to log in using the updated email address.

![Figure 6-2: My Profile](image)

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 log in page and then follows the Lost Password link below the log in boxes (Figure 6-3).
3. The user enters the email address he/she registered and a new password will be sent to the registered email 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 the Archive.

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.
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.

![Keyword Search](image)

*Figure 6-5: Keyword Search*

6.3.2 Search Archive

The **Search Archive** feature at the top right of the SHRP 2 Reliability Archive page is similar to an advanced search feature in a library 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. 30s to daily);
- Data sets that were collected on a specific day of the week;
- Data sets that include or exclude holidays;
- Specific types of non-data sets including analysis tools, computer code, data dictionaries, final reports, guide books, presentations, RFPs or surveys.
After the user chooses conditions to filter the search by, he/she can display the remaining artifacts geographically on a map or in a list. See below.

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 on 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 to the data set, the user clicks on it on the map.
6.3.4 Search through a List of Artifact Titles

Users can utilize 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 on 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 the SHRP 2 Reliability Archive, the user clicks on 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 on List of Reliability Focus Area Related Projects to view the list of projects.
6.3.6 View Latest Artifacts
The list of latest artifacts is shown on the SHRP 2 Reliability Archive homepage (Figure 6-10).

6.3.7 View Your Own Artifacts
If the user is a Principal Investigator or Creator on a project, he/she can use this option to view all the artifacts that he/she has uploaded (Figure 6-11).

1. At the top of the SHRP 2 Reliability Archive web page, the user clicks My profile to view his/her profile information.
2. The user chooses My Artifacts to view a list of the artifacts that he/she has uploaded.
6.4 WORKING WITH ARTIFACTS

This section provides information about working with artifacts. Once the user has found the artifact he/she is interested in, he/she 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 at the top right of the metadata tab by clicking on Download File. To download the data dictionary, the user can click on 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.

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. When the user opens this tab, he/she will see a central panel and a left and right side panel (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, create a filter that finds all the data points with speeds between 40-50 miles per hour. 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 side panels can be collapsed or expanded by clicking on the vertical gray bars.

Figure 6-12: General Navigation of Data Tab

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 then make selections to show or hide columns (Figure 6-14).

Figure 6-13: Grid View

Figure 6-14: Show or Hide Columns of Data
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. First, the user selects the type of graph and 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.

![Figure 6-15: Graph View](image)

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.

6.4.2.3.1 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 chosen to plot, 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 dropdown 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.

![Figure 6-16: Displaying Data Series](image)

6.4.2.3.2 Types of Graphs

The user can choose from scatter, line, spline, area spline, column, bar, and pie graphs (Figure 6-17).

![Figure 6-17: Types of Graphs](image)
6.4.2.3 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 on the icon located at the top right of the graph (Figure 6-18).

![Printing and Downloading Options](image)

Constraints in graphing are:
- 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.

6.4.2.4 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.
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).

**6.4.2.5 Filter**

The SHRP 2 Reliability 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 AM Peak between 7-9am;
- 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 miles per hour);
- Latitude and longitude coordinates to view data at one location only;
- Weather information column, to view data gathered under snowy conditions only.

6.4.2.5.1 Selecting a Filter

There are three types of filters available:

- Slider filter.
- Number filter.
- 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 he/she wants to apply.
2. The user clicks Add.
3. The user chooses filter conditions and clicks Update View to update the visualization in the central panel.

Figure 6-21: Defining Filter Criteria

6.4.2.5.2 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.
6.4.2.5.3 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 comma 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 & 828.

6.4.2.5.4 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 US states of DE, CO, DC, AL, ID, MD or FL. The text criteria can be separated by a comma.

6.4.2.5.5 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 filter – ‘AND’ vs ‘OR.’ ‘OR’ logic is applied within a filter, while ‘AND’ logic operates between 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 AM peak period AND
• Volumes between 100 – 500 OR 2000 – 2400.

Figure 6-25: Example of Search with AND & OR Operators

Other details:

• 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 to all artifacts (i.e. non-data sets and data sets). For data sets, this feature will download the .csv file.

6.5.2 Download Filtered Artifact

If the user has just filtered a data set, he/she 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 the user an opportunity to express his/her opinion or ask questions about an artifact. The comment feature is also capable of blocking inappropriate words and phrases automatically. However, it is not a fully proofed 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 out of a possible 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](image)

6.7 UPLOADING ARTIFACTS

6.7.1 Who Can Upload Artifacts?

The SHRP 2 Reliability 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 currently 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/her name, email address and the name of the SHRP 2 Project.

6.7.2 Artifact Upload Wizard

Prior to uploading an artifact, the file needs to be prepared into the required format (See Section 5.6). To upload artifacts, the SHRP 2 system must recognize the account as SHRP 2 PI and the uploading user must be logged in.

To start the upload process the user should navigate to the My Profile page, select the My Artifacts tab and then click on + Upload a New Artifact (Figure 6-27).
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

If the user uploads a data set (i.e., a .csv file), then he/she 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).

- Firstly, the user gives each column a heading. Headings should be between 1 and 80 characters long, unique, and only contain permitted characters including a-z, A-Z, 0-9, dashes, spaces and underscores. The heading should be user-friendly.
- Next, the user should choose the type of data in each column. Options include: Number, Text, Date-Time, Date & 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 any columns in excess of this number.
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 this data allows the search functions of the SHRP 2 Reliability Archive to find the uploaded artifact (Table 6-1).

<table>
<thead>
<tr>
<th>Name</th>
<th>Instructions</th>
<th>Data Sets?</th>
<th>Non-data Sets</th>
<th>Input Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Short descriptive title</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>5-10 line description which may include: purpose, origin of data, processing techniques, observations, findings or acknowledgements. Be concise.</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Project</td>
<td>Select a SHRP 2 project that your artifact is associated with.</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Class</td>
<td>Select SHRP 2 Primary for files that were produced as part of a SHRP 2 research project. Select User-submitted for all other files.</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Artifact Type</td>
<td>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.</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Related Artifacts</td>
<td>This field acknowledges that relationships can exist between artifacts. Further information provided below the table.</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years described</td>
<td>Year range described in this file. For example, the traffic data in your data set may be collected during 2011 and 2012.</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locations</td>
<td>This field gives you the opportunity to specify up to ten locations for the artifact. You may determine the location based on</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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;
- Someone on one project may use a data set from a different project.

This section can be utilized 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 also related to the artifacts from Northern Virginia.

<table>
<thead>
<tr>
<th><strong>Data dictionaries</strong></th>
<th>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.</th>
<th>✓</th>
<th>✓</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data sources(s)</strong></td>
<td>The organization that provided the data. This could include government bodies or third parties, e.g. traffic.com.</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Corridors</strong></td>
<td>This field captures the names of the road in a data set. For example, US101.</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Data Types</strong></td>
<td>Check the types of information that is included in this data set</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Collection Technologies</strong></td>
<td>How was the data collected? Choose the 'on site' field data collection technology</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Collection Frequency</strong></td>
<td>Time interval of data collection</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Days of Week</strong></td>
<td>The day of week that data was collected</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Holidays</strong></td>
<td>Indicate whether this data includes holidays or not</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6-29: Example of Related Data sets**
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. If the user would like to make changes before submitting, he/she 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 under My Profile page.

**6.7.4 Unsuccessful Artifact Processing**

At times, it may not be possible for the SHRP 2 Reliability 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 Edit Artifact Metadata**

The metadata of a processed or in-process artifact can be edited anytime by clicking on the edit pencil on the My Artifacts tab under the My Profile page (Figure 6-30).

![Figure 6-30: Editing an Artifact’s Metadata](image)

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 Delete Artifacts

The PI or the Creator can delete an uploaded artifact via the My Artifacts tab by clicking on the black and white cross next to the artifact (Figure 6-31). The My Artifacts tab is located on the My Profile page.

Figure 6-31: Deleting an Artifact

6.8 AUTOMATICALLY GENERATED EMAIL NOTIFICATIONS

This section lists the automatically generated email notifications that are used in the Archive. Usually these emails are used to: (1) Validate the identity of a registered user, (2) Notify the status of an uploaded artifact during the ingestion process, and (3) Inform if someone has commented on a PI’s or Creator’s artifact.

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 log in screen]

Kind regards,
Archive Administrator
SHRP 2 Reliability Program
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 to view all the artifacts you’ve uploaded; click ‘My profile’ and then ‘My Artifacts’ on the SHRP 2 Reliability Archive webpage.

Thank you for your contribution to the SHRP 2 Reliability Program archive!

Kind regards,
Archive Administrator
SHRP 2 Reliability Program

6.8.3 Artifact has Finished Processing
Subject: SHRP2 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 webpage.

Thank you for your contribution to the SHRP 2 Reliability Program archive!
6.8.4 Artifact Processing was Unsuccessful

Subject: SHRP2 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]
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

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)

AWS is a bundle of remote computing services that provides cloud-computing platform that is 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 our team’s point of view, the proposed L13 architecture (See Section 3.1.10) was slightly outdated. To that end, we
modified the proposed architecture and leveraged the extensive cloud-based services Amazon provides to the public. We 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 is a 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 our design we have not implemented a hot standby instance as a backup for cases the operation of the EC2 module fails. Amazon guarantees uptime of more than 99%. In case of any potential failure the admin 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, etc.) is 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 avoid errors. RDS supports three popular relational databases, i.e., MySQL, SQL Server, and Oracle. The Archive utilizes MySQL for managing its database system. As of April 2014, the size of the database was 500GB. The service is elastic. Therefore scaling up the storage is very easy.

- **Simple Storage Service (S3)/Glacier.** This service is used to backup the database and the file system. The Archive backs up the contents of the EBS daily and the RDS bi-weekly on S3. S3 keeps the data for one month and moves them to Glacier, a cost-efficient archival storage with very high availability and very low failure rate. It should be noted that sending/and retrieving data to/and from Glacier is slow. The size of the S3 storage service is 2TB (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 developers normally implement with WordPress because it provides PHP interpretation and URL rewriting.

#### 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 based on 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. SHRP 2 Archive theme is a child theme of WordPress’ Twentyeleven general theme. The SHRP 2 Archive theme was customized for the Archive user interface.

### 7.2.1.1 Key Open Source JavaScripts 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 Iteris 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.

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

### 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>
<thead>
<tr>
<th>Plugin</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes Plugin</td>
<td>A plugin that is used to handle inappropriate content, ratings, and such. This feature was implemented into the system but is not being used.</td>
</tr>
<tr>
<td>Custom Email</td>
<td>Sends custom email from SHRP 2 Archive plugins, and adds a custom registration email</td>
</tr>
<tr>
<td>L13a Ingestion</td>
<td>Implements the custom file ingestion process for the L13a reliability data archive.</td>
</tr>
<tr>
<td>L13A WP-Admin Restriction Mod</td>
<td>Hides the WordPress Admin banner on top of the site.</td>
</tr>
<tr>
<td>Meteor Slides</td>
<td>Easily creates responsive slideshows with WordPress that are mobile friendly and simple to customize. In the SHRP 2 Archive system, the admin has the ability to insert a slideshow at the homepage.</td>
</tr>
<tr>
<td>S2 Comment Form</td>
<td>A plugin to add custom fields to the comment form.</td>
</tr>
</tbody>
</table>
SHRP 2 Custom Meta | This plug-in 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 login, 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 WordPress existing tables and adding new tables. Section 7.3.1 and Section 7.3.2 will review the native WordPress tables and the SHRP 2-specific tables in more details.

Please note the Archive stores data sets in two formats:

- Flat file, which is the original .csv format kept in WordPress’ file system.
- 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,’ where ArtifactID is a unique identifier that is assigned to each file (artifact) by WordPress. Please 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 7-3: List of the SHRP 2 Database Tables

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Description</th>
<th>Created By</th>
</tr>
</thead>
<tbody>
<tr>
<td>wp_commentmeta</td>
<td>Each comment features information called the meta data and it is stored in the wp_commentmeta.</td>
<td>WordPress*</td>
</tr>
<tr>
<td>wp_comments</td>
<td>The comments within WordPress are stored in the wp_comments table.</td>
<td>WordPress</td>
</tr>
<tr>
<td>wp_links</td>
<td>The wp_links holds information related to the links entered into the Links feature of WordPress. <em>(This feature has been depreciated, but can be re-enabled with the Links Manager plugin.)</em></td>
<td>WordPress</td>
</tr>
</tbody>
</table>
The Options set under the Administration > Settings panel are stored in the wp_options table. See Option Reference for option name and default values.

Each post features information called the meta data and it is stored in the wp_postmeta. Some plugins may add their own information to this table.

The core of the WordPress data is the posts. It is stored in the wp_posts table. Also Pages and navigation menu items are stored in this table. This table is also 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 has been modified.

The categories for both posts and links and the tags for posts are found within the wp_terms table.

Posts are associated with categories and tags from the wp_terms table and this association is maintained in the wp_term_relationships table. The association of links to their respective categories are also kept in this table.

This table describes the taxonomy (category, link, or tag) for the entries in the wp_terms table.

Each user features information called the meta data and it is stored in wp_usermeta.

The list of users is maintained in table wp_users.

The tables stores the relationships among artifacts.

The column types of each data set are stored in this table.

Artifact_ID represents artifact ID of a data set (automatically generated by WordPress). This table stores the content of a data set and is used for visualizing and filtering.

This table stores the ingestion state of all the artifacts.

* For more information on WordPress database description visit [http://codex.wordpress.org/Database_Description](http://codex.wordpress.org/Database_Description).

Table 7-4 to Table 7-8 show fields in tables created or modified for the Archive. Table names starting with ‘s2’ represents relations specifically created for the Archive system.

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
<th>Extra</th>
</tr>
</thead>
<tbody>
<tr>
<td>artifact_id</td>
<td>int(11)</td>
<td>NO</td>
<td>PRI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relation_id</td>
<td>int(11)</td>
<td>NO</td>
<td>PRI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
<th>Extra</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>bigint(20)</td>
<td>NO</td>
<td>PRI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>idx</td>
<td>int(10)</td>
<td>NO</td>
<td>PRI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>Type</td>
<td>Null</td>
<td>Key</td>
<td>Default</td>
<td>Extra</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------</td>
<td>------</td>
<td>-------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>name</td>
<td>varchar(80)</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>type</td>
<td>int(10)</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>label</td>
<td>varchar(80)</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>min_range</td>
<td>varchar(80)</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>max_range</td>
<td>varchar(80)</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7-6: S2_dset_ArtifactID table fields**

- `_rowid_` bigint(20) NO PRI auto_increment
- Data Set Column*
- * This table stores data sets columns. The field and type varies depending on the data set.

**Table 7-7: wp_posts table fields**

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

Table 7-8: s2_events table fields

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 till the moment it becomes available in (or gets deleted from) the Archive. S2A core functionalities are listed in Section 5.2 (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 stored in 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 artifacts. See Section 5.2 (Step 4. back-end processing) for more information.
- **s2_proc_msg** – Provides processing outcomes in a message for the Creator. The massage is displayed on ‘My Artifact’ list located on ‘My Profile’ page.
<table>
<thead>
<tr>
<th>State Variables</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
</table>
| 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 pressed in Step 4 of the ingestion process  
2=Processing  
  ● Triggered by: admin reviews and approves  
1=Published, available for public use  
  ● Triggered by: s2a server completes processing  
-1=Pre-trash  
  ● Triggered by: admin action, sending 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.” |
Figure 7-3: The 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:

- 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 give confidence that system functionality is not broken by code changes.
- Integration – Verify that the SHRP 2 Archive functions in the required target environments.
- Security – Verify that necessary software security updates have been applied.

Please 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. We recommend the readers check the latest version of the Test Plan document that will be provided separately after completion of the project.

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 utilize the Bugzilla platform to keep track of the bugs and enhancement stories.

8.1.1 Unit Testing

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. Since the team is following the agile approach, the programmers are mandated to:

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

It should be noted 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.
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 is displayed correctly. Errors, if any, must be caught by the corresponding plugin and logged or shown to the Administrator only, but 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 non-numeric 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

Objective
The objective of this section of tests is to 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 meta data
4. Visualization of previously loaded artifacts
5. Downloading previously loaded artifacts

Test Cases

8.2.1.1.1: Verify that the userid: L13ATestUser, password: L13AsECURITYiSaWESOME! can login.

8.2.1.1.2: Verify that the userid: L13ATestUser, password: L13AsECURITYiSaWESOME! can login. Note that userids are not case-sensitive.

8.2.1.1.3: Verify that the userid: .L13ATestUser, password: L13AsECURITYiSaWESOME! cannot login.

8.2.1.1.4: Verify that the userid: L13ATestUser, password: L13aSecurityIsNotAwesome cannot login.

8.2.1.1.5: Verify that a user’s login credentials are cleared when browser is restarted.

Steps:
a) Log into the SHRP 2 Archive
b) Close the browser used to login.
c) Restart the browser
d) Go to the SHRP 2 archive URL. Verify that the user must login again.

8.2.1.2.1: Search criteria (values subject to change): By Project: Project L02

Returns: Orange 5 over Atlanta, Georgia
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

8.2.1.2.2 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

8.2.1.2.3 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

8.2.1.2.4 Search criteria (values subject to change): Non-Data 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…

8.2.1.3.1 Upload artifact A1 and assign the following metadata. Steps:

a) Go to Upload a New Artifact Wizard in the My Profile dialog
b) Select File -> Click Browse… and use chooser to locate and load the artifact.
c) Select “Save and Continue”
d) Complete Set Metadata dialog
   i) Set Description to “Test Artifact”
   ii) Set Project to “Project L15 – Innovative IDEA Projects”
   iii) Set Artifact Type to “Final Report”
   iv) Set Locations to “Sacramento, California”
   v) Click Save and Continue
e) The Publish Artifact dialog appears
   i) Click Submit

8.2.2 Performance Tests
Objective
The objective of this section of tests is to verify that the performance requirements are satisfied. They include:

1. UI responsiveness (UIR) as follows:
   a. Basic features return in less than 3 seconds for 90% of requests
   b. Search returns in less than 5 seconds for 90% of requests
c. Queries on the Data tab return in less than 30 seconds per GB of content for 90% of time

d. Visualization returns in less than 5 seconds per GB of content

e. Ingest completes in less than 30 seconds per GB of content.
   i. Assuming an upload feed of 3 MB/second or more.

2. Newly ingested artifacts appear in search results within four hours of upload

3. Deleted artifacts no longer appear in search results within four hours of deletion

Test Cases

8.2.2.1.a.1 Time interval from SHRP 2 Login page to SHRP 2 Landing page meets UIR objective.

8.2.2.1.a.2 Time interval from SHRP 2 Landing page to Artifact page meets UIR objective.

8.2.2.1.a.3 Time interval from Artifact page, Metadata tab to Discussion tab meets UIR objective.

8.2.2.1.a.4 Time interval from Artifact page, Discussion tab to Home tab meets UIR objective.

8.2.2.1.a.5 Time interval from SHRP 2 Landing page, Home tab to Search Archive tab meets UIR objective.

8.2.2.1.a.6 Time interval from SHRP 2 Landing page, Search Archive tab to Explore by Focus Area / Project tab meets UIR objective.

8.2.2.1.a.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.

8.2.2.1.a.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.

8.2.2.1.a.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.

8.2.2.1.a.10 Time interval from SHRP 2 Landing page, Explore by Focus Area / Project Tab; navigate from Safety Focus Area to Other meets UIR objective.

8.2.2.1.a.11 Time interval from SHRP 2 Landing page, Explore by Focus Area / Project Tab; navigate from Safety Focus Area to Home meets UIR objective.

8.2.2.1.b.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.

8.2.2.1.b.2 Starting from SHRP 2 Landing page, Search Tab; orange circle #11, above San Francisco, CA selected, 11 bubbles visible. Click on 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.
8.2.2.1.c.1 Starting this artifact page located at: 
http://shrp2archive.org/?attachment_id=848; go to Data tab; select date range from 07/01/08 to 07/02/08; the query result will show up within the time interval UIR 1.c.

8.2.2.2.1 Verify ingested artifacts appear in search results within UIR 2 timeframe. The steps are:

a. Ingest artifact A2 into the SHRP 2 Archive.
b. As the Administrator, process the artifact.
c. Start a timer and wait for time interval UIR 2 to pass.
d. Complete a SHRP 2 search and verify that artifact A2 appears in the result set.

8.2.2.3.1 Verify that deleted artifacts do not appear in search results within the UIR 3 timeframe. The steps are:

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

Objective

The objective of this section of tests is to verify that the procedures designed to guarantee 99% of availability function correctly. They include:

1. Time required to detect that login is not functional limited to no more than five minutes
2. Time required to detect that database is no longer communicating to no more than five minutes
3. Time required to fail over from primary server to backup server completes in less than four hours
4. Time required to rebuild a failed database is less than two hours
5. Data loss limited to no more than the artifacts loaded in the last 24 hours.

Test Cases

8.2.3.1.1 Verify that the login monitoring process detects the inability to login within the timeframe listed in the objectives description above. The steps are:

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 Objectives section above

8.2.3.2.1 Verify that the database connection monitoring process detects the inability to connect to the database. The steps are:

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.
8.2.3.3.1 Verify that the server failover process completes within the time frame described in the Objectives description above. The steps are:

  a. Stop the primary server
  b. Start a timer
  c. Verify that the backup server takes over within the time frame described in the Objectives section. A backup server takeover is successful if a user can login and perform Test Case 4.1.2.1.

8.2.3.4.1 Verify that the database reconstruction process can complete a DB rebuild within the time frame described in the Objectives description above. The steps are:

  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 section 8.2.1 Functional Tests.

8.2.3.5.1 Verify that an artifact loaded over 24 hours ago is included in a backup server’s archive inventory. The steps are:

  a. Load artifact A3 into the Primary SHRP 2 server.
  b. Wait 24 hours
  c. Verify that artifact A3 is present in the artifact list in the backup server.

8.2.4 Scalability Tests

Objective
The object of this section of tests is to verify that the system’s scalability requirements are satisfied. They are:

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 timeframe described in the Performance Testing section. Artifacts larger than 2.5 GB are rejected.

Test Cases
8.2.4.1.1 Run five instances in parallel of the tests described in Section 8.2.2 Performance Tests. Verify that the Performance requirements are satisfied.

8.2.4.2.1 Verify that artifact A4, which is 2.5 GB in size, can be ingested into the archive

8.2.4.2.2 Verify that artifact A5, which is greater than 2.5 GB in size, is rejected within the ingest process.

8.2.5 Maintainability Tests

Objective
The objective of this section is to 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:
1. Inspecting each class to verify that unit tests exist that:
   a. Verify that invalid values of each incoming parameter are detected and the
      further processing is prevented.
   b. Verify positive operation of the class with at least one test case.
   c. Verify invocation of each “catch” block in at least one test case.

**Test Procedure**

8.2.5.1 Verify unit test existence, the steps are:

   a. Procure the source code for the project along with the test cases
   b. For each class inspect the unit tests and verify:
      i. There are test(s) that verify validity of code processing the incoming
         parameters.
      ii. There is at least one test that verifies positive operation.
      iii. 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**

**Objective**

The object of this section is to 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 9 on Windows 7
2. Safari 5.1 on Mountain Lion (10.8)
3. Firefox 25.0 on Windows 7
4. Chrome 30.0 on Windows 7

**Test Procedure**

8.2.6.1 Run the tests listed in sections 8.2.1 Functionality Tests and 8.2.2 Performance
Tests using IE9 on Windows 7.

8.2.6.2 Run the tests listed in sections 8.2.1 Functionality Tests and 8.2.2 Performance
Tests using Safari 5.1 on Mountain Lion (10.8)

8.2.6.3 Run the tests listed in sections 8.2.1 Functionality Tests and 8.2.2 Performance
Tests using Firefox 25.0 on Windows 7

8.2.6.4 Run the tests listed in sections 8.2.1 Functionality Tests and 8.2.2 Performance
Tests using Chrome 30.0 on Windows 7.

**8.2.7 Security Tests**

**Objective**

The Objective of this section is to 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
6. Lucene and SOLR

**Test Procedure**
For each of the software modules listed in the objectives section:

b. Apply remediation within the timeframe described in the objectives section.

**8.2.8 Visual GUI Testing**

**Objective**
The objective of this test is to verify the visual appealing of all graphical user interface elements of the Archive.

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

8.2.8.1. Homepage

8.2.8.1.a. Slider on the home page

8.2.8.1.b. Latest artifacts on the Recent Artifacts list. Long titles and descriptions should be truncated.

8.2.8.2. Top menus

8.2.8.3. Metadata page

8.2.8.3.a. Metadata page and position of the text when the artifact metadata contains a lot of information.

8.2.8.3.b. Metadata map

8.2.8.3.c. Leaflet credentials has to be viewable on the map.

8.2.8.3.d. Metadata information. Check the metadata information to make sure they are consistent.

8.2.8.3.e. Data tab’s left and right containers.

8.2.8.3.f. Text on the filter dropdown menus has to be readable.

8.2.8.4. Grid View
8.2.8.4.a. Scroll bar on the grid page.

8.2.8.4.b. Scroll bar on the filter page when the user enters too many filtering criteria.

8.2.8.5. Graph View

8.2.8.6. Map View

8.2.8.7. Discussion Tab

8.2.8.7.a. Location of the rating stars

8.2.8.8. Ingestion Page

8.2.8.9. User profile page

8.2.8.10. Search Archive page

8.2.8.10.a. Leaflet credentials has to be viewable on the map.

8.2.8.10.b. Filter check boxes.

8.2.8.10.c. Search results on the map.

8.2.8.10.d. List results.

8.2.9  Testing Automation

Functionality and Performance test cases were automated using Selenium framework. The testing code was written in Python. After each modification (according to the stories submitted through Bugzilla platform), the testing team had to run the code to make sure other elements of the system are not affected. The code is written in a way that it can be run against different browsers, i.e., Firefox, Chrome, and IE.
### 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 the Test Plan*

<table>
<thead>
<tr>
<th>Artifact Number</th>
<th>Test</th>
<th>Relevant characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>8.2.1.3.1</td>
<td>Used to demonstrate artifact upload function.</td>
</tr>
<tr>
<td>A2</td>
<td>8.2.2.1</td>
<td>Used to demonstrate artifact upload time</td>
</tr>
<tr>
<td></td>
<td>8.2.2.3.1</td>
<td>Used to demonstrate artifact deletion</td>
</tr>
<tr>
<td>A3</td>
<td>8.2.3.5.1</td>
<td>Used to demonstrate artifact upload propagates to backup server</td>
</tr>
<tr>
<td>A4</td>
<td>8.2.4.2.1</td>
<td>2.5 GB in size</td>
</tr>
<tr>
<td>A5</td>
<td>8.2.4.2.2</td>
<td>Greater than 2.5 GB in size</td>
</tr>
</tbody>
</table>
CHAPTER 9: NOTES ON OPERATION AND MAINTENANCE OF THE ARCHIVE

The design and operation of the Archive system is dependent on not only the requirements driven by the users but also the financial, technical and policy-related constraints. Although the L13 report attempted to shed light on those issues (i.e., life-cycle costs, archiving approaches, etc.), 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, operation and maintenance of the Archive system by developing white papers that were put into discussion among the members of T-ETG and the SHRP 2 team. The papers were structured in a manner that would provide solution alternatives and were intended for obtaining stakeholders’ feedback. They only reflected the Iteris team’s perspectives at the time of development and were designed to trigger internal discussions at 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 based on the papers did not exactly follow the suggestions provided in the papers due to the evolving nature of the project.

This chapter summarizes the project team’s assessment of various issues and final conclusions made based on the papers. It also includes concerns that were raised in the white papers in order to draw the SHRP 2 management team’s attention to the Archive’s key operation 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. Operation and Maintenance
3. Data Ownership and Personally Identifiable Information

9.1 INCLUSION OF USER-SUBMITTED DATA IN THE ARCHIVE

One of the outcomes of the June 4th, 2012 L13A workshop was the Subject Matter Expert (SME) panel’s suggestion to add a feature that allows the Archive users to submit their processed/transformed data sets and objects, derived from the original archived data, back into the Archive. SHRP 2 believed feeding back the user-generated products was aligned with the SHRP 2 strategic goals and was 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?

We propose three user submission scenarios:

**Scenario 1 (All the users can upload flat files only):**
- Users can only submit a flat file (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 is not corrupted or infected, but no preprocessing step would be required.
- Users would be able to submit their objects under the community pages.

**Scenario 2 (All the 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 pre-processed 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) that includes data extracted from various sensors/segments, or a set of sensor-level/segment-level data sets (in .csv or .xls format) where each set represents data collected from a single sensor/segment.

**Scenario 3 (Trusted users can upload any files with no file type limit):**
- In terms of file upload constraints, this scenario is similar to Scenario 2. The only difference is that only a trusted group of users (in addition to PIs) can upload artifacts. At the time of writing the white paper this scenario was not discussed an option. It was added later after in depth discussion with SHRP 2 and FHWA team.

Please note that any artifact submitted via any scenario is grouped as “User-Submitted” artifact.

### 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, Scenario 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.

### 9.1.3 Conclusion

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

- It provides all of the envisioned functionality for the SHRP 2 archived data (see Table 9-1).
- It uses 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 assist the Administrator with identifying irrelevant artifacts. However, 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 through monitoring process to prevent users submitting
PII data is too high for SHRP 2. Therefore, per SHRP 2 request, the team only implemented Scenario 3 in which only a trusted group of users, name SHRP 2 contractors, can upload artifacts for the time being.

Lastly, the team believes the adverse implications of an excessive storage space requirement for hosting user-submitted data is not significant enough, when compared to the benefits, as long as users submit valuable artifacts to the Archive. As a result, we propose 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.

Table 9-1: Supported functionality for each user submission scenario

<table>
<thead>
<tr>
<th>Features</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>List Search</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Map Search</td>
<td>●</td>
<td>●**</td>
<td>●**</td>
</tr>
<tr>
<td>Full Download</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Sub-set Download</td>
<td>●</td>
<td>●**</td>
<td>●**</td>
</tr>
<tr>
<td>Visualization</td>
<td>●**</td>
<td>●**</td>
<td>●**</td>
</tr>
<tr>
<td>Collaboration</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

* No sensor location.

** Pre-processing of the data set is needed to leverage the feature.
Table 9-2 Comparison of Scenario 1, Scenario 2, and Scenario 3 for Key Criteria

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHRP 2 Strategic Alignment</td>
<td>Alignment with SHRP 2 strategic goals</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Some features are not supported in Scenario 1.</td>
</tr>
<tr>
<td>Cost</td>
<td>Direct cost of hardware</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Team will use cloud-computing model.</td>
</tr>
<tr>
<td></td>
<td>Cost of software development</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Project team will leverage the existing data ingestion feature for Scenario 2 and 3.</td>
</tr>
<tr>
<td></td>
<td>Cost of internet and web services (i.e. cloud)</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Scenario 2 cost is higher because it requires more storage space.</td>
</tr>
<tr>
<td></td>
<td>Recurrent/operation and maintenance costs</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Scenario 2 requires ample admin time to review the artifacts submitted by the regular users.</td>
</tr>
<tr>
<td>Changes for provision of backup services</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and equipment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>Back-end coding effort</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>See “Direct cost of software development”</td>
</tr>
<tr>
<td></td>
<td>Alignment with user requirements</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UI development / implementation effort</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Metadata entry effort</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Required database size</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>Administration</td>
<td>Admin staff effort/cost for processing and</td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
<td>Scenario 2 requires more admin time to review submitted artifacts by the users.</td>
</tr>
<tr>
<td></td>
<td>validation of artifacts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Risk to the project success (budget /schedule</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td>risks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adverse effect of technology evolution on the</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td>system operation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
9.2 KEY ISSUES ASSOCIATED WITH OPERATION AND MAINTENANCE OF THE ARCHIVE

The core objective of this section is to review the various alternatives, as well as their implications, for the operation 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 operation and maintenance requirements?
- What are hosting options for the SHRP 2 L13A system operations and maintenance phase?
- What are the operation and maintenance costs?

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 System and Operations Requirements?

The Archive is currently designed for 99% availability using routine backup and recovery systems and processes. This guarantees for 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 needs to be recoverable, and the annual availability needs to be met
  - No outage will be greater than 72 hours (only would occur with major system failure where the new system would need to re-index 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
  - Backup frequency: Daily
- Software Maintenance
Server Patches & Updates:
- CentOS, MySQL DB (Quarterly)

Software Patches & Updates:
- WordPress, PHP, Highchart (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 Terabytes

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

- Option 1: Server Based with Server Backup (Using Existing Data Center)
- Option 2: Server Based with Server Backup (Hosted at Highly Available Data Center)
- Option 3: Server Based with Cloud Backup (Hybrid)
- 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, with the risks being a single point of failure and the timeliness of recovery in the event of a catastrophic issue, which 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. Option 4 is the recommended option. Each option is described in more detail below.

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

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 already paid-for equipment, 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
  o New Equipment
  - Additional Disk Space for Existing Server
  - Backup Server
  - Backup System for Backup Server
  - Installation
  - Total: $6,500
  o Support for two server based systems (8 hours/wk): $75,000/yr
    ▪ Review & Maintenance
      • Weekly deployment review
      • Patches
      • Functionality bug fixes
    ▪ Backup & Recovery
      • Annual recovery verification
      • Full backup (Every 3 months) (1-2 TB)
      • Incremental backups
        o User related tables – (Daily) (500 GB)
        o New artifacts – (Upon upload) (10GB) triggered by admin process

Option 2: Server Based with Server Backup (Hosted at Highly Available [HA] 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
  o Purchasing a second server to support backup and hot recovery
  o Moving server location to HA Data Center
• Benefits
  o Uses already paid-for equipment, with only a marginal cost for a secondary server
  o Redundant power and network
  o Space is available for servers at HA Data Center
  o Could add warranties to extend the expected lifetime of server up to 7 years (approximately $140/yr)
  o Spare systems available on site
  o The systems are configured with OS installations on a dedicated Redundant Array of Independent Disks (RAID)-1 pair and data storage on a separate RAID-10 array
  o 24/7 Support Staff
• Limitations
  o Equipment will need to be replaced every 3 years (unless warranty extension is used)
  o One time additional cost to install system in the data center
• Cost
  o New Equipment
    ▪ Additional Disk Space for Existing Server
    ▪ Backup Server
    ▪ Backup System for Backup Server
    ▪ Installation
    ▪ Total: $6,500
  o Transition to HA Data Center
    ▪ Cost to move and reinstall: $2,000
  o Support for two server-based systems (8 hours/wk): $75,000/yr
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 backup 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-hours 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/yr
  - Transition to Cloud Database Backup
    - Cost to move and reinstall: $1,000
  - Support for two systems (8 hours/wk): $75,000/yr
    - Review & Maintenance
      - Weekly deployment review
      - Patches
      - Functionality bug fixes
    - Backup & Recovery
      - Annual recovery verification
      - Full backup (Every 3 months) (1-2 TB)
      - Incremental backups
        - User related tables – (Daily) (500 GB)
        - New artifacts – (Upon upload) (10GB) triggered by admin process
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/yr
  - Transition to Cloud Hosting and Database Backup
    - Cost to move and reinstall: $2,000
  - Support for two cloud-based systems (8 hours/wk): $75,000/yr
    - Review & Maintenance
      - Weekly deployment review
      - Patches
      - Functionality bug fixes
    - Backup & Recovery
      - Annual recovery verification
      - Full backup (Every 3 months) (1-2 TB)
      - Incremental backups
        - User related tables – (Daily) (500 GB)
        - New artifacts – (Upon upload) (10GB) triggered by admin process

9.2.4 What are Operation & 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.

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.
- 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.
- Project Management: Costs to manage the transition to the O&M phase.

**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.
- Support Years: Number of years used to calculate annual cost values (Amazon hosting costs).

**Table 9-3: Archive System Options Summary**

<table>
<thead>
<tr>
<th>Option</th>
<th>Type</th>
<th>Primary System</th>
<th>Backup</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Server Based</td>
<td>Server</td>
<td>Server</td>
<td>Existing Data Center</td>
</tr>
<tr>
<td>2</td>
<td>Server Based</td>
<td>Server</td>
<td>Server</td>
<td>High Availability Data Center</td>
</tr>
<tr>
<td>3</td>
<td>Hybrid</td>
<td>Server</td>
<td>Cloud (Data Only)</td>
<td>Existing/HA Center/Amazon</td>
</tr>
<tr>
<td>4</td>
<td>Cloud Based</td>
<td>Cloud</td>
<td>Cloud</td>
<td>Amazon</td>
</tr>
</tbody>
</table>

**Table 9-4: Archive System Operation and Maintenance Costs Summary**

<table>
<thead>
<tr>
<th>Item</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Disk Space</td>
<td>$500</td>
</tr>
<tr>
<td>Backup Server</td>
<td>$3,000</td>
</tr>
<tr>
<td>Backup System</td>
<td>$2,000</td>
</tr>
<tr>
<td>Installation Costs</td>
<td>$1,000</td>
</tr>
<tr>
<td>Server Transition Costs</td>
<td>$0</td>
</tr>
<tr>
<td>Training Transition</td>
<td>$2,500</td>
</tr>
<tr>
<td><strong>One-Time Costs</strong></td>
<td><strong>$9,000</strong></td>
</tr>
<tr>
<td>Annual Support</td>
<td>$75,000</td>
</tr>
<tr>
<td>Server Warranty</td>
<td>$0</td>
</tr>
<tr>
<td>Annual Cloud On-Demand</td>
<td>$0</td>
</tr>
<tr>
<td>Annual Cloud Heavy Reserved</td>
<td>$0</td>
</tr>
<tr>
<td>Support Years</td>
<td>1</td>
</tr>
</tbody>
</table>
9.2.4.3 Hosting Option

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. Utilizing 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.

![Figure 9-1: L13A Options Uptime Risks](image_url)
9.3 MANAGING DATA RIGHTS AND PERSONALLY IDENTIFIABLE INFORMATION IN A DYNAMIC SYSTEM CONTAINING NON-SHRP 2 DATA

It was stated earlier that data in the SHRP 2 Archive comes just from SHRP 2 Reliability related Projects. The current Archive is currently configured to house a static data set although in the future 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 many vehicles that pass over a loop in a time slice such as a five minute period. Thus it is not possible to identify individual vehicles from the loop data in the SHRP 2 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 there are no personal identifiers 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 was 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 – is 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.

Further assurance regarding the absence of PII data in the SHRP 2 Archive is accumulating because 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.

In this section we provide a review of industry practices for data rights protections, author attribution and options for protecting PII that might be inadvertently added by users to the Archive in the event that the Archive was opened up to non-SHRP contractors in the future to enable users to submit any Reliability-related artifacts. We offer options to manage data licenses and address PII data protections (in case the user artifact upload feature were to become 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 and specifically that is would be desirable if the SHRP 2 Archive were 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 of uncontrolled data that requires 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 the any contributors to, and users of, the data archive 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 system 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 five years, many new open data sites have been created by the public sector. As data rights provisions are legal terms, we examined data rights protections used in open data implementations that follow the same or similar legal structure that would apply to SHRP 2 data.

9.3.2 Open Data Licensing Options
We propose 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 so that individuals can include or exclude data with specific license types from their queries.

9.3.2.1.1 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 the data artifact that is public domain is shown in Figure 9-2. If the artifact is not in the public domain, there are six other licenses that 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:

- **Attribution**: Ensures that authors of the artifact are mentioned appropriately in derivative works for commercial or non-commercial 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.
- **Non-Commercial**: Allows non-commercial 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).

9.3.2.2 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 non-profit organization based in Great Britain that supports open data projects around the globe. The OKF 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, compared to content licenses provided by CC, 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”
- Open Database License — “Attribution Share-Alike for data/databases”

Unlike the CC licenses, OKF licenses do not provide any limitations for commercial use or non-derivatives. The foundation also provides a narrative that illustrates the differences in database vs. 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 non-homogenous 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, 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 “Database” and “Contents”)

When the owner of the database and the content of the database are different, OKF calls this a non-homogenous database and uses the following rights:

- Share-alike: use ODbL for Database qua Database plus whatever license you want/can for Contents
- Public domain: use PDDL for Database qua Database plus whatever license you wish/can for Contents

Note the CC licenses could be used in conjunction with the OKF licenses in this case to appropriately license the content.

9.3.2.3 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, it is suggested that one have a business process in place to request 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 prior to posting the data, database or other artifact to the system.

To manage data rights of an open archive, many sites utilize 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 the users select the license to submit, and an administrator is able to review the submission before the data is available to public access. 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 datasets 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 (United States Federal Chief Information Officers Council 2012) expands on a strategy document called “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. 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 2012)
   b. Define PII for the SHRP 2 Archive (McCallister et al. 2010)
   c. Determine if data are linked or can data be linked (“linkable”) to a specific individual.
   d. Use an existing System of Records Notice (SORN) or draft a new SORN, if required (US Department of Transportation 2014).
   e. Determine the role of PII in the inventory.
   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.
   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, 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?

To check whether PII exists in the SHRP 2 Archive and whether files that are added by users contain PII, there are several other methods recommended in the NIST Guide, including “reviewing system documentation, conducting interviews, conducting data calls, using data loss prevention technologies (eg. 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. How to measure risk for PII is also reviewed in the NIST guide including impact level definitions for low, moderate and high risk PII and factors for determining PII confidentiality impact levels 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 of the data protection.
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 datasets 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. Include in the language that the data being uploaded is free of PII and that the data has become unlinked or anonymous. Additionally, the data should be posted to an administrative area for a review of the data set to determine if the data contains any PII prior to becoming accessible to users of the archive.

Once data is in the Archive, the Administrator must serve as a data steward that performs an initial review of all the data. The data steward role 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, although each step reduces the risks and assigns traceability to the appropriate parties. However, realistically no one is going to have a reason and 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 where 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


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, etc.). Summarize the major points about the data collection. This might include:

- Detector type (Bluetooth, loop, etc.);
- Road / Road authority that owns the road (i.e. state);
- Number of stations;
- Date / duration of data collected;
- Other relevant information i.e. whether it was in a construction site, poor weather etc.

**Processing Techniques**

Quickly summarize any processing techniques used in 1-2 sentences. Any users that 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.
Table A-1: Column Descriptions

<table>
<thead>
<tr>
<th>CSV Column header title</th>
<th>Ingested data set column header title</th>
<th>Column Description</th>
<th>Units of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED</td>
<td>Example – Speed</td>
<td>Average speed of vehicles passing the detector station</td>
<td>Miles / hour</td>
</tr>
<tr>
<td>OCC</td>
<td>Occupancy</td>
<td>Average occupancy of the detector station</td>
<td>Percentage</td>
</tr>
</tbody>
</table>

Acknowledgments

Acknowledge those who:

- Supplied the detector data i.e. road authorities or other organizations;
- Primarily processed the data; and
- Others than contributed (but no personal acknowledgements).
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 provide 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
  - Authorize and Monitor Security State:

- NIST 800-37: Guidelines for Applying the Risk Management Framework to Federal information Systems