Requirements and Feasibility of a System for Archiving and Disseminating Data from SHRP 2 Reliability and Related Studies

DETAILS

0 pages | null | PAPERBACK
ISBN 978-0-309-43039-5 | DOI 10.17226/22881

AUTHORS

The Second
STRATEGIC HIGHWAY RESEARCH PROGRAM

SHRP 2 REPORT S2-L13-RW-1

Requirements and Feasibility of a System for Archiving and Disseminating Data from SHRP 2 Reliability and Related Studies

ZONGWEI TAO, JEFFREY SPOTTS, AND ELIZABETH HESS
Weris, Inc.
Reston, Virginia

TRANSPORTATION RESEARCH BOARD

WASHINGTON, D.C.
2011
www.TRB.org

Subscriber Categories

Highways
Data and Information Technology
Operations and Traffic Management

The Second Strategic Highway Research Program

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

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

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

SHRP 2 Report S2-L13-RW-1

ISBN: 978-0-309-12900-8

Copyright Information

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

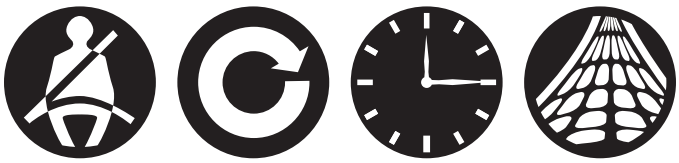
Notice

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

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

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

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



SHRP 2 Reports

Available by subscription and through the TRB online bookstore:

www.TRB.org/bookstore

Contact the TRB Business Office:
202-334-3213

More information about SHRP 2:

www.TRB.org/SHRP2

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by the 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. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg 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. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

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

www.national-academies.org

SHRP 2 STAFF

Neil F. Hawks, *Director*
Ann M. Brach, *Deputy Director*
Kizzy Anderson, *Senior Program Assistant, Implementation*
Stephen Andrle, *Chief Program Officer, Capacity*
James Bryant, *Senior Program Officer, Renewal*
Mark Bush, *Senior Program Officer, Renewal*
Kenneth Campbell, *Chief Program Officer, Safety*
JoAnn Coleman, *Senior Program Assistant, Capacity*
Eduardo Cusicanqui, *Finance Officer*
Walter Diewald, *Senior Program Officer, Safety*
Jerry DiMaggio, *Implementation Coordinator*
Charles Fay, *Senior Program Officer, Safety*
Carol Ford, *Senior Program Assistant, Safety*
Elizabeth Forney, *Assistant Editor*
Jo Allen Gause, *Senior Program Officer, Capacity*
Abdelmename Hedhli, *Visiting Professional*
Ralph Hessian, *Visiting Professional*
Andy Horosko, *Special Consultant, Safety Field Data Collection*
William Hyman, *Senior Program Officer, Reliability*
Linda Mason, *Communications Officer*
Michael Miller, *Senior Program Assistant, Reliability*
Gummada Murthy, *Senior Program Officer, Reliability*
David Plazak, *Senior Program Officer, Capacity and Reliability*
Robert Raab, *International Coordinator*
Monica Starnes, *Senior Program Officer, Renewal*
Noreen Stevenson-Fenwick, *Senior Program Assistant, Renewal*
Charles Taylor, *Special Consultant, Renewal*
Dean Trackman, *Managing Editor*
Hans van Saan, *Visiting Professional*
Pat Williams, *Administrative Assistant*
Connie Woldu, *Administrative Coordinator*
Patrick Zelinski, *Communications Specialist*

ACKNOWLEDGMENTS

This work was sponsored by 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. The project was managed by David J. Plazak, Senior Program Officer for SHRP 2 Capacity and Reliability.

The research reported herein was performed by Weris, Inc. Dr. Zongwei Tao, PE, Founder and President of Weris, Inc., was the Principal Investigator. The other authors of this report are Jeffrey Spotts and Elizabeth Hess, Weris associates with expertise in digital archiving and information technology, respectively.

FOREWORD

David J. Plazak, *SHRP 2 Senior Program Officer*

This report provides an assessment of the feasibility of developing and populating an online archive for the great variety and volume of data being produced by the SHRP 2 Reliability focus area research program. The goal of the archive, if feasible, is to provide researchers and other interested parties with ready access to data needed to independently validate the results of SHRP 2 Reliability research and to conduct follow-on research. For this project, the term "data" was defined in the broadest way possible to include statistical data, analytical tools and models, written reports, pictorial data, and video data.

Extensive, high-quality data resources are critical to the understanding of nonrecurring highway congestion and travel time reliability. The reliability of transportation facilities can only be assessed in the context of a statistical distribution of travel times. A number of things affect travel times on a day-to-day basis, including fluctuations in travel demand, inadequate base capacity, weather, traffic incidents, special events, work zones, and poorly functioning traffic control devices. Months and months of travel time data and related data such as weather conditions are needed to understand reliability problems and how they can be addressed on a regional or corridor basis.

The report establishes several alternative information technology architectures that could be used to develop an online reliability data archive, and it analyzes their advantages, disadvantages, and costs. The report recommends a solution based on cloud computing data storage and a mixture of open-source and commercial, off-the-shelf software. This alternative was further assessed through the construction of a prototype data archive, which is also described in the report. The prototype archive was populated with a variety of data from SHRP 2 Reliability Project L03, which created a large quantity of data holdings of different types. Finally, the report finds that a SHRP 2 Reliability Archive is feasible and recommends that SHRP 2 move ahead with projects to design, build, and populate an online archive for the Reliability focus area. Work on the data archive is planned to be under way in 2011. It will ultimately be populated with data from all of the SHRP 2 Reliability research projects and closely related projects from other SHRP 2 research focus areas.

CONTENTS

Executive Summary

Introduction

The SHRP 2 Reliability focus area aims to benefit society by sponsoring research projects that seek to reduce highway congestion through incident reduction, management, response, and mitigation.

The data and analytical approach behind these research products can also be of significant long-term value if researchers can build new research and products based on this foundation. This requires that the base data, derived data, analytical models and tools, and so forth that make up the intellectual framework of the SHRP 2 Reliability research projects be preserved and made accessible. Transportation practitioners can also benefit from having access to much of the same information, particularly access to the research products that reflect the implementation of the SHRP 2 research efforts.

The experience of the first SHRP program, where 13 of the 14 major databases collected during that program are no longer accessible, illustrates the need to plan for data preservation and access well in advance. The preservation of so-called born-digital information is much harder than the preservation of traditional paper records because of rapid changes in technology, which, in turn, lead to rapid obsolescence of data formats and storage media.

Reliability Project L13, Requirements and Feasibility of a System for Archiving and Disseminating Data from SHRP 2 Reliability and Related Studies, was designed to assess the technical, economic, and business aspects of developing, operating, and maintaining a long-lived archival system that preserves and makes readily available to researchers and practitioners the data from SHRP 2 Reliability and related projects for a period of 20 to 50 years.

The research team interviewed the SHRP 2 program director and senior program officers for the Capacity, Reliability, Renewal, and Safety programs to establish the overall goals for the system, gather business requirements, and understand organizational and governance issues. The team also interviewed all active Reliability project contractors and the contractors for the related SHRP 2 Capacity Projects C04 (Improving Our Understanding of How Highway Congestion and Pricing Affect Travel Demand) and C05 (Understanding the Contribution of Operations, Technology, and Design to Meeting Highway Capacity Needs) to understand the nature of the data that would eventually be archived and to gauge the contractors' preparedness for organizing and providing that data for archiving. Finally, the research team met with the general counsel of the National Academies to understand any legal or institutional issues with respect to data rights that might impact the feasibility of the Reliability Archive.

On a parallel track, the team conducted a literature survey to identify available and emerging technologies that might be applicable to the archive. The team also researched

similar systems that have been deployed in government and industry and issues and best practices that have been identified in these systems.

From this foundational work, the research team developed a vision for the Reliability Archive that captured key high-level goals. Once the SHRP 2 program management validated them, the high-level goals provided guiding principles for the development of a conceptual design and a detailed set of requirements for the Reliability Archive.

Once agreement on the vision, concept, and requirements was established, the research team enumerated desirable user interfaces for the Reliability Archive based on the likely users of the system and their business requirements. The research team also analyzed a range of technical and administrative issues that needed to be considered before alternative solutions could be explored. The team then developed three alternative solution approaches that could satisfy the requirements specified through the research project.

A life-cycle cost model was developed to identify all of the initial and recurring costs likely to be incurred over the service life of the Reliability Archive, and a 25-year life-cycle cost analysis was performed for each of the three solution alternatives. The benefits to major stakeholders and the technical and business risks of each solution alternative were also analyzed. The output of this cost/benefit/risk analysis was a final recommendation on the feasibility of deploying the proposed Reliability Archive.

Findings

Input from the meetings with SHRP 2 stakeholders and Reliability project contractors led the research team to conclude that a single, conventional relational database system would not be adequate to build the Reliability Archive and achieve the goals of preserving data and enabling access to its contents by transportation practitioners and researchers.

The following were observed:

- Among projects that are under way, diverse file types are being collected and produced that embody the intellectual product of each research project.
- Some projects are purchasing and/or collecting structured data sets (databases) and, in some cases, aggregating such data, any of which may form the basis for analytical models and resultant predictions and conclusions.
- The nature of the structured data sets varies from project to project. Incident; extraordinary events; roadway information; and volume, occupancy, and speed (VOS) are some of the data set types that exist. The formats of these data sets also vary. Some are in "flat" binary files, while others are in various database formats, some of which are proprietary to particular software vendors.
- There is variability among structured data sets containing the same kind of information, even within a single project. For example, some VOS data may be purchased from a vendor such as INRIX, while other VOS data may be collected from a state DOT. Subtle yet significant differences between them might exist.
- There are some data, methodology, and outcome dependencies among projects. For example, SHRP 2 Reliability Project L05 (Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes) will build on the statistical relationships between countermeasures and reliability performance measures developed in Reliability Project L03 (Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies). The contractor will develop corridor- and network-level strategies using countermeasures and strategies from Reliability Projects L03, L07 (Evaluation of Cost-Effectiveness of Highway Design Features), and L11 (Evaluating Alternative Operations Strategies to Improve Travel Time Reliability); integrated business processes identified in Reliability Project L01 (Integrating Business Processes to Improve Reliability); model results from Reliability Project L04

(Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools); and information from other sources. Any relationships or linkages will exist at the conceptual or knowledge level, not only at the data item level.
- The archival system will need to preserve raw data as well as the data and conclusions derived from it. How raw data lead to conclusions is a combination of data, methodology, and the conclusions per se.
- Some data (particularly raw data) that will be archived have specific rights and restrictions governing access. A fairly granular access control mechanism may be required.
- Without exception, all of the projects expect to produce a range of document-centric, or semi-structured files, including reports and presentations in various formats. The work product of some projects will consist entirely of documents.
- At the time of this writing, some projects in the SHRP 2 program are just starting and others are being planned. It is impossible to know what all of the data from all of the SHRP 2 projects will look like by the time this report is published.

These observations led the research team to conclude that the Reliability Archive cannot be thought of as a database, which presupposes that all aspects of the structure be known up front. Rather, a much more flexible, generalized approach is needed. Thus, the research team focused on a vision of an "active" archive system that could serve as a repository capable of managing files and metadata from different content sources.

Since the purpose of the archive is to preserve a diverse but related collection of digital artifacts and to make them accessible to practitioners and subsequent generations of researchers, the research team proposed that the conceptual design pattern for the archival system follow that of a digital library or museum. Libraries and museums focus on preserving information, maintaining the provenance of information, and putting information into context. These essential curatorial principles reflect what the Reliability Archive needs to support in the digital realm.

The research team proceeded to identify three main goals for the Reliability Archive:

- Preserve, for up to 50 years, all of the valuable digital assets collected and produced by SHRP 2 Reliability focus area research projects.
- Provide transportation researchers and practitioners with a way to discover and then access these digital assets in standard, open formats.
- Establish an extensible architecture that facilitates future expansion of the archival system to
  - Preserve digital assets from other projects;
  - Enhance discovery by integrating related data, e.g., for data visualization;
  - Provide data integration or mashup services; and
  - Create a collaborative community.

This framework led to a conceptual design for the system as depicted in Figure ES.1.

This conceptual design is based on standards and best practices in library and archival science as applied to the digital realm. The concept embodies the life-cycle management of digital content from submission by the project research team, to further classification and contextualization by the institutional staff, to ingestion into the repository for long-term preservation, and, ultimately, to access by transportation researchers and practitioners. A key concept of the design is a flexible mechanism for encoding metadata that describes and links content, and encodes policies such as access and ownership rights.

From this conceptual design, detailed system requirements were created and three solution alternatives were formulated based upon them. Each alternative was evaluated with respect to how well it met the system requirements and conformed to the concep-

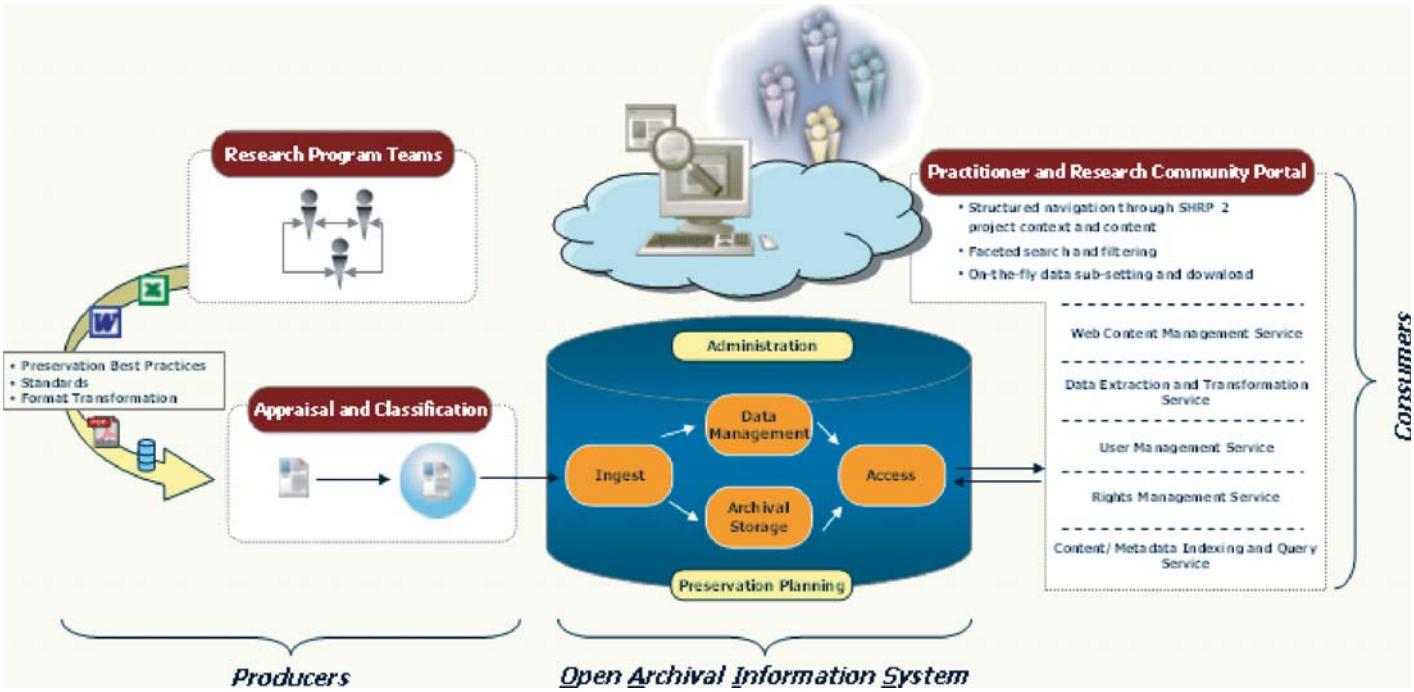


Figure ES.1. Conceptual design.

tual design. Initial and life-cycle costs, benefits to stakeholders, risk mitigation, and schedule were also assessed.

The most compelling alternative was based on commercially available digital object repository management software designed for universities, libraries, museums, archives, and information centers. The functionality provided by these software suites maps very closely to the functional requirements and conceptual design of the archival system. These systems enable institutions to manage digital entities end-to-end, from submission to access, while ensuring their integrity over time through continuous preservation actions.

Another ingredient of the recommended alternative is the use of so-called cloud storage, which is based upon a pay-as-you-go, web-based access model. The use of this kind of utility computing service lowers up-front costs, eliminates periodic storage hardware upgrades, and simplifies system management.

The initial cost of this alternative is estimated to be about 37% below the $1.2 million budget threshold that SHRP 2 has stipulated for the development and implementation of the Reliability Archive. The other alternatives are about 15% above this budget limit. This alternative also has the lowest project recurring costs because the use of a storage service takes advantage of economies of scale while avoiding system maintenance costs and periodic hardware upgrade costs for storage during the system's life.

Conclusions

The research team concludes that it is highly feasible for the SHRP 2 program to cost-effectively deploy a data archival system that meets all of the goals and objectives envisioned by its major stakeholders.

Recommendations

The research team recommends that the SHRP 2 program proceed with the L13A Reliability Archive project as planned, following the approach described in this report.

CHAPTER 1

Background

The goal of the SHRP 2 Reliability focus area is to reduce road congestion through incident reduction, management, response, and mitigation. To achieve this goal, it has sponsored more than a dozen research projects that focus on travel time reliability and operations strategies. The nature of reliability research dictates that many such projects deal with large amounts of traffic data. This data is often aggregated and analyzed in various ways to produce research products that include data and methods to support decision making, guidance on institutional change needed to support agencies' increased focus on operations, and analyses of the effectiveness of highway designs and operational countermeasures to support incorporation of reliability into planning, programming, and design manuals and procedures.

The implementation of these research products can have great societal benefits. The data and analytical approach behind the research products can also be of significant long-term value if researchers and practitioners can build new research and products based on this foundation. This requires that the base data, derived data, analytical models and tools, and other information that constitute the intellectual framework of the SHRP 2 Reliability research projects be preserved and made accessible.

Accordingly, SHRP 2 sponsored the subject of this report to assess the technical, economic, and business aspects of developing, operating, and maintaining a long-lived archival system that preserves and makes readily available to researchers and practitioners the data from SHRP 2 Reliability and other related projects for a period of 20 to 50 years.

The need for preservation planning well in advance of the actual need for preservation is illustrated by lessons learned during the first SHRP program. Today, 13 of the 14 major databases from that program are no longer accessible. This loss of knowledge, which was gained at significant expense, was clearly a motivation for providing funding for this feasibility study and, contingent on a convincing outcome from the present research, for the planned follow-on Reliability Project L13A, which budgets $1.2 million for the actual development, deployment, and operation of the Reliability Archive.

This budgetary figure established a financial constraint against which projected costs were evaluated. Another constraint is the 18-month implementation timeline expected for Reliability Project L13A. Any alternative would have to be implementable within this time frame.

CHAPTER 2

Research Approach

To establish the feasibility of developing, operating, and maintaining a long-lived Reliability Archive, the research team assessed the technical, economic, and business aspects of the proposed system.

The project began with two parallel tasks. The research team interviewed the SHRP 2 program director and senior program officers for the Capacity, Reliability, Renewal, and Safety programs in order to establish the overall goals for the system, gather business requirements, and understand organizational and governance issues. The team also interviewed all active Reliability project contractors, plus the contractors for related Capacity projects C04 and C05, to understand the nature of the data that would eventually be archived and to gauge the contractors' preparedness for organizing and providing that data for archiving. Finally, the research team met with the general counsel of the National Academies in order to understand any legal or institutional issues with respect to data rights that might impact the feasibility of the Reliability Archive.

On a separate track, the team conducted a literature survey to identify available and emerging technologies that might be applicable to the archive, find where similar systems have been deployed in government and industry, and identify issues and best practices.

Based on this foundational work, the research team developed a vision for the Reliability Archive that captured key high-level goals. In turn, these provided guiding principles for the development of a conceptual design and a detailed set of requirements for the Reliability Archive.

Once agreement on vision, concept, and requirements was established, the research team enumerated desirable user interfaces for the Reliability Archive based on the likely users of the system and their business requirements. The team also analyzed a range of technical and administrative issues that needed to be considered before alternative solutions could be explored.

The research team then developed three alternative solution approaches that could satisfy the requirements developed through the research project. The team strove to be consistent with the spirit of the L13 project request for proposal (RFP), considered a broad range of concepts—outsourcing and/or hosting of data outside the National Academies was mentioned specifically as a concept to consider—and developed at least one alternative that was simple and straightforward while ensuring that all the alternatives proposed were practical. The team also took into account the practicality of various technical implementation options, considering the institutional framework under which the Reliability Archive would be deployed and managed.

A life-cycle cost model was developed to identify all of the initial and recurring costs likely to be incurred over the service life of the Reliability Archive. A storage capacity sizing model was also developed for the Reliability Archive, inasmuch as data storage is a significant contributor to cost in any archive. Using these models, a 25-year life-cycle cost analysis was performed for each of the three solution alternatives. The benefits to major stakeholders as well as the technical and business risks of each solution alternative were analyzed. The output of this cost/benefit/risk analysis was a final recommendation on the feasibility of the proposed Reliability Archive.

CHAPTER 3

Findings

SHRP 2 Management Perspective

The research team interviewed the SHRP 2 program director and senior program officers for the Capacity, Reliability, Renewal, and Safety program areas early in the project in order to understand their perspective on the issues.

Goal and Targeted Audience for the Reliability Archive

The primary objective of the Reliability Archive is to allow users to validate the research results from relevant SHRP 2 projects and to refine and build on research results in the future. Currently, the archive is mainly targeted to serve transportation researchers, such as university professors, transportation engineers, and planners. Thus, its expected immediate and long-term benefit is to give these researchers access to the data in the archive so that they can reproduce research results or build new research on the data.

The feasibility of the archive hinges largely on the actual size of the downstream user base, which is a key issue in justifying long-term preservation of a research database. While everyone agrees that there is a definitive need for an archive of research data, it is difficult to identify or guess the size of the future user base.

The archive could also be used to support other transportation communities in the future. For example, state DOTs can use data in the archive to augment their own traffic data collection programs. State DOTs currently spend vast resources to collect traffic data. A national system like the Reliability Archive that can provide state DOTs similar data could reduce state-level data collection efforts and costs.

The focus of the archive should be on data rather than on documents, although associated documentation will also need to be archived in conjunction with the data. It needs to be further clarified whether or not all documentation type project deliverables, such as reports and presentations, will be archived for the entire SHRP 2 program.

From the "data perspective," the Reliability Archive may need to focus on data sets that are more oriented toward future researchers rather than toward practitioners whose data needs tend to change frequently.

User Access to the Reliability Archive

One of the major concerns of SHRP 2 senior management is how easily users will be able to access the data in the archive. Under the first SHRP program, 14 databases were built to capture a wide range of data. At the time of this writing, 13 of these 14 databases are no longer accessible. Part of the reason is that data was saved in old formats that cannot be easily made available with today's technology. Data frequently becomes inaccessible after being collected because of a number of reasons, including the following:

- Technology obsolescence;
- High costs of maintaining and managing the data; and
- Availability of newer and better data.

Initial ideas were exchanged during the meetings with respect to making the archived data more widely accessible and more easily available to those who need it. One approach is an on-the-fly extraction and transformation service as a function available through a portal-type user interface. The following specific examples were discussed:

- Each program under the SHRP 2 has established a business framework that describes its program vision and underlying business concepts and principles that guide individual projects. These frameworks could be used as a business context or portal to construct future user interfaces for the Reliability Archive. For example, the Capacity Program's Collaborative Decision Framework includes 50 decision

points intended to work for all states. It basically rewrites the entire transportation planning process. Each decision point can have its own archive.
- Another approach is to provide users with more contextual information about data sets. Commercial products are emerging that are designed to allow users to search, display, and consume information without having to know anything about how and where it is structured and stored.

Views on the Data to Be Archived

The primary objective of the Reliability Archive will be to preserve research project data. However, archiving data is much more complicated than simply saving data in a file system or a conventional database. The research team's interviews yielded the following views on the data to be archived in the system:

- The interviewees agreed that conclusions as well as data need to be archived; these conclusions are in the form of research products based on the data.
- Projects relevant to the archive will typically involve either collecting data or analyzing or mining data. As such, the data expected to be archived will include base or raw data that represent the original information collected as well as research or subject data that are outcomes of analyses.
- Sometimes the base or raw data sets may be proprietary and may not be used outside the associated projects; only institutional staff may have access to this type of information.
- Data to be archived will either be collected by the project contractors or purchased from vendors such as INRIX. For example, arterial data is hard to collect and most probably will be purchased from vendors.
- The Federal Highway Administration (FHWA) is expected to have nonexclusive rights to subjective data. The original producers may retain commercial rights.
- Metadata is critical because people often call two different things the same name; hence, it is necessary to define metadata standards. It is therefore desirable to have a data dictionary and other means to describe the data to the researchers.
- The use of XML to represent metadata was discussed as a promising way to describe data consistently.

Project Contractor Perspective

The research team also interviewed the contractors of active Reliability projects and relevant Capacity projects (C04 and C05) to understand the data used and produced by those projects that would need to be archived. Examples of such data include travel time and speed used for modeling and simulation, travel- or highway-related data collected during a project, analytical models developed by a project, and reports and presentations produced as a result of a project.

Appendix A summarizes the key characteristics of reliability and other SHRP 2 projects that are relevant to the Reliability Archive. The characteristics are organized according to the following categories:

- Raw Data: Almost every project listed in Appendix A is required to collect or obtain some form of raw data to support its analysis. In order for future researchers to validate the results of these projects, it is desirable that the raw data be retained in the archive. Two aspects of the raw data, namely, Data Sources and Data Rights, are examined and summarized in Appendix A.
- Research Outcome: There are a diverse range of outputs produced from these SHRP 2 projects. Each type of output is an integral part of the entire set of outcomes toward building an enriched knowledge repository. The following are typical types of outputs:
  - Derived Data: This includes data derived from any kind of analysis on the raw data via mathematical models, modeling and simulation, and computer programs. Such derived data may be saved in a variety of formats such as spreadsheets, text files, and databases.
  - Models: This includes mathematical models and formulas, simulation models, business process flows, strategies, methodologies, and analytical frameworks. They can be in text, graphics, and equation formats.
  - Tools: This refers to small to medium types of computer-based tools or applications developed by the contractors to support their analysis. These tools will typically be tied to an application development environment such as SAS, Excel, or Access.
  - Code: This refers to any software code that is developed and used by the contractors to derive their outcomes and conclusions. An example would be a JAVA or C++ program. The programs will typically be saved in both a source file format and an executable format.
  - Reports: This includes written reports, technical memoranda, presentations, and training materials, most of which are expected to be in Microsoft Word, Excel, PowerPoint, and PDF formats.
- Reliance on Other Projects: This captures the data dependencies among projects, which must be retained so that future users can have a complete and thorough understanding of the outcomes produced from these projects.
- Metadata: This summarizes what metadata standards, if any, are planned to be used in each project.

Figure 3.1 presents an overall project timeline based on information available from the SHRP 2 program. The projects are organized into the following categories:

- Reliability Archive-related projects;
- Active projects;

C	ID	Project Name	Duration	Start	Finish
#	L13	Requirements and Feasibility of A System for Archiving and Disseminating Data from SHR 2 Reliability and Related Studies	18 mons	Sep 2008	Mar 2010
#	L13(A)	Design and Implement the L13 Data Archive System	12 mons	Sep 2010	Aug 2011
#	L16	Assistance to Contractors to Archive Their Data for Reliability Projects	30 mons	Mar 2010	Sep 2012
#		**Active Projects**	**65.2 mons**	**Feb 2007**	**June 2012**
#	L01	Integrating Business Processes to Improve Reliability	18 mons	Feb 2008	Aug 2009
#	L02	Establishing Monitoring Programs for Travel Time Reliability	36 mons	Feb 2009	Jan 2012
#	L03	Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies	24 mons	Feb 2007	Jan 2009
#	L04	Incorporating Reliability Performance Measures in Planning and Operational Modeling Tools	36 mons	Mar 2009	Feb 2012
#	L06	Institutional Architecture to Advance Operational Strategies	42 mons	Feb 2007	Aug 2010
#	L07	Evaluation of Costs and Effectiveness of Highway Design Features to Improve Travel Time Reliability	48 mons	July 2008	June 2012
#	L10	Feasibility of Using In-Vehicle Video Data to Explore How to Modify Driver Behavior that Causes Non-recurring Congestion	36 mons	Mar 2009	Feb 2012
#	L11	Evaluating Alternative Operations Strategies to Improve Travel Time Reliability	18 mons	Sep 2008	Feb 2010
#	L12	Training and Certification of Traffic Incident Responders	27 mons	Mar 2008	June 2010
#	L14	Effectiveness of Different Approaches to Disseminate Traveler Information on Travel Time Reliability	24 mons	Sep 2009	Aug 2011
#	C04	Improve Our Understanding of How Highway Congestion and Pricing Affect Travel Demand	28 mons	Sep 2007	Jan 2010
#	C05	Understanding the Contribution of Operations, Technology, and Design to Meeting Highway Capacity Needs	25 mons	Jan 2008	Jan 2010
#		**Pending Contracts (Estimated to Start in Early 2009)**	**24 mons**	**Jan 2010**	**Dec 2011**
#	L05	Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes	24 mons	Jan 2010	Dec 2011
#		**2009 Planned Projects (Estimated to Start in Late 2009 or Early 2010)**	**30.03 mons**	**Jan 2010**	**June 2012**
#	L08	Incoprating Non-recurrent Congestion Factors into the Highway Capacity Manual Methods	24 mons	July 2010	June 2012

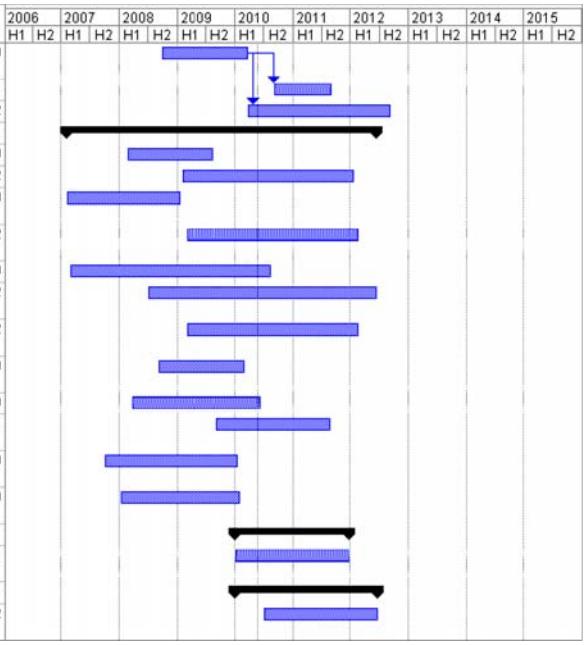


Figure 3.1. Project timelines.

- Pending projects; and
- Future planned projects.

The following conclusions were drawn from this project timeline:

- Most reliability projects are already under way. At the time this report was written, only four more projects have yet to be started. These include one pending project (L05) and three planned projects.
  - One of these four projects will become active in early 2010.
  - The three remaining projects (L08, L09, and L15) will become active in 2010.
- All projects will be completed by the end of 2012.
  - Two projects were completed by the end of 2009.
  - Six projects will be completed by the end of 2010.
  - Two projects will be completed by the end of 2011.
  - Six projects will be completed by the end of 2012.

The implementation of the Reliability Archive will be completed by August 2011, assuming it will start within 12 months of the completion of the feasibility study. By the time the Reliability Archive is deployed, two-thirds of the projects will be completed. Data from these projects should be available to be moved into the Reliability Archive.

Currently, no single metadata standard is applied across all these projects. It appears that each project develops its own approach to data collection and organization. The research team believes it is highly desirable that the efforts of assisting project contractors to prepare their data for archiving (which is the intent of planned Reliability Project L16, now a part of Reliability Project L13A) be started earlier, preferably early in 2010, well before a number of projects are expected to be completed.

Literature Research

The findings from the meetings with SHRP 2 stakeholders and Reliability project contractors reinforced the research team's initial impression that a single, conventional relational database system would not be adequate to build the Reliability Archive. Thus, the research team started to focus on a vision of an active archival system that could serve as a repository capable of managing files and metadata from different content sources.

Digital Archiving Technology

A survey of the literature in the public domain reveals that the issue of archiving digital resources has been discussed actively for the last decade and that the intensity of the discussion and corresponding volume of research, opinions, experience, and technologies pertaining to the subject have grown dramatically in the last few years.

It is not hard to understand why this has happened when one considers the explosive growth of digital information that has occurred, which shows no sign of abating. One industry analyst forecast (*1*) projected a compound annual growth rate (CAGR) of digital content of 57% worldwide between 2006 and 2010. As organizations accumulate vast amounts of digital content, they are becoming increasingly aware that it is

actually much easier to lose digital information than it is to lose traditional paper records.

The Storage Network Industry Association's (SNIA) *100 Year Archive Requirements Survey* (*2*) identified this so-called "Digital Crisis," which involves the risk of losing digital information over time because one

- Cannot read it;
- Cannot interpret it correctly;
- Cannot validate its authenticity; or
- Cannot find it.

The executive summary of the same report also describes succinctly the two grand technical challenges of logical and physical migration that must be dealt with in an archival system:

> Logical migration is the practice of updating the format of the information into a newer format that can be read and properly interpreted by future applications or readers without losing the authenticity of the original. Physical migration means to copy the information to newer storage media to preserve the ability to access it and to protect it from media corruption. Best practices today require logical and physical migration every 3–5 years. (*2*)

Until recently, a digital archive was generally thought of as a library of tapes containing backups. It is now more widely understood that storage backup and disaster recovery technologies designed for operational continuity do not address the issue of long-term data preservation. *Backups* take a snapshot of information at a given point in time that may be restored as quickly as possible. They are a short-term data recovery solution after data loss or corruption. *Digital archiving,* on the other hand, preserves the authentic digital document of record for a specified period of time (or even indefinitely) to keep it accessible even as technology advances.

Preservation is thus a primary focus of a digital archive, but the preservation of digital records presents numerous challenges because, unlike paper or microfilm, digital information can easily be corrupted, disseminated, copied, or altered beyond recognition. Also, the hardware and software needed to access digital records change rapidly, and storage media such as tapes and discs can deteriorate quickly even if they do not appear to be damaged. Finally, the context of a digital record and its relation to other records can easily be lost.

The sheer volume and the volatility introduced by digital content impose a new set of requirements capable of scaling and of preventing accidental changes to the records. Procedures need to be put in place to identify, classify, move, evolve, access, and occasionally dispose of digital records. Both library science and traditional archival practice provide an extensive body of knowledge that is being leveraged with technology to provide solutions to this digital preservation dilemma.

Research and Standards Initiatives in Archiving

Research and innovation around digital preservation are occurring in both the public and private sectors. Organizations with a vested interest in the subject, such as national libraries and archive agencies of many countries, museums and libraries, major research universities, standards organizations, and industry associations are active in the field. Information technology companies now have second- and third-generation hardware and software products on the market.

The most influential standards initiative for archival systems is the *Reference Model for an Open Archival Information System* (OAIS) (*3*), which has been adopted as an International Organization for Standardization (ISO) standard (ISO 14721) that identifies the processes required for long-term preservation within an archival repository and establishes a common framework of terms and concepts.

The OAIS model was developed by the Consultative Committee for Space Data Systems (CCSDS) in response to a need for standards in support of the long-term preservation of digital information obtained from observations of terrestrial and space environments. The research team believes that it is relevant to the SHRP 2 program because of the following reasons:

- It was developed by agencies confronted with very long-term and very large-scale data preservation problems.
- It emphasizes long-term access in addition to long-term data preservation. Governmental agencies, museums, libraries, and other institutions have an inherent understanding of the long-term value of the content they are archiving and its importance to future researchers. (In contrast, many initial deployments of archiving technology in commercial enterprises were driven by compliance mandates, which created a "fix it fast" mentality to put in place a system that allowed the organization to prove it was meeting its legal obligations to retain information.)
- It is widely accepted as a reference model that provides a common vocabulary useful for framing requirements and in assessing implementation and operational feasibility.

One of the most important concepts of OAIS is the idea of an information package, which is essentially a container (or object) that encapsulates both the archived data itself (content information) and the various categories of metadata that describe the data, its relationships to other data, and other descriptive information. Figure 3.2 shows the information package concept. Managing content and context in this self-contained and self-describing way makes it easier to apply the traditional archivist's governing principles, such as provenance, in the digital realm.

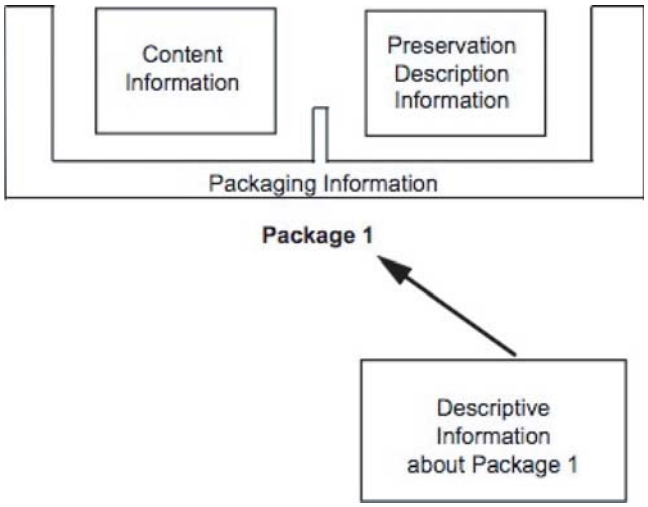


Figure 3.2. OAIS information package concept.

The OAIS model also defines the major entities and functions of a digital repository constructed to maintain safe, long-term custody of digital objects, as illustrated in Figure 3.3. The major functions are as follows:

- Ingest: Accepting digital objects into the archive;
- Archival storage: Storing, managing, and retrieving objects; managing the storage hierarchy; and refreshing the media on which the objects are stored;
- Data management: Writing, reading, and updating both administrative data and descriptive metadata;
- Administration: Managing of the overall operation of the archive;
- Preservation planning: Managing the logical and physical integrity of the archive over time; and
- Access: Locating, applying access controls, and generating responses to requests for archived objects.

The items labeled SIP (Submission Information Packages), AIP (Archival Information Packages), and DIP (Dissemination Information Packages) in Figure 3.3 refer to the different kinds of self-contained, self-describing information packages that might exist in a repository.

Producers submit SIPs to a repository. In the Reliability Archive, the producers will be the research teams conducting the various research projects in this focus area of the SHRP 2 program. AIPs will be managed by the Reliability Archive for the duration of their valuable life cycle. DIPs are retrieved when transportation researchers and practitioners access the Reliability Archive.

It is important to note that the OAIS model is a conceptual framework that does not prescribe any specific implementation at any level. The OAIS model defines what is needed for a modern digital archive but not how to build it. It is the conceptual foundation for many important digital preservation initiatives as well as many archival products.

Role and Importance of Metadata

Metadata literally means data about data; its importance in a long-lived digital archive cannot be overemphasized. Metadata is important for context, description, and discovery, and it also encodes policies related to administration, accessioning, preservation, and use of information.

Over time, many standards have been developed to represent different categories of metadata for specific object types. The transportation sector is no exception. However, there is no catch-all standard that accommodates every type of digital object. The research team expects objects managed by the Reliability Archive to be tagged with different kinds of descriptive and technical metadata as appropriate for their content.

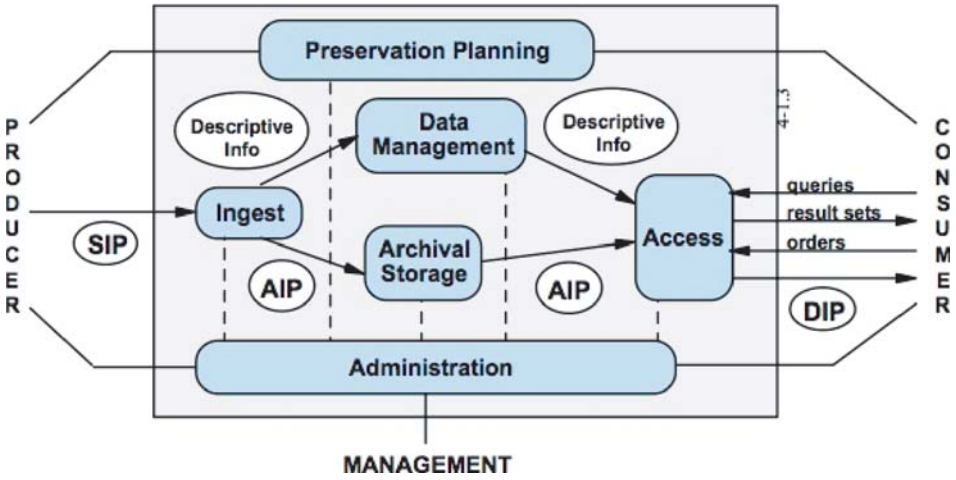


Figure 3.3. OAIS functional entities.

In addition to metadata specific to particular object types, all digital objects require different levels and types of metadata at different points in their life cycle; all of this diverse metadata needs to be associated or packaged with the object it describes.

The Metadata Encoding and Transmission Standard (METS) (*4*) was developed to deal with these issues. METS was developed by the Digital Library Foundation and is supported by the Library of Congress as its maintenance agency.

Metadata Encoding and Transmission Standard

METS was designed as an overall framework within which all the metadata associated with a single digital object can be stored or referred. METS is an Extensible Markup Language (XML) schema that provides a mechanism for recording the various relationships that exist between pieces of content and between the content and the metadata that make up a digital object. It enables effective management of digital objects within a repository, acts as a standard for transferring metadata within repositories, facilitates access and navigation by the researcher, and links the digital object and its metadata inextricably together.

METS was specifically designed to act as an OAIS information package. It can deal with all categories of metadata cited by OAIS (content, preservation, packaging, and descriptive metadata). Packaging all of this metadata with the digital object it describes ensures that the object is self-documenting over time.

An Aside About XML

XML is sometimes called a metalanguage, which means that it can describe other languages. It is extensible because the markup elements are user-defined. XML and HTML are sometimes confused. HTML was designed to specify how data are presented, whereas XML was designed to transport and store data and says nothing about their presentation.

Figure 3.4 shows an excerpt from an XML document, which is one row (record) from a table exported from the Long-Term Pavement Performance (LTPP) database as it might be encoded in XML format. The structure and data are quite easy to deduce since XML is eye-readable (and machine-readable, too). The tags in the example (like <SHRP_ID> and <STATE_CODE>) delimit the data contained in the document. (Each tag has a corresponding closing tag starting with a slash, e.g., </SHRP_ID>.)

XML is considered a robust archival format and is readily interchangeable because it uses standard ASCII code rather than a binary format to encode data. Because it is neutral and flexible, it has become a de facto standard way to express both data and metadata in a structured manner.

```
<INV_LAYER>
    <SHRP_ID>0500</SHRP_ID>
    <STATE_CODE>1</STATE_CODE>
    <CONSTRUCTION_NO>1</CONSTRUCTION_NO>
    <LAYER_NO>1</LAYER_NO>
    <DESCRIPTION>7</DESCRIPTION>
    <MATERIAL_TYPE>52</MATERIAL_TYPE>
    <LAYER_TYPE>G</LAYER_TYPE>
    <RECORD_STATUS>E</RECORD_STATUS>
</INV_LAYER>
```

Figure 3.4. XML encoding example.

METS Document Structure

The following section provides an overview of the structure of a METS document and its application.

As Figure 3.5 illustrates, a METS document consists of seven major sections:

1. METS Header: Contains metadata describing the METS document itself, including such information as creator and editor.
2. Descriptive Metadata: May point to descriptive metadata external to the METS document, or contain internally embedded descriptive metadata, or both. Multiple instances of both external and internal descriptive metadata may be included in the descriptive metadata section.
3. Administrative Metadata: Provides information regarding how the files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information

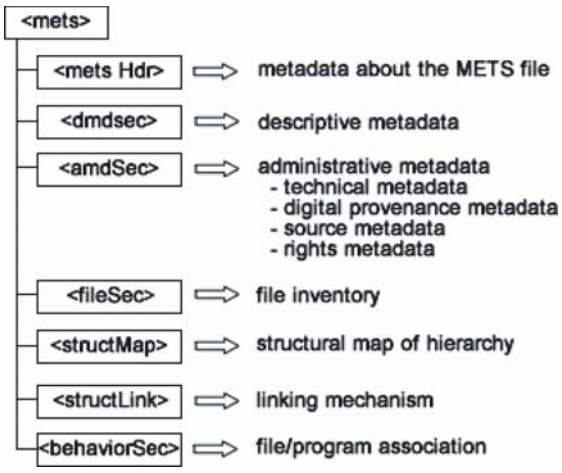


Figure 3.5. METS document structure.

regarding the provenance of the files constituting the digital library object (i.e., master/derivative file relationships, and migration/transformation information). As with descriptive metadata, administrative metadata may be either external to the METS document or encoded internally.
4. File Section: Lists all files with content that constitutes the electronic versions of the digital object. Files may be grouped by object version. METS can deal with simple digital objects containing a single file and with complex objects composed of many files.
5. Structural Map: Outlines a hierarchical structure for the digital library object, and links the elements of that structure to content files and metadata that pertain to each element. METS can express the hierarchical structure common to digital objects, which may have been created originally in multiple directories and folders.
6. Structural Links: records the existence of hyperlinks between nodes in the hierarchy outlined in the Structural Map.
7. Behavior: can be used to associate executable behaviors with content in the METS object, including the ability to identify a module of executable code that implements and runs the behaviors defined abstractly by the interface definition.

An advantage of METS is that it does not dictate the content of metadata. Other metadata schemas can be incorporated into a METS file or referred to from it. It is also extensible inasmuch as new versions of metadata may be incorporated alongside older versions of metadata.

Because it was conceived as a framework for packaging disparate metadata, METS has strong advantages in this area. The research team believes that such a framework is needed. Any system capable of handling XML documents can be used to create, store, and deliver a METS file, thereby mitigating problems of software obsolescence. METS offers strong capabilities, flexibility, and extensibility. Finally, it has strong worldwide adoption, particularly in preservation repositories.

Other Descriptive Metadata Sources

In the context of OAIS and a METS "wrapper," descriptive metadata is and can be thought of as being associated with a digital object. There are two other kinds of descriptive metadata that can be gleaned from digital objects and used to find items of interest:

- Embedded metadata is descriptive metadata contained within the file itself. A familiar example is the information displayed when a Windows user right-clicks a file and chooses Properties. In addition to file system metadata such as size or date created, many document formats embed author, subject, and keyword information. Image files typically embed metadata related to height, width, color depth, and so forth. This embedded metadata can be very useful in helping future users find and filter content of interest.
- Derived metadata is information that can be gleaned about a file through content inspection. This can be simple keyword indexing or it can involve much more sophisticated techniques that can identify entities, relationships, and contextual linkages. The advantages of derived metadata as a finding aid in a long-lived archive are numerous. It is impossible to anticipate all the ways that future researchers will wish to access information. Derived metadata is not dependent on fixed metadata schema that effectively requires that all classification be done up-front. It can be extracted at any time. Moreover, it is reasonable to assume that more and more sophisticated information extraction technologies will become available over the life of the archival system.

Embedded and derived descriptive metadata can be used, along with metadata stored or referred to in the object's wrapper, to facilitate user access to information in the archival system.

Vision for the Archival System

In the research team's experience, any successful requirements analysis and feasibility study must be guided by a clear and accurate vision. This project is no exception.

The initial vision for the proposed archival system was set in the RFP for the L13 project. This RFP made it clear that long-term preservation of Reliability focus area project data is important ("in the range of 20 to 50 years") and that the reasons for preserving these data are (1) "to allow others to validate the research results," and (2) to make the data "available for researchers in the future to refine and build on the research results."

In the course of this research, the team talked to a wide range of stakeholders and has come to a more expansive vision of the archival system. From the research, the team has identified three main goals for the Reliability Archive:

- Preserve for up to 50 years all of the valuable digital assets collected and produced by SHRP 2 Reliability focus area research projects.
- Provide transportation researchers and practitioners with a way to discover and then access these digital assets in standard, open formats.
- Establish an extensible architecture that facilitates future expansion of the archival system to:
  - Preserve digital assets from other projects;
  - Enhance discovery by integrating related data (e.g., for data visualization);
  - Provide data integration or mashup services; and
  - Create a collaborative community.

The first goal expresses that all of the digital information being purchased, collected, aggregated, analyzed, and produced across all of the Reliability focus area projects is a potentially valuable asset that must be protected for the long-term.

The second goal captures the understanding that the prospective users of this digital information are both practitioners and researchers, and that they have certain expectations, shaped by the Internet and other experiences, of how users should be able to find and exploit information in the system.

The third goal builds on the second to consider how broader connections might be made. At the data level, it is clear that data from other focus areas either relate to the Reliability focus area or are similarly valuable and perhaps should be preserved in the archival system. It seems likely that users will want to connect data from the archival system to other external services or with other external data. Finally, it is natural that users might want to connect with each other to learn from each other and share their experiences in using the information in the archival system, and that the system itself might facilitate these connections.

Conceptual Design for the Archival System

After interviewing project teams across the Reliability and other focus areas of SHRP 2, the research team observed the following:

- Among projects that are under way, diverse file types are being collected and produced that embody the intellectual product of each research project.
- Some projects are purchasing or collecting structured data sets (databases) and, in some cases, aggregating such data, any of which may form the basis for analytical models and resultant predictions and conclusions.
- The nature of these structured data sets varies from project to project. Incident, extraordinary events, VOS (volume, occupancy, and speed), and roadway information are some of the data set types that exist. The formats of these data sets also vary. Some are in flat binary files, while others are in various database formats, some of which are proprietary to particular software vendors.
- There is variability among structured data sets containing the same kind of information, even within a single project. For example, some VOS data may be purchased from a vendor such as INRIX, while other VOS data may be collected from a state DOT. Subtle yet significant differences between them might exist.
- There are some data, methodology, and outcome dependencies among projects (e.g., Reliability Project L05 will build on the statistical relationships between countermeasures and reliability performance measures developed in Reliability Project L03). The contractor will develop corridor- and network-level strategies using countermeasures and strategies from Reliability Projects L03, L07, and L11, integrated business processes identified in Reliability Project L01, and model results from Reliability Project L04, as well as information from other sources. Any relationships or linkages will exist at the conceptual or knowledge level, not only at the data item level.
- The archival system will need to preserve raw data as well as the data and conclusions derived from it. How raw data leads to conclusions is a combination of data, methodology, and the conclusions themselves.
- Some data (particularly raw data) that will be archived have specific rights and restrictions governing access. A fairly granular access control mechanism may be required.
- Without exception, all of the projects expect to produce a range of document-centric, or semi-structured, files, including reports and presentations in various formats. The work product of some projects will consist entirely of documents.
- Some SHRP 2 projects are just starting and others are being planned as of this writing. It is impossible to know what all of data from all of the SHRP 2 projects will look like by the time the present feasibility study concludes.

These observations lead to the conclusion that the proposed archival system cannot be thought of as a database, which presupposes that all aspects of the structure be known up front. Rather, a much more flexible, generalized approach seems warranted.

Since the purpose of the archive is to preserve a diverse but related collection of digital artifacts and make them accessible to practitioners and subsequent generations of researchers, the team proposes that the conceptual design pattern for the archival system follow that of a digital library or museum (see Figure 3.6). Among the advantages of this approach is that there is a growing body of standards, software tools, and best practices gaining worldwide adoption, which could be leveraged should the archival system be deployed.

The remainder of this section describes the major elements and functions of the conceptual design, using concepts and terminology from the OAIS reference model.

Producers

The born-digital files that will eventually constitute the digital objects preserved in the archival system originate with the work of the research project teams. It is expected that the teams will play an important initial role in assembling and organizing

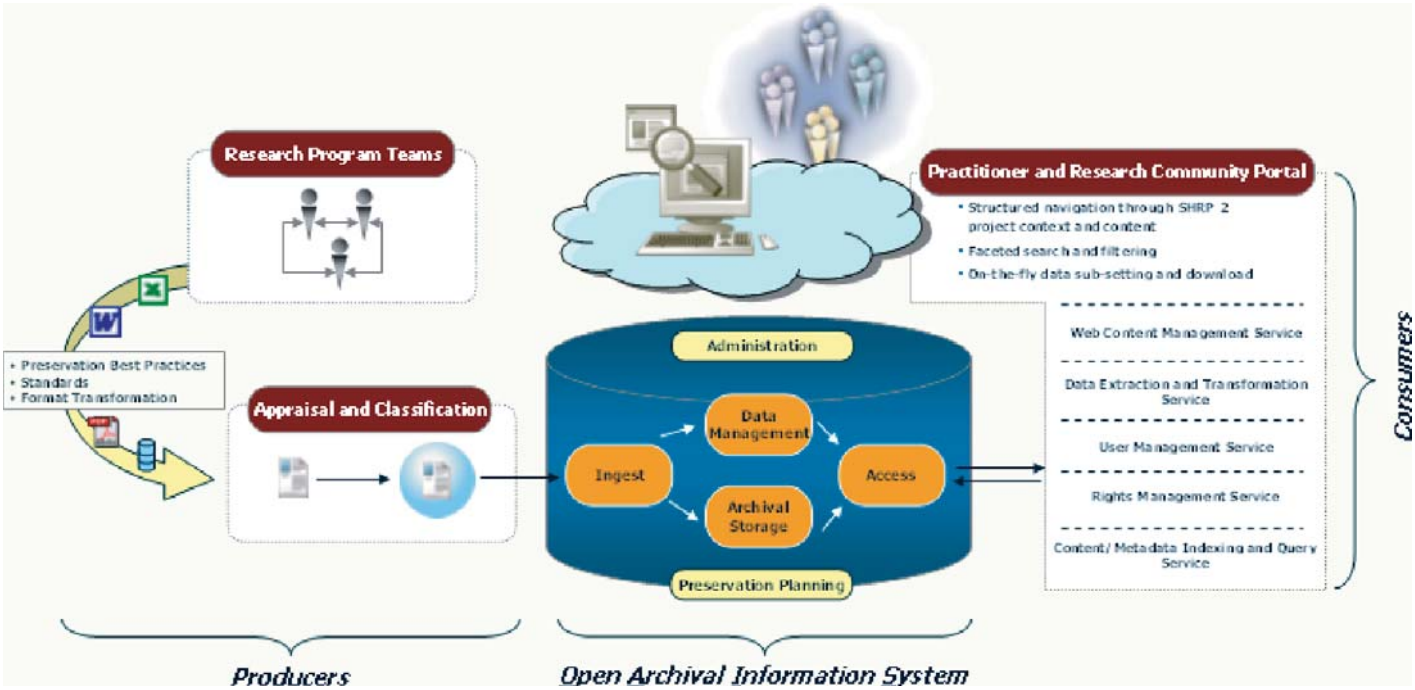


Figure 3.6. Conceptual design.

content for submission to the archive. In OAIS terms, this means the project teams will create initial submission information packages (SIP) for conveyance to the archival system.

While submission for archiving will logically occur upon completion of the project, planning and preparation for archiving will need to occur sooner. This includes selecting, where available, the most preservation-friendly formats for files, and creating basic descriptive metadata. All aspects of copyright, privacy, and proprietary rights must be documented. A Reliability focus area research project is anticipated to assist project teams with such submission-related work. The project teams will need training, standards, best practices, and tools.

Depositing a file or collection would begin by completing a submission agreement and an inventory of the file or collection. It is expected that all aspects of submission could be supported by a web-based application (submission portal) and the SIP could be transported electronically using Internet protocols such as HTTP and FTP.

SIPs would be staged and processed in an accessioning workbench function prior to ingest into the archival system. Tasks performed would be the typical work of an archivist, which includes appraising a submission as worthy of preservation, and cataloging it. This work essentially involves establishing what OAIS describes as preservation description information (PDI), which is often called preservation metadata.

Once the necessary pre-accessioning work has been done, the object or collection would be passed to the ingest function of the archival information system.

Core Archive Functions

The OAIS model defines ingest, data management, archival storage, access, administration, and preservation planning as the six core functions in an archival information system. At a summary level, these functions are responsible collectively for the following:

- Preserving the collection of digital artifacts;
- Monitoring and insuring the integrity of digital artifacts across physical migrations and any format migrations (transformations);
- Maintaining the physical security of digital artifacts;
- Facilitating the discovery of information; and
- Enforcing access control.

Consumers

The primary consumers of the information housed in the archival system are expected to be a worldwide community of transportation practitioners who will use the information directly, as well as researchers who will validate and build on this information base.

It is expected that users will interact with the archival system through a web-based portal. This portal would provide a structured way of navigating through SHRP 2 project content as well as context. A logical way of providing structured navigation is by project.

As researchers become more accustomed to operating in the online realm, they would expect to be able to search for specific items, navigate through information "bottom-up" as well as "top-down," and to follow lateral relationships, including those that may span repositories. At a minimum, users will expect to be able to search by keyword or phrase.

Users will also expect the system to provide faceted search, also called faceted navigation or faceted browsing, which is a technique whereby search results are organized dynamically into categories. A count is often displayed so users can see how many results match each category, or facet. The user can then "drill down" into the search results by category.

The technique is familiar to anyone who has visited an online retailer. Faceted search is also being used extensively in online libraries and is gaining adoption in enterprise search applications. It follows that search and metadata are inextricably linked subjects. Metadata will be discussed in some detail in the following section of this report.

The navigation techniques described above will help users find what they need. Other services will be needed to provide this information to them. Access to information will only be provided to those authorized to see or download it. The entire archive system will be accessed through an authentication and authorization system. Users will be named and be assigned a role of administrator, submitter, or consumer. (There will be more granular levels within each role.)

In terms of access, the two major functions the access portal would provide are the download of single files or file collections, and on-demand data subsetting and download of structured data sets and databases. In concept, the latter would provide what users of the LTPP database have desired, which is a self-service ability to extract and download data subsets of interest.

The access portal would be dependent on a stack of widely available software services available commercially and in open source, including the following:

- Web server;
- Content management;
- Data extraction and transformation;
- User management;
- Rights management;
- Content/metadata indexing and query; and
- Collaboration.

Online Community

Online access to a transportation-related digital library service provides an opportunity to connect the consumers of the information it houses. The same social networking techniques and technologies used across many disciplines and organizations to foster collaboration could be employed in the proposed archival information system. For example, an online message board, or Internet forum, could be a feature of the archival system's online presence. The forum could be organized into top-level categories reflecting the four focus areas of the SHRP 2 program, with top-level folders for every research project in each category.

Forum content itself is entirely driven by its members through their posts and replies. Participants can build connections with each other and groups can form naturally around discussion subjects. Among other things, a forum can help members find answers to questions, share findings and best practices, and identify needs and opportunities for further research. Because of its global reach, leveraging Internet technology in such a way is potentially useful for facilitating international cooperation with other research organizations.

Internet forum software is widely available at little to no cost and the research team recommends considering its role as part of the overall archival information system.

System Requirements

A broad set of requirements was generated from discussions with stakeholders and research into best practices and technical capabilities, both current and expected. These detailed requirements are listed in Appendix C.

These requirements were generated and reviewed as part of the second task of the research project. In a subsequent task the requirements were used to evaluate the solution alternatives that were formulated. The scoring of alternatives versus requirements is also shown as additional columns in the table contained in Appendix C.

For convenience, these requirements are organized into categories and subcategories largely following OAIS nomenclature:

- Producers: Requirements pertain to the preparation and submission of digital artifacts to the archival system.
- Ingestion: Requirements pertain to the acceptance of digital objects into the archive.
- Archival storage: Requirements pertain to storing, managing and retrieving objects, managing the storage hierarchy, and refreshing the media on which the objects are stored.
- Data management: Requirements pertain to writing, reading, and updating both administrative metadata and descriptive metadata.
- Preservation planning: Requirements pertain to managing the logical and physical integrity of the archive over time. In general, these reflect the expression of (setting) policies that are typically enforced by other functions of the archival system. Because there are numerous aspects of preservation, the team uses the subcategories of retention, deletion, replication, logical migration, and backup and recovery to organize these requirements. (Although it is desirable to not

require routine backups of the data in an archival system, under certain circumstances the ability to back up data to tape or other removable media in a standards-based format is useful, such as for data migration and device relocation.)
- Administration: Requirements pertain to managing the overall operation of the archive. The system must operate in a predictable manner, be easily managed, and capable of issuing alerts regarding status and health.
- Access: Requirements pertain to locating, applying access controls, and generating responses to requests for archived objects.
- Consumers: Requirements pertain to facilitating controlled access to information in the archival system.
- Systemwide: General requirements with broad scope.

User Interfaces

The research team analyzed desirable user interfaces for the Reliability Archive based on a review of the system's likely users and their needs. The team focused exclusively on end-user interaction with the proposed archival system (e.g., future researchers and practitioners) and not on administrative interfaces, since the latter will be largely determined by the implementation path that will be recommended.

User Profiles

TRB Special Report 296: Implementing the Results of the Second Strategy Highway Research Program, defines four broad user groups for the Reliability products of the program (*5*). The research team assessed and evaluated these groups' interests, preferences, and desired features and functions with respect to user interfaces for the Reliability Archive, as follows:

1. Leaders of transportation agencies are concerned primarily with strategic issues related to transportation and its role in the economy and society.
   – Primary interests: They would be interested in a small but critical set of products, such as business processes, strategies, institutional structures, and performance measures.
   – Desirable user interface: They need to quickly find the conclusions of each project. They are interested in viewing and downloading business process diagrams, executive summaries, and presentations.
2. The technical staff of transportation agencies is the largest group of potential users of the Reliability products. They are responsible for delivering transportation programs and services to their customers within legal, regulatory, and financial constraints.
   – Primary interests: They would be interested in different sets of the Reliability products, depending on their technical roles. Overall they are interested in applying the end products such as tools, reports, and training programs to their day-to-day responsibilities. They will also be interested in using the raw data sets to complement their own data and integrate the L13 data sets with theirs to develop their own unique tools or products.
   – Desirable user interface: They need to quickly find the end products of individual projects according to their roles and responsibilities. The end products will need to be organized accordingly, such as by the categories of planning, design, and operations. It should be recognized that trying any new technologies, operating strategies, and procedures can be difficult and risky. Thus, this group of users will need to be convinced of the usefulness of the SHRP 2 products. The online community of the Reliability program will be an excellent place for them to learn and share with each other the experience of using SHRP 2 products.

   This user group will also be interested in downloading tools and training programs. Specifically, they will be very interested in downloading raw data sets if they do not have their own jurisdiction-specific data, or they will combine downloaded data with their own data.
3. Nontransportation professionals with some relationship to transportation operations usually have very different scopes of responsibility, such as law enforcement, firefighting, or management of a special event venue.
   – Primary interests: They would be interested in the end products about operational strategies in incident management, travel time reliability improvement, and special event coordination and collaboration.
   – Desirable user interface: They need to quickly find any conclusion, results, and strategies that are related to transportation operations, incident management, and travel reliability improvement. This is the user group that will also be very interested in using the online community to communicate with users from other disciplines.
4. Researchers and analysts are interested in understanding transportation operations and in developing innovative approaches to meet operational challenges.
   – Primary interests: This group of users will be interested in the entire set of Reliability programs. In particular, they want to understand how conclusions and results are derived from each project. Therefore, they will be interested in raw data and research methodologies. Their goals are to verify the research results and try to build and create addenda research programs.
   – Desirable user interface: Their focus will primarily be on the interface to individual projects. They want to be able to understand the traceability among different parts of the projects from raw data sets to final results.

Lessons from Relevant Systems

The research team surveyed a wide range of systems from the transportation sector, other research disciplines, and even consumer sites that shape user expectations, to find relevant user interface examples that might inform the conceptual design of the proposed archival system. Appendix D contains information on those sites and what was gleaned from reviewing them. This review revealed the following general characteristics of these sites:

- Almost all provide multiple methods for users to navigate to information. These methods include direct access to data sets, complete or partial views of the data, and access to information from a particular business process or decision point.
- When information is deemed to be sensitive, users must be registered to access it. Users are assigned to specific profiles with appropriate privileges.
- Collaborative functions that facilitate connections among users are common on most of the websites. A community is an expected "Web 2.0" feature that is valuable and also relatively easy to implement technically.
- Search is one of the most often used approaches for users to find information. Basic and advanced search functions are common. Filtering of outcomes is becoming more prevalent. Search scope often covers both content and metadata.
- Any system providing data access has numerous provisos related to legal information, privacy policy, program disclaimers, and accessibility aids.

Guiding Principles for User Interfaces

From the user profile analysis and review of relevant representative systems, the research team identified the following four principles that apply to the proposed archival system generally and to the user interface specifically:

- Openness: An open system is one that may be accessed by users operating on differing platforms, other application languages, and independent network infrastructures. The operational system should not impose any undue restraints on the user regarding hardware, software, and connectivity other than those currently used by the user to access the Internet and that are widely available within the industry.
- Zero client administration: The delivered system should not require any special administration on the client side other than the availability of the most basic requirements such as an operating system and a standard web browser. As the operating system continues to grow and mature, it must be able to do so without having to manually modify the client system and/or manually install new software on the client.
- Expandable: The system must be able to continually grow and expand in both the content and the services it provides. As much as possible, these changes should be transparent to the client. New services should be able to come online with little or no impact on existing services.
- Easy to use: Finally, the system must be easy to use. Overly complicated user interfaces tend to fall into disfavor and end up not being used.

In addition to the above, the user interface of the system must comply with the accessibility standards of Section 508 of the Rehabilitation Act of 1979, thereby insuring that the system can be used by persons with disabilities through the use of various assistive technologies.

Conceptual User Interfaces and Requirements

The following sections discuss methods various mechanisms might use to navigate and use the system, as a means to define some basic requirements for the user interface.

Home Page

The home page of an online system typically establishes the top-level navigation scheme for the site. The home page of the archival system should provide various navigation paths to information, as illustrated in Figure 3.7 and explained in the following sections.

Navigation of Reliability Research Projects

As described earlier, because of their varied roles and responsibilities, users will be interested in different aspects of the Reliability products, and the system should provide different and flexible navigation alternatives.

Direct Project Lists

This approach is similar to the way information is currently organized on the SHRP 2 section of the TRB website. Users can click on a Reliability Project Database link to find lists of Reliability projects. Each project name is another link that will lead to the project information. This navigation mechanism would be useful to users already familiar with the SHRP 2 projects, who want to find particular information about a specific project.

Reliability Themes

The SHRP 2 Reliability research plan defines four subject matter themes. Each of them directly links to the four strategic

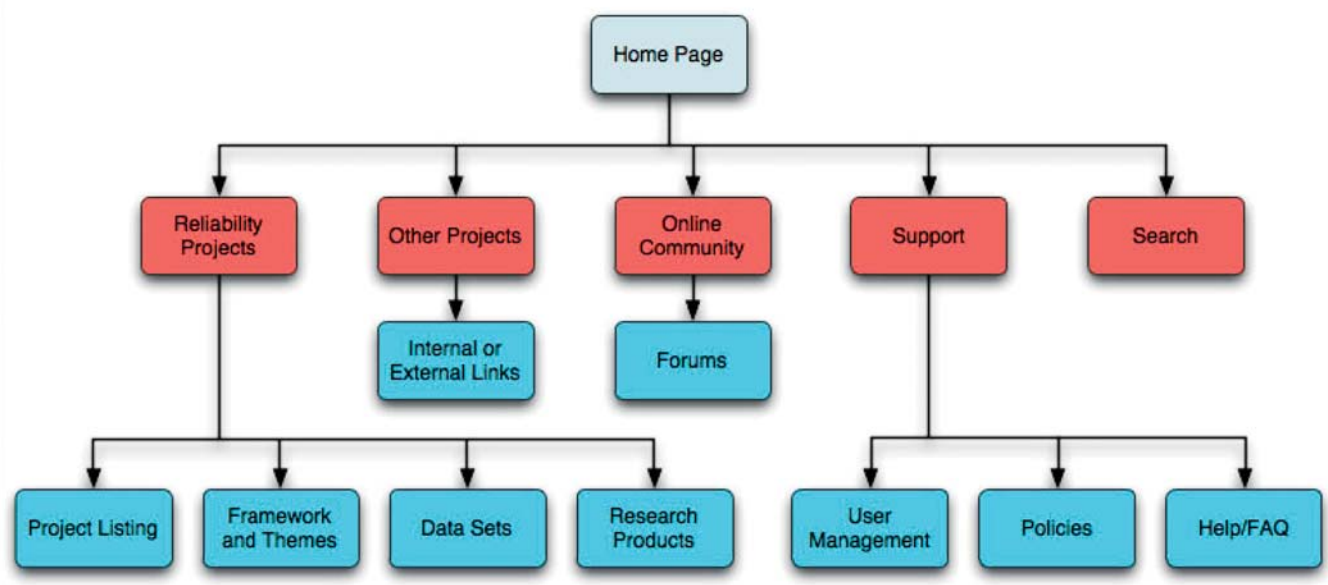


Figure 3.7. Archive home page navigation paths.

objectives, namely, reduction of nonrecurring incidents, improved incident management, improved incident response, and mitigation of the effects of incidents on highway users. The four themes are as follows:

- Theme 1. Data, Metrics, Analysis, and Decision Support;
- Theme 2. Institutional Change, Human Behavior, and Resource Needs;
- Theme 3. Incorporating Reliability in Planning, Programming, and Design; and
- Theme 4. Fostering Innovation to Improve Travel Time Reliability.

Each theme consists of a group of relevant Reliability projects. The research plan describes in detail the scope of each theme and its related projects, which are not repeated here. However, these four themes can serve as another excellent grouping of individual projects. The advantages of this approach are that users can find similar projects in one place and see how they address different aspects of each strategic objective.

Reliability Program Framework

The Reliability research plan includes project L17, which is expected to develop a framework for improving travel time reliability. Based on the initial project description, the framework intends to "package the results of the SHRP2 Reliability portfolio of projects in a concise and accessible manner, and to provide a graphic illustration of how the projects fit together." This framework, once developed, might be another excellent navigational device to individual project information.

However, according to the research plan at the time this report was written, the RFP for this framework will not be published until March 2010. It is expected that the earliest the work will begin would be in the fall of 2010. Given the project's planned 18-month duration, the framework will not be completed until mid-2012. The proposed implementation plan for the Reliability Archive is an 18-month project expected to be completed in mid-2011. Therefore, incorporating the concept of the L17 framework into the L13A Archive system user interface design will be challenging. One possibility is for the L17 task plan to include an interim deliverable that provides design input for this aspect of the L13A user interface.

Data Set Organization

The previous approaches are centered on how to help users find the projects they are seeking. This might not be the only way for users to access information. Sometimes users may want to find out what particular sets of source data are available or were used in a project. To enable users to easily find raw data sets, the Reliability Archive should provide a "source data sets" navigation method. The data sets can be organized by projects, locations, subjects, sources, and collection methods. By clicking on a data set link, users can find the following information:

- Data set name;
- Collection method;

- Related project;
- Location of the data set;
- Format and size of the data set; and
- Derived data and research results.

To assist users in finding the data sets they are interested in, the system can provide a map-based alternative showing the locations of these data sets. If appropriate, geo-locator metadata could be used in conjunction with an external mapping service such as Google Earth to visualize the locations.

Grouping of Research Products

As analyzed earlier, a large group of users will be merely interested in the end products of the Reliability program and how to apply them to their day-to-day responsibilities. Therefore, the Reliability Archive user interface should be able to provide these users with a direct access to the end products.

From another perspective, as suggested in the SHRP 2 implementation report (*5*), implementation of Reliability products will deliver the most benefits when the products are used together as part of an integrated, systemic approach that includes institutional, analytical, and technological components.

There can be different ways to group the end products. One way is to group them according to the following business functions of typical transportation agencies:

- Planning;
- Design; and
- Operations.

Alternatively, the products can be grouped according to these detailed subject interests:

- Quantitative relationships;
- Analytical tools;
- Performance measures;
- Operational strategies;
- Dissemination strategies;
- Best practices;
- Effective organizational and institutional structures;
- Training programs;
- Concepts of operations;
- Framework;
- Business processes; and
- Portfolio of innovative ideas.

The products may also be grouped in a way that links certain relevant projects. For example, the anticipated Reliability Project L05, Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes, is to develop procedures for the transportation planning and programming process that demonstrate the benefits of operational strategies aimed at improving mobility and reliability. According to its initial work plan, the project will build on the statistical relationships between countermeasures and reliability performance measures developed in Reliability Project L03. In the first phase, the L05 contractor will also develop corridor- and network-level strategies using countermeasures and strategies from Reliability Projects L03, L07, and L11, integrated business processes identified in Reliability Project L01, and model results from Reliability Project L04, as well as information from other sources.

Navigation of Project-Level Data and Results

Since the SHRP 2 Reliability program is carried out via individual projects, project-level navigation is expected to be one of the main navigation paths for users.

A significant amount and variety of content might be archived for each project, including raw data, methodologies, and research outputs. One approach for presenting the project information is through simple lists. However, this approach will not best convey the knowledge produced by the project, and as a result will not help users understand the implications of the results; nor will it assist in verifying results or in building new research upon these results.

From a knowledge management perspective, it would be more effective for each project to have a home page that presents a project-focused navigation and traceability chart similar to that presented in Figure 3.8.

Figure 3.8 illustrates the relationship between different concepts, principles, and outcomes from the project. It also shows the traceability of the final results or conclusions drawn from raw data via using the methods, programs, and formulas defined in the project.

This approach aligns with the concept of Resource Framework Diagram (RFD) technology in the W3 specifications. RFD is intended to link loosely coupled data or contents in order to model and share distributed knowledge. These linkages among objects in the Reliability Archive would be encoded as a kind of descriptive metadata.

A mock-up of such a project knowledge map based on Reliability Project L03 as an example is shown in Figure 3.9. Other project-level page mock-ups illustrating different modes of information discovery are shown in Figures 3.10–3.12.

The mockup shown in Figure 3.10 lists all the raw data sets used in the Reliability Project L03. Users would come to this screen by clicking the "Raw Data" item on the left navigation menu. Clicking the details link on an individual item would take the users to a page showing additional details pertaining to that item.

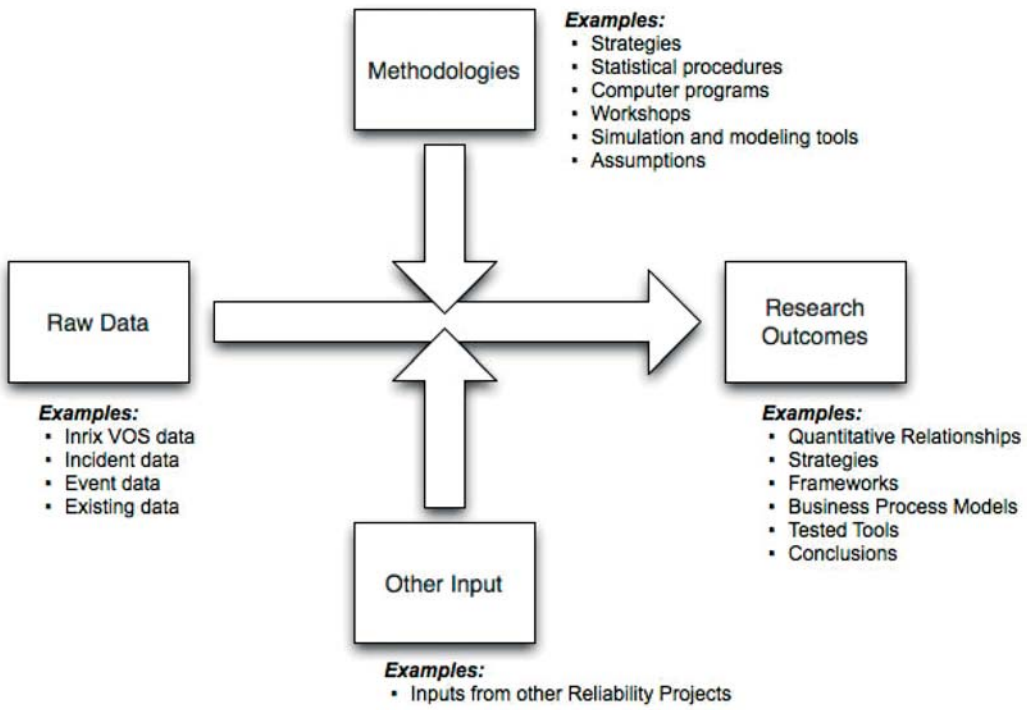


Figure 3.8. Project-level navigation and traceability.

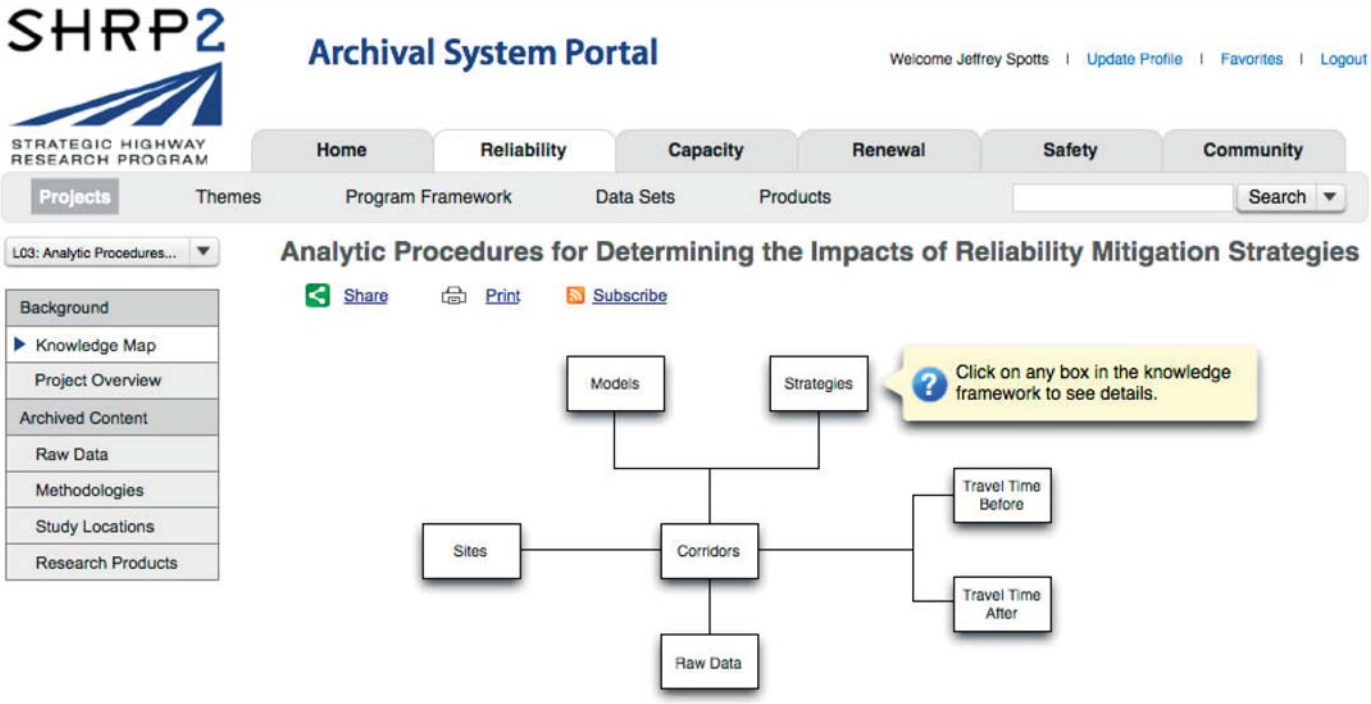
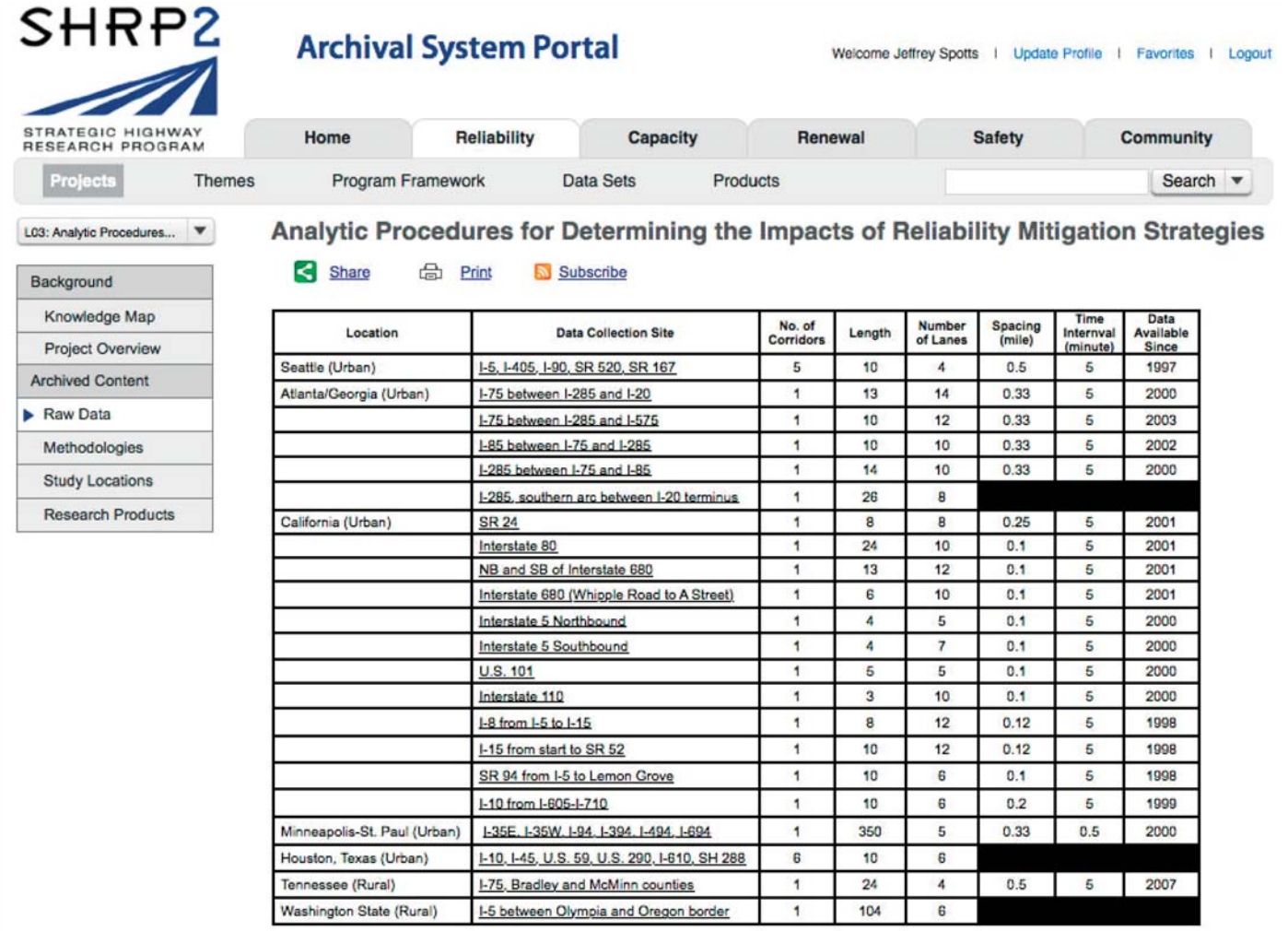


Figure 3.9. Project-level knowledge map mock-up.



Location	Data Collection Site	No. of Corridors	Length	Number of Lanes	Spacing (mile)	Time Interval (minute)	Data Available Since
Seattle (Urban)	I-5, I-405, I-90, SR 520, SR 167	5	10	4	0.5	5	1997
Atlanta/Georgia (Urban)	I-75 between I-285 and I-20	1	13	14	0.33	5	2000
	I-75 between I-285 and I-575	1	10	12	0.33	5	2003
	I-85 between I-75 and I-285	1	10	10	0.33	5	2002
	I-285 between I-75 and I-85	1	14	10	0.33	5	2000
	I-285, southern arc between I-20 terminus	1	26	8			
California (Urban)	SR 24	1	8	8	0.25	5	2001
	Interstate 80	1	24	10	0.1	5	2001
	NB and SB of Interstate 680	1	13	12	0.1	5	2001
	Interstate 680 (Whipple Road to A Street)	1	6	10	0.1	5	2001
	Interstate 5 Northbound	1	4	5	0.1	5	2000
	Interstate 5 Southbound	1	4	7	0.1	5	2000
	U.S. 101	1	5	5	0.1	5	2000
	Interstate 110	1	3	10	0.1	5	2000
	I-8 from I-5 to I-15	1	8	12	0.12	5	1998
	I-15 from start to SR 52	1	10	12	0.12	5	1998
	SR 94 from I-5 to Lemon Grove	1	10	6	0.1	5	1998
	I-10 from I-605-I-710	1	10	6	0.2	5	1999
Minneapolis-St. Paul (Urban)	I-35E, I-35W, I-94, I-394, I-494, I-694	1	350	5	0.33	0.5	2000
Houston, Texas (Urban)	I-10, I-45, U.S. 59, U.S. 290, I-610, SH 288	6	10	6			
Tennessee (Rural)	I-75, Bradley and McMinn counties	1	24	4	0.5	5	2007
Washington State (Rural)	I-5 between Olympia and Oregon border	1	104	6			

Figure 3.10. Raw data overview mock-up.

Figure 3.11 provides specific information about a particular data set such as the data set "I-8 from I-5 to I-15 in California" depicted in the mock-up. The tabular information included on the page is the metadata for this data set prepared by the project contractor. Clicking the "Download" button would download the file over HTTP to the user's system, assuming they have access rights. The "Download" button would be gray and inoperable if a user's role does not allow such access.

Figure 3.12 is a mock-up of a map-based view of a project's study sites that enables users to access the raw data and research products associated with these individual sites.

Other Projects

Users should be able to navigate to content related to other SHRP 2 Capacity, Renewal, and Safety projects, whether or not the content is preserved in the archival system. The same types of navigational schemes described for the Reliability focus area could be applied to the other three focus areas.

Online Community

The system should provide an online community environment where users can pose questions, get answers, and share their experience and expertise. This is another example of an environment users may want to enter independently of how they are navigating the system. For example, a user may be viewing information about a particular project or data set and might wonder how another researcher interpreted a result or used the data. A natural way to find answers to such questions would be to go to the community and see what has been posted relevant to that topic, or to write an inquiry post. Thus, community access should be available on every page of the system in a consistent location.

Search

All of the navigation mechanisms discussed thus far are based on deterministic paths. Fixed hierarchies are necessary and useful, but are insufficient alone. It is not possible to predict

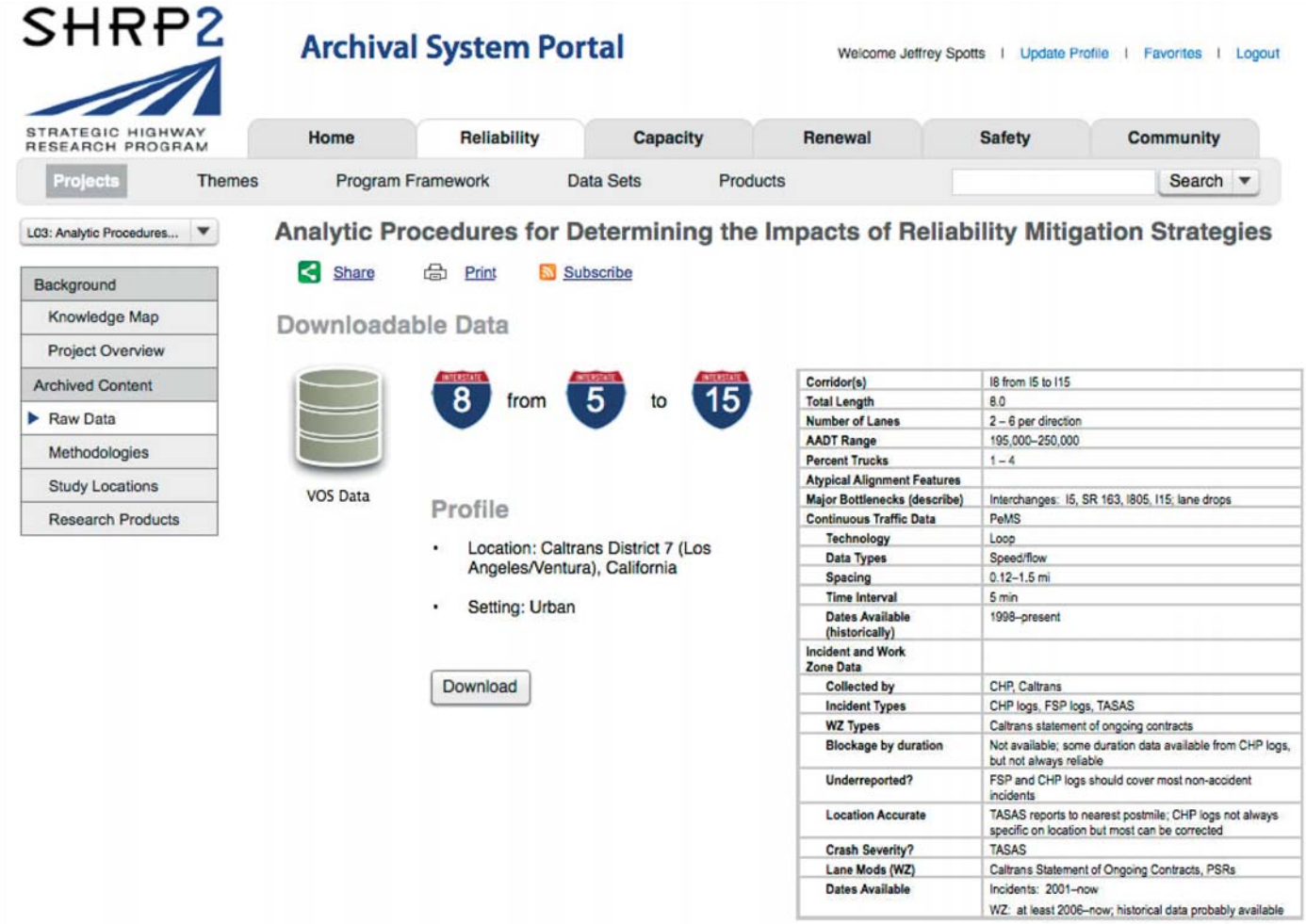


Figure 3.11. Project raw data overview page mock-up.

all of the ways that future users may wish to seek and connect information.

Users should be able to search the archive based on both content and metadata. Since this is a general capability that a user might want to invoke at any point in his or her interaction with the system regardless of the navigation path he or she has taken, the search function should be a capability that would be accessible from anywhere in the access portal independent of any other more structured navigational schemes. The mock-ups serve to illustrate how some of these capabilities might be realized in an actual system.

Simple and Advanced Searching

Simple searches might be done by typing a keyword or phrase into the search box on the main navigation bar and clicking the Search button, as shown in the Figure 3.13.

Instead of clicking the Search button, the user might pull down the combo control and find an Advanced Search option that leads to a page such as that depicted by Figure 3.14, which would allow him or her to build a complex search expression. Clicking the plus sign at the end of a line adds a new statement to the expression. (The minus sign would delete a statement.) A similar interface is used in Apple's popular iTunes software for building so-called Smart Playlists.

Viewing and Refining Search Results

Independent of the search mode used (simple or advanced), the results of a search would be returned in an interface such as that depicted in Figure 3.15. This example illustrates faceted searching. Additional search facets that are derived from a content and metadata analysis of the result set would appear in the left hand column. Clicking these links in succession would allow the user to filter these results.

Customer Support and Administration

The system should provide self-service interfaces for routine user management tasks. These include new account

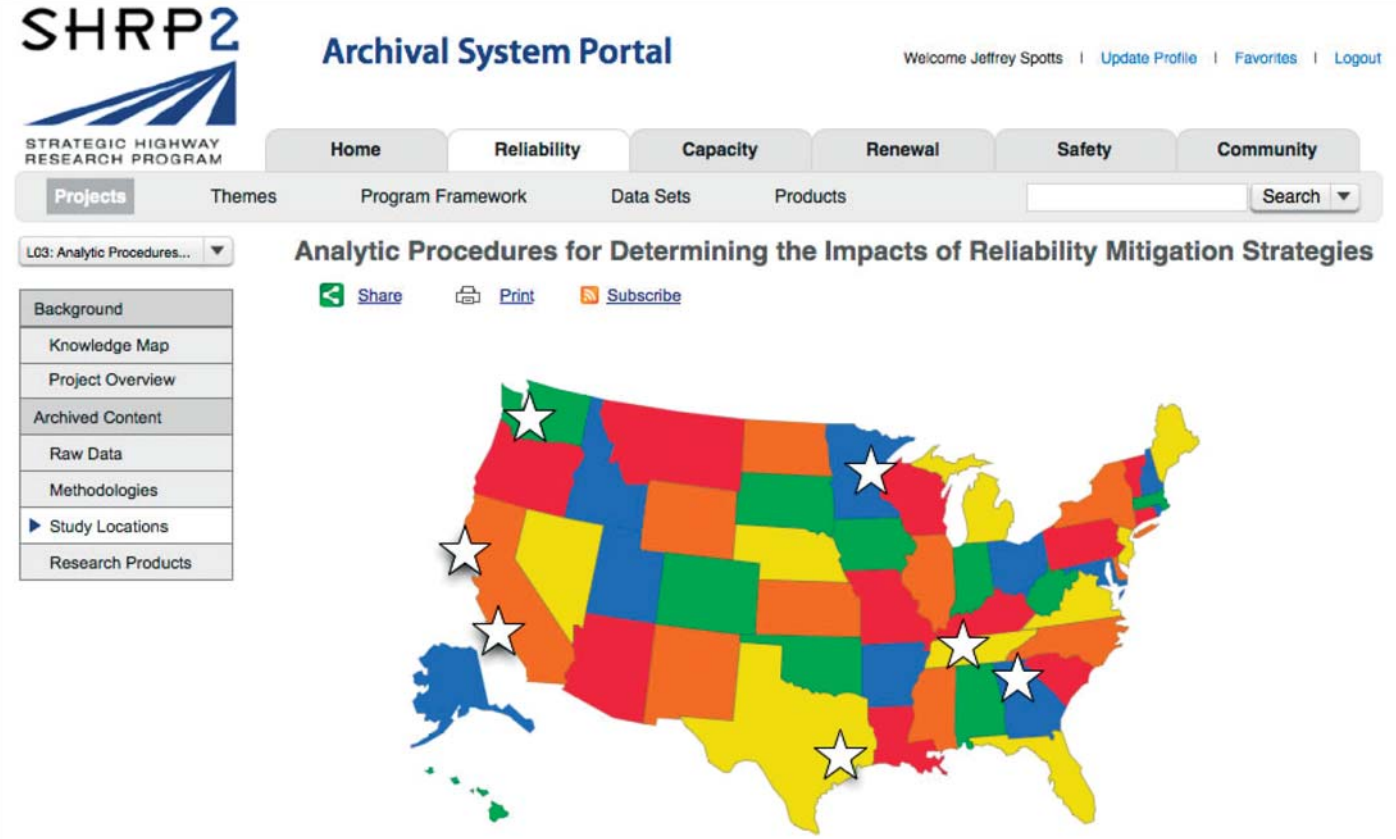


Figure 3.12. Project study locations page mock-up.

registration, user profile management, and password reset requests. These user interfaces often provide Help and FAQ content. A common practice for any site providing downloadable content is to make visible at the top-level of the site any policy statements with respect to privacy, data rights, warranty disclaimers, and other such policies.

Data Integrity and Quality

Based on the conceptual design discussed earlier, the data in the Reliability Archive will consist of the data to be preserved and the metadata associated with the archived data. Both types of data are critical to the success of the archive. Thus, it is important to evaluate and control the quality of both.

Based on how data are collected, used, and produced by individual Reliability projects, and on how the data are then prepared for, submitted to, and preserved in the archival system, there are three logical points of data quality control:

- Within individual Reliability projects;
- Through Reliability Project L16 (now a part of Reliability Project L13A) (designed to assist Reliability project contractors in preparing their data for submission to the archive); and
- By active enforcement of the preservation policy within the archival system.

Figure 3.13. Simple search box.

The approach to addressing data quality will vary with the type of data. Figure 3.16 depicts a digital object (in OAIS terms, an Archival Information Package, or AIP) as it might logically exist in the L13A Archive. An AIP in the Reliability Archive will include three types of information: content information; preservation description information; and packaging information and descriptive information.

Content Information

This consists of the original data sets or data objects. In this example, it is the VOS data sets collected at the I-8 site. These data sets might be in text, binary, or spreadsheet format. The

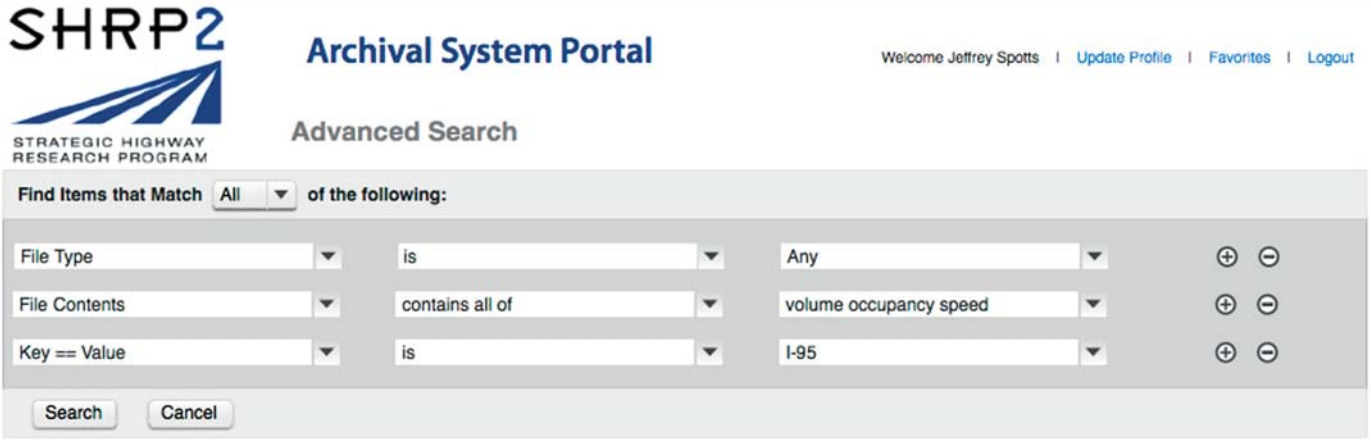


Figure 3.14. Advanced search.

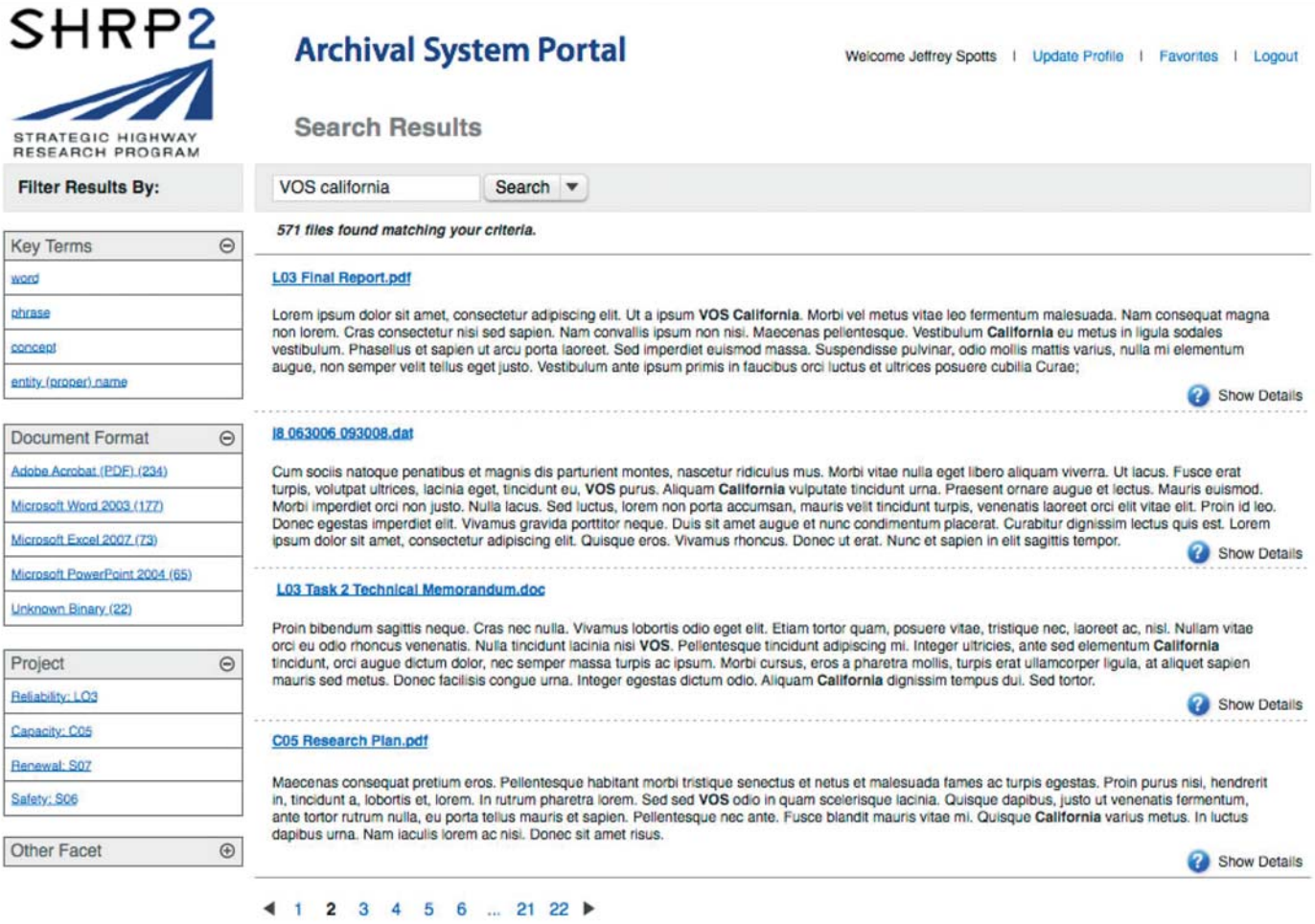


Figure 3.15. Viewing and refining search results.

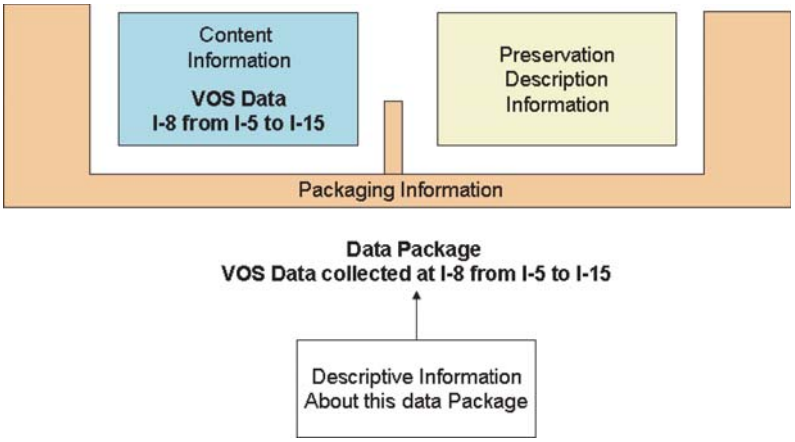


Figure 3.16. Example of archival information package.

quality of this content will be controlled within each individual project. Interviews with the Reliability project contractors indicate that almost all of the projects have robust data quality control standards and processes for the data they collect and produce. For example, the Reliability Project L03 team has developed quality control checks used in the FHWA's Mobility Monitoring Program for identifying suspect or invalid data that will be applied to all roadway-based traffic measurements.

The FHWA's Traffic Data Quality Measurement report (*6*) is one of the most common standards used in these projects. This report describes a data quality framework on six fundamental measures: accuracy, completeness, validity, timeliness, coverage, and accessibility.

When a Reliability project is going to deliver the data to be archived, the project contractor is expected to submit the data along with its data quality control standards, methods, and assessment.

Reliability Project L16, which is designed to assist Reliability contractors in preparing data for archiving, should review the data quality assessment prepared by the contractor and either confirm or modify the quality rating. Given the wide acceptance of the FHWA Traffic Data Quality Measurement report, Project L16 should apply the data quality measurement framework from this report to evaluate and assign the quality rating on the data delivered by the projects. This quality rating would be a metadata attribute that would be part of the preservation description information (PDI) described next.

Project Metadata: Preservation Description Information

Preservation description information is the metadata information to be prepared and collected by individual projects. In the example illustrated in Figure 3.16, this is to clearly identify and understand the environment in which the "VOS Data at the I-8 from I-5 to I-15" (content information) was created. It would include the following information:

- The source of the data collected;
- The context in which the collected data is related to other information from the project;
- The reference by which the content information can be uniquely identified; and
- The fixity that acts like a wrapper or protective shield, to protect the content information from undocumented alteration.

There will be two types of data quality issues with the project metadata. One is that each project will probably use and collect different metadata elements. The other is that some metadata information may be inaccurate or incomplete.

Reliability Project L16 must play a critical role to ensure the quality of the project metadata. For example, L16 should prepare detailed guidelines on what core or mandatory metadata must be provided, along with specifications on data quality. A quality control screen should be set up to assess the project metadata. Feedback should be prepared and sent to contractors in case their metadata is not accurate or complete.

Once the project metadata passes the data quality screen test, they will be saved to the metadata database in the L13 data archive.

System-Generated Metadata: Packaging Information and Descriptive Information

System-generated metadata refers to how the data package is stored in the data archive and how it is referred to with respect to its contents. The critical aspects of data quality will still be data accuracy and completeness.

L16 is expected to create descriptive information for the data package. The tools or technologies selected for the Reliability Archive will save the descriptive information and also automatically generate other system or storage-related information. Table 3.1 summarizes the data quality management process for the Reliability Archive.

Table 3.1. Data Quality Management Summary

Data in Reliability Archive	Aspects of Data Quality	Data Quality Control		
		Reliability Project/ Contractor	Project L16	Reliability Archive System
Content data	Accuracy, completeness, validity, timeliness, coverage, and accessibility	Provide quality assessment	Review the quality assessment and assign a quality rating based on the FHWA Traffic Data Quality Measurement framework	Save the rating; data quality and integrity control at the metadata database level
Project metadata	Accuracy, completeness, and accessibility	Prepare and submit the project metadata	Set up the project metadata standards and guidelines Screen the quality of the project metadata	Save the project metadata; data quality and integrity control at the metadata database level
System metadata	Accuracy, completeness, and accessibility	Not applicable	Create quality descriptive information	Data quality and integrity control at the metadata database level

Data Rights

From interviews with the Reliability project contractors, the research team found the following with respect to the issues of data rights:

- There are few or no restrictions on the derived data from these projects.
- The raw data used in these projects typically come from the contractors' existing data sets, a state DOT or other transportation agencies' detectors and accident data programs, as well as from the private sector.
- Currently, about half of the projects have not identified the sources of the data that will be used.
- As of the date of this report, it appears that INRIX is the only data provider from the private sector. Its agreement with the Reliability Project L03 contractor includes stipulations on the use of raw and derived data.

It was equally important to acquire a good understanding on the same subject from the contract administration and legal perspectives of the National Academies and Transportation Research Board. During the project, the research team met with the general counsel of the National Academies to discuss this matter. The discussion mainly centered around Reliability Project L03's agreement with INRIX with respect to data rights clauses on raw data and derived data. The following summarizes the group consensus from the meeting:

- The goal of the Reliability Archive is to provide future end users with access to SHRP 2 Reliability project data without restrictions. In general, there is no perceived negative impact with respect to data rights affecting the feasibility of building the Reliability Archive.
- The majority of the raw data used by the Reliability projects comes from the public sector, so it poses no data rights issues.
- In any case where there are usage restrictions on raw or base data, the Reliability Archive needs to focus on archiving the derived or aggregated data.
- The omission of original base and raw data from the archive might impact the ability of the future end users to efficiently validate the results of a project. In such a case, the project contractor will need to leverage and maximize the utility of metadata to explain how the derived data was aggregated. The knowledge map described earlier can be another means to guide end users in validating the research results.

Institutional Framework and Governance

Given the size and level of complexity of the Reliability Archive, a proven and reliable institutional framework is warranted in order to provide long-term stewardship of the archive. This section explores a set of key principles that could become the building blocks of this institutional framework.

Best Practices of National Systems

Numerous national systems similar to the Reliability Archive have been developed. Successful systems all have mature institutional frameworks or governance models with a clear, long-term stewardship mission. These mature frameworks possess the following characteristics:

- Clear and well-communicated vision that is shared by stakeholders and participating organizations;
- Well-defined multitiered organization structures, roles, and responsibilities;

- Dedicated funding models to ensure the continuity of available funding to support ongoing administration, maintenance, and technology upgrade;
- Global reach to all possible user groups;
- Strong commitment from key stakeholders, dominant industry players, and influential organizations;
- Willingness to collaborate with relevant standards development bodies and professional associations to leverage existing and emerging technologies, standards, and services; and
- Clear and enforced policies and procedures that are monitored constantly.

SHRP 2 Implementation Report

The SHRP 2 implementation report (*5*) includes specific recommendations on the overall strategies for implementing the SHRP 2 research products. These recommendations encompass an array of issues, such as who is responsible for implementing the results, where and how much funding is needed, and how to set up implementation priorities. The report also discusses potential roles and responsibilities of national transportation organizations such as FHWA, TRB, and the American Association of State Highway and Transportation Officials (AASHTO). The right direction for the Reliability Archive is to develop its institutional framework under the guidance of these recommendations.

Principal Implementation Agent

One of the key recommendations of the SHRP 2 implementation report is to establish a principal implementation agent that is a national organization that will lead and support SHRP 2 implementation (*5*). A similar role should also be established for the archival system. The Reliability Archive principal implementation agent will be responsible for the following tasks:

- Implementing the Reliability Archive to a production environment once its development is completed under SHRP 2;
- Long-term managing of the data archive, including system administration, maintenance, and upgrade;
- Communicating with the user community on matters such as updates on the implementation and new contents added;
- Establishing relevant policies and procedures for using the archival system; and
- Maintaining coordination with stakeholders at both the strategic and technical levels.

As recommended in the SHRP 2 implementation report, FHWA should serve as the principal implementation agent for SHRP 2, in partnership with AASHTO, National Highway Traffic Safety Administration (NHTSA), and TRB.

This recommendation is similar to the approach taken with LTPP. Currently, LTPP is administrated and maintained under FHWA with support from contractors who provide technical resources and system production support. This model could be applicable to the Reliability Archive.

Stakeholder Advisory Group

To support the principal implementation agent, a formal stakeholder advisory group should be established to provide strategic guidance and technical advice on the long-term stewardship and use of the archive. This advisory group should operate under the SHRP 2 implementation oversight committee to coordinate overall implementation strategies.

The advisory group should include the principal users of the archive and broad stakeholder representation such as leaders of state DOTs, technical staff, nontransportation professionals, academic researchers, as well as experts in information technology and knowledge management.

The advisory group should be responsible for the following:

- Coordinating with the SHRP 2 implementation oversight committee to ensure that the archive implementation approach aligns with the overall implementation strategies;
- Providing both policy and technical guidance to the principal implementation agent;
- Setting priorities for maturing, maintaining, and upgrading the archive;
- Developing communication strategies with user groups to maximize the awareness, access, and usage of the archive; and
- Monitoring progress on the archive implementation and reporting it to the SHRP 2 implementation oversight committee.

Use of Private Sector IT Services

A key part of the institutional framework for the archival system is to ensure that it will be available to users on a 24/7 basis. This requires that system administration and maintenance processes follow rigorous standards, which demands reliable information technology infrastructures and skilled personnel. Although FHWA is an ideal candidate as an implementation agent and has strong IT resources, FHWA is not an IT service shop and does not specialize in providing product system support services. Thus, alternatives need to be explored. A practical option is for system administration and maintenance to be outsourced but managed by the Reliability Archive's principal implementation agent.

Technical Issues

Some specific technical issues were cited explicitly for analysis in the L13 Reliability Project RFP. The research team explored the applicability of each technical issue to the Reliability Archive. These issues are data normalization and

denormalization, online analytical processing (OLAP) and user-defined functions, service-oriented architectures (SOA), and virtualization.

Normalization and Denormalization

The term "normalization" originated in 1970 with the work of E. F. Codd at IBM, considered by many to be the father of the relational database (*7*). Virtually all modern transactional database applications strive to represent data in what Codd called first normal form (1NF), essentially meaning that no table should contain any repeating groups (arrays). Of course, arrays are pervasive in real-world data, so they are handled in relational database systems via relationships between tables. (One row in a master table might be related to N number of rows in a details table, thus obviating the need to fix the maximum number of detail items, which is the case in a denormalized data structure.)

As relational databases became widely adopted, performance problems began to be observed in highly query-intensive applications with fully normalized data models. A recent trend in the database market has been the development of specialized databases for "read-mostly" applications such as OLAP, which employ selective denormalization to speed up query performance.

This entire subject area is a large and fairly complex one that can be dealt with here only in summary fashion. The bottom-line question is whether or not data normalization or denormalization has any application in the proposed archival system. The research team believes that the answer is no, at least in terms of normalizing or denormalizing data postresearch as part of the process of preparing it for preservation.

As discussed in Chapter 2, a fundamental purpose of an archive is to preserve unchanged the information entrusted to its care, and to facilitate access to this information when needed. Basic preservation principles argue against such an obvious structural reorganization of data in order to preserve it.

That said, an investigator might normalize or denormalize data in the routine course of his or her project. For example, normalized raw data might be the basis for denormalized, aggregate data used in an analytical model. As pointed out previously, all data sets and the relationships among them are important in establishing the traceability of results; thus, all should be part of the collection submitted to the archival system. Another way of saying this is that normalized and denormalized data should be able to coexist, and be linked, if appropriate, in the archival system.

OLAP and User-Defined Functions

The purpose of the Reliability Archive, based on its guiding principles and user requirements, is to serve transportation researchers and decision makers by preserving transportation project information and facilitating lookup, presentation, and downloading of such information. Therefore, 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. Analyses such as OLAP and user-defined functions are domain-specific and should be addressed by each user based on his or her specific needs. Any attempt to provide such analyses as a function of the archival system would likely miss the mark. They would be costly to build and maintain and, absent any concrete requirements, would likely be ineffective. Beyond the preservation mission of the archive, the appropriate emphasis should be on facilitating the finding of the correct information and getting it into the user's hands for any subsequent manipulation. Toward this end, one potentially useful technology is mashups, which are discussed in the following section.

Mashups

The focus of the Reliability Archive is to provide users easy access to project information, which includes not only SHRP 2 Reliability projects but also other projects from the Capacity, Renewal, and Safety focus areas (some of those projects may have their data and metadata stored in the archival system, while others may have their own storage facilities). In addition, the system may also facilitate the search and downloading of other relevant information outside the SHRP 2 focus areas.

As shown in Figure 3.17, the Reliability Archive will potentially need to provide access to information from multiple sources. It is likely that users, particularly researchers, will want to aggregate data from the archive, or even aggregate data from the archive with data found elsewhere. This could be achieved by using mashup technologies, which would provide aggregated data from the archival system and various other sources.

Mashups have the following three fundamental, defining characteristics:

- They are lightweight composite applications that employ a web-oriented architecture to provide quick information integration for end users;
- They source content or functionality from established systems and have no native data store or content repository; and
- The mashup result is an explicit mixture of source content and functionality, where the sourced content and functionality retain their original essence or purposes.

A mashup environment will enable the construction and use of three fundamental mashup entities: mashup components, mashups, and mashup applications (see Figure 3.18 for the architecture). A mashup application consists of one or more mashups; a mashup consists of two or more mashup components.

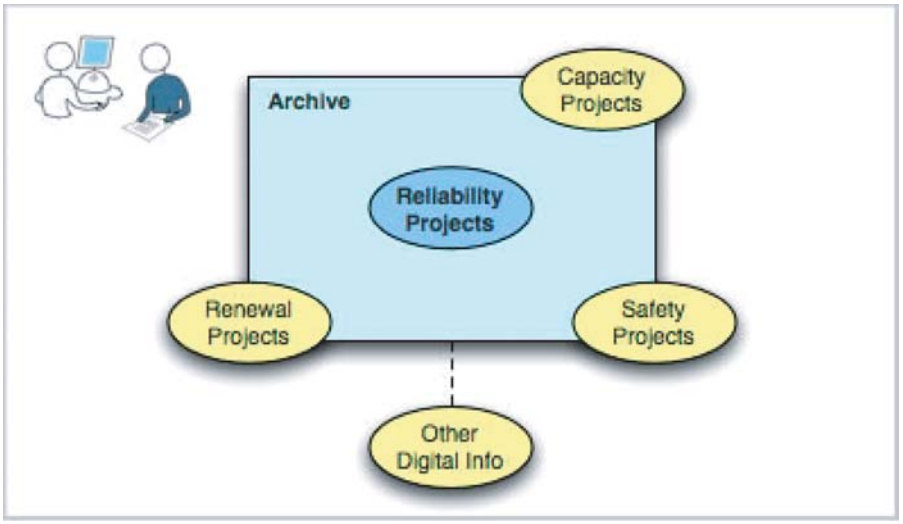


Figure 3.17. User view of mashup service.

Mashup Sources (Information and Function)

Mashups source their content and functionality from established information systems. In the case of the SHRP 2 Reliability projects, this would include the Reliability Archive and other relevant information sources, some of which may not be web based.

Information Access, Augmentation, and Delivery

Non-web-based sources are transformed and made available for mashups.

Mashup Assembly

The mashup assembly process provides access to mashup components, the means to assemble these components into a mashup, and the ability to preview the result. Mashup assembly should also provide search capability of the mashup components and their metadata.

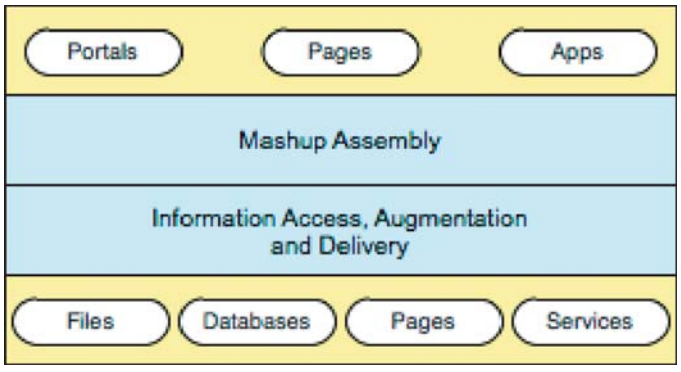


Figure 3.18. Mashup reference architecture.

Mashup Visualization

Mashup visualization delivers a mashup to its destination, usually a web page, portal, or web-based application. Like the other technical issues discussed in this section, the key question regarding mashups is, Does it have applicability to the Reliability Archive? The research team found that while it might be applicable (i.e., some prospective users might like the system to provide a general data aggregation service), there was no clear requirement to include such capability in the system. Mashup technology, moreover, would only make the Reliability Archive more complex. Not including a mashup service as a requirement today does not of course preclude it from being added to the Reliability Archive at some future date.

Service-Oriented Architecture (SOA)

SOA refers to a method for systems integration where systems expose functionality as interoperable services. The concept goes back to the first examples of distributed computing systems and is now associated with web services, making the concept practical on a wide scale. Web services provide the capability to integrate disparate data by exposing the data as discrete web services accessible over open, standardized protocols. This provides a unified means of accessing information from a diverse set of sources and platforms.

Mashups are an example of functionality that can be delivered by the archival system using a SOA. SOA and web services can be expected to play other roles in the Reliability Archive. The search function of the system could span other repositories (known as a federated search) if other repositories expose their indexes as a web service. Similarly, the archival system could expose its index as a web service so that it can be the target of a federated search invoked on another system.

Virtualization

Virtualization is a popular topic today in information technology circles. Virtualization uses software to abstract a hardware environment. It is best known for its application in insulating an operating system from the underlying 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 form of platform virtualization is prevalent in server environments for shared hosting and is now quite common on desktop environments. This application of virtualization is 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.

Virtualization is also an interesting possibility for certain archival situations. For example, archivists in museums and libraries who catalog the personal papers of artists, politicians, scientists, and others are now confronting the possibility that the collections donated to their institutions will include removable storage media and even complete computer systems, in addition to the usual journals, files, and other paper records they have received historically. Archiving a virtual machine image is a possible means of preserving information and the execution environment on which access to that information depends.

This approach to archiving would introduce other problems. Accessing an archived virtual image successfully would now be dependent on having a version of the virtualization software that can run the archived virtual machine image. The attraction is reducing the vicious cycle of format dependencies from many (all the applications needed to access data on a given machine) to one (the virtualization software). Although no case has been identified where this technique might be applicable in the Reliability Archive, it could be considered should such a requirement emerge.

In storage, virtualization is used to abstract logical storage from physical storage. Some form of storage virtualization could be used in the actual deployment of the proposed archival system, since the technique would facilitate the physical migration of archived content to new storage media over the life of the system.

Establishing Solution Alternatives

The research team began to map the 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 described in Figure 3.19.

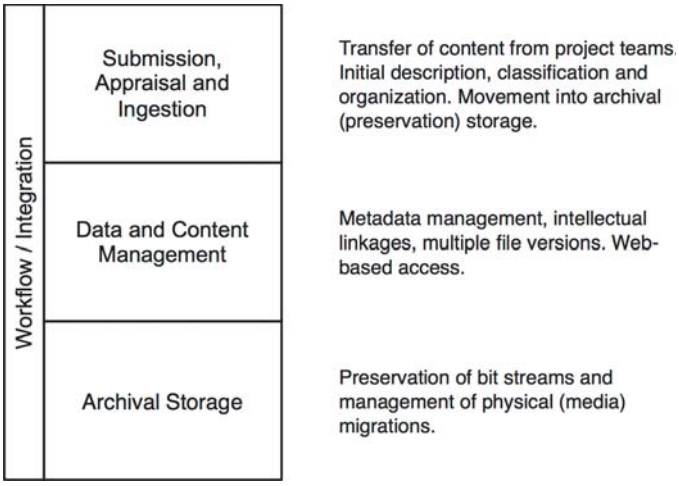


Figure 3.19. Functional blocks of the proposed archival system.

The following were identified as critical issues that influence the selection of potential alternatives:

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

Both issues are analyzed in the following sections.

Importance of System Functionality

Over the expected life of the Reliability Archive—more than 25 years—the relative importance of functionality will change (see Figure 3.20). The trustworthiness, reliability, and durability of archival storage are constants throughout the life of any archive; these are areas where trade-offs should be avoided, if possible.

Submissions to the archive will be made by project teams as their respective projects conclude. These submissions will be assessed and then ingested into the archival system. This process will conclude perhaps three years after the system becomes operational; this is thus an area where the long-term sustainability of this functionality is of lesser importance.

Content and data management is very important throughout the life of an archive. Arguably, the importance of this function grows over time because this function impacts the curation of the archive and how effectively the information it contains is exploited by practitioners and researchers.

Estimated Data Volume of the Archive

The overall data volume that has to be managed over the 25-year expected life cycle of the Reliability Archive presents

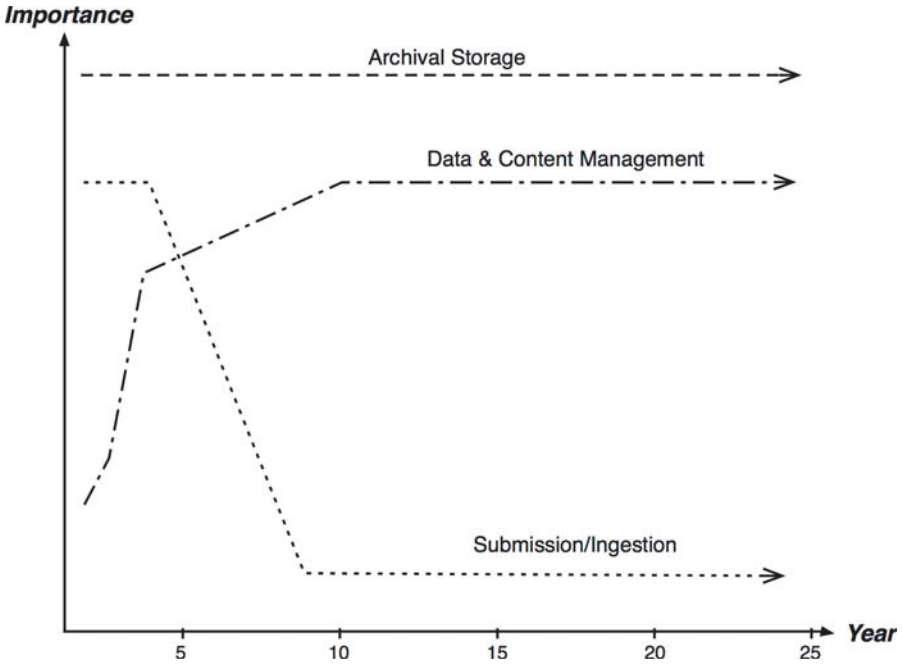


Figure 3.20. Relative importance of functionality over time.

certain challenges and will influence the ultimate choice of a storage system. Because no SHRP 2 research project is completed yet and many have not even begun, the research team had to come up with a reasonable set of assumptions to build a model for the estimated storage capacity needed in the archival system.

This model categorizes each project into one of the following types:

- Type 1: Mostly documents;
- Type 2: 50% data and 50% documents;
- Type 3: 75% data and 25% documents; and
- Type 4: Over 95% data.

A capacity "base value" was assigned for each of these types, as shown in Table 3.2. This base value was derived from information gathered from interviews with all the project contractors.

Because some data sets *may* be stored in XML format, an XML overhead factor was included that takes into account additional space typically needed for encoding binary information as text in XML files. A metadata factor was also incorporated in the model to account for the need to store metadata for each object. The value of this factor will increase with the level of complexity of the project and the volume of data to be archived. Finally, a headroom factor was provided to ensure that there will be a certain amount of additional space available to satisfy unanticipated storage needs and to ensure the system is running at less than 100% of storage capacity.

Figure 3.21 summarizes the estimate of usable storage capacity required. (Usable capacity refers to space needed to store user files. Raw capacity will be higher because of formatting overhead, RAID overhead, hot spaces, and other factors, depending on the system implementation.) The research team used 70 TB of usable capacity as the basis of the life-cycle cost estimates across all of the solution alternatives.

Solution Components and Implementation Approaches

As part of the solution-visioning process, the research team considered a range of potential sources of technology. Candidate application suites were identified that provided end-

Table 3.2. Storage Model Parameters

	Model Parameters		
Project Type	**XML Capacity (GB)**	**Overhead Factor**	**Metadata Factor**
Type 1: Mostly documents	100	1%	2%
Type 2: 50% data and 50% documents	500	5%	7%
Type 3: 75% data and 25% documents	1,000	10%	10%
Type 4: Over 95% data	20,000	15%	15%
Headroom Factor	20%		

Reliability and Relevant Projects	Status	Project Type	Capacity (GB)	XML Overhead Factor	XML Overhead Space (GB)	Total Storage (w/ headroom)	Metadata Factor	Metadata Space(GB)	Total (GB)
L01 Identification and Analysis of Best Practices	Active	Type 1 - Most Document	100	0.01	1	121.2	0.02	2	123.2
L02 Establishing Monitoring Programs for Mobility and Travel Time Reliability	Active	Type 2 - 50% data and 50% document	500	0.05	25	630	0.07	35	665
L03 Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies	Active	Type 4 - Over 95% data	20,000	0.15	3000	27600	0.15	3000	30600
L04 Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools	Pending	Type 1 - Most Document	100	0.01	1	121.2	0.02	2	123.2
L05 Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes	Planned	Type 1 - Most Document	100	0.01	1	121.2	0.02	2	123.2
L06 Institutional Architectures to Advance Operational Strategies	Active	Type 1 - Most Document	100	0.01	1	121.2	0.02	2	123.2
L07 Evaluation of Cost-Effectiveness of Highway Design Features	Active	Type 2 - 50% data and 50% document	500	0.05	25	630	0.07	35	665
L08 Incorporation of Non-recurrent Congestion Factors into the Highway Capacity Manual Methods	Planned	Type 2 - 50% data and 50% document	500	0.05	25	630	0.07	35	665
L09 Incorporation of Non-recurrent Congestion Factors into the AASHTO Policy on Geometric Design	Planned	Type 2 - 50% data and 50% document	500	0.05	25	630	0.07	35	665
L10 Feasibility of Using In-Vehicle Video Data to Explore How to Modify Driver Behavior that Causes Non-Recurring Congestion	Pending	Type 3 - 75% data and 25% document	1,000	0.1	100	1320	0.1	100	1420
L11 Evaluating Alternative Operations Strategies to Improve Travel Time Reliability	Active	Type 3 - 75% data and 25% document	1,000	0.1	100	1320	0.1	100	1420
L12 Improving Traffic Incident Scene Management	Active	Type 1 - Most Document	100	0.01	1	121.2	0.02	2	123.2
L14 Effectiveness of Different Approaches to Disseminating Traveler Information on Travel Time Reliability	Planned	Type 4 - Over 95% data	20,000	0.15	3000	27600	0.15	3000	30600
L15 Reliability Innovations Deserving Exploratory Analysis (IDEA)	Planned	Type 2 - 50% data and 50% document	500	0.05	25	630	0.07	35	665
C04 Improving Our Understanding of How Highway Congestion and Pricing Affect Travel Demand	Active	Type 3 - 75% data and 25% document	1,000	0.1	100	1320	0.1	100	1420
C05 Understanding the Contribution of Operations, Technology, and Design to Meeting Highway Capacity Needs	Active	Type 3 - 75% data and 25% document	1,000	0.1	100	1320	0.1	100	1420
		Total	**47,000**		**6,530**	**64,236**		**6,585**	**70,821**

Figure 3.21. Estimated storage capacity needed for the archival system.

to-end coverage of submission, appraisal, ingestion, and data and content management. These suites generally abstracted the interface to the storage tier, thus allowing freedom of choice for the archival storage layer. Within the archival storage tier itself, there were suboptions. The research team also identified candidate software tools that focused on specific tasks that could be considered as components of a system.

In addition to identifying *what* software and hardware technology might address the functional and operational requirements of the archival system, the research team also needed to consider the question of *how* the technology could be acquired and implemented. These options include commercial off-the-shelf technology (COTS), open-source software (OSS), in-house software development, hosting, and software and storage as a service (SaaS).

The following sections discuss the kinds of solution components considered, the various technology implementations available, and how the research team analyzed which choices are appropriate for the institutional framework in which the system will be deployed and managed.

Commercial Off-the-Shelf Technology

COTS technology is software and hardware that is ready-made and available for sale, lease, or license to the general public. The research team considered both COTS software and hardware products as potential sources of technology.

Commodity versus Specialized Hardware

It is useful to distinguish between commodity hardware (e.g., servers or generic storage) that is readily interchangeable from vendor to vendor, from specialized hardware that is unique to a given vendor (and therefore more proprietary in nature).

Open-Source Software

Open source has become one of today's most popular models of software development. OSS is created and maintained via a collaborative model. Larger open-source projects often have primary sponsors, which include commercial, governmental, and nonprofit entities. Contributors to open source projects may be motivated individuals, but many are employees of technology companies assigned to work on such projects.

With OSS, users can go to a trusted repository on the web to obtain a copy of the source code, which is distributed under one of several licenses (e.g., the GNU General Public License, or GPL) that provides users the freedom to run the software for any purpose, to study and modify the source code, and to freely redistribute copies of either the original or modified software without royalty payments or other restrictions on who can receive them.

The lines between commercial and open-source software are blurring. Many proprietary software products today incorporate some components that are licensed under open-source terms. And many OSS packages are available as commercial

distributions where the distributor adds value in terms of testing, integration with other technology, certification on certain hardware, and support.

When considering OSS, the research team found it useful to distinguish between products that are supported under commercial terms and packages that are only available on a community-supported basis, meaning that users are essentially on their own and have to figure out problems with the assistance of the community using that package.

In-House-Developed Software

Inasmuch as the National Academies and TRB and any likely implementation agent for the proposed Reliability Archive have little to no in-house software development capability, the research team looked at in-house development in the context of this discussion as a potential responsibility of the L13A project contractor.

Software and Storage as a Service

SaaS has become a popular deployment model for certain software applications. It is based on an on-demand, pay-as-you-go model that eliminates up-front acquisition costs and variable operational expenses. SaaS is often talked about in the context of cloud computing, in which various computing services are made available to the user from the cloud, which is a metaphor for the Internet.

A recent development is the availability of cloud storage services from vendors such as Amazon that are based on a similar pay-as-you-use model. These services have addressed enough of the issues and risks relative to security, integrity, availability, and quality of service to be considered for a wide range of storage applications, and there is considerable interest in the archiving community in using cloud storage services as part of a long-term preservation strategy.

Hosting

Hosting is generally understood to mean the operation and maintenance of a computer system on someone's behalf as a commercial service. It is a means of deploying and managing software and hardware, whether COTS, OSS, commodity, or proprietary. It is discussed here because it may be applicable to the proposed archival system and because it is important to differentiate between hosting and SaaS/cloud storage.

With the latter, the customer (at least theoretically) enjoys cost savings because he or she is using a small fraction of a massive, Internet-scale technology deployment. With hosting, provisioning of hardware is more fixed and hardware is often dedicated to a customer, particularly if the computer and storage requirements are significant. Generally speaking, a customer has more freedom of choice in software components and configuration in hosting than in SaaS and cloud storage.

Pros and Cons of These Approaches

When the research team considered these different approaches to technology implementation against the backdrop of the institutional framework in which the proposed Reliability Archive is to be deployed, the following conclusions were reached:

- In-house software development should be considered only as a last resort and only for limited functionality where the need is short-term. It cannot be considered for core functionality that must be sustainable over the life of the archival system.
- Community-supported OSS should be considered only in similar circumstances, since it generally requires developing significant in-house expertise to implement and support it.
- COTS software (which may or may not include OSS components) seems to be the most attractive option for the application and infrastructure software portion of the system because of the availability of commercial support services, eliminating the burden and issues that arise with self-support of either in-house-developed or community-supported OSS.
- Cloud storage is a solution component that should be considered if for no other reason than the cost of acquiring and managing storage (including replacing the hardware on a 3–5-year basis), which is likely to be the single largest cost over the system's lifetime.
- Hosting should also be considered as a technology deployment option principally because it offloads certain operational burdens.

Solution Framework

The visioning and filtering process the research team went through led to the conceptual solution framework (see Figure 3.22). This builds on the functional blocks concept introduced at the beginning of this section and maps it against the implementation options the team judged to be most viable.

The research team identified two different classes of COTS application software that meet most of the end-to-end functional requirements of the proposed system (with respect to submission, appraisal, ingestion, and data and content management) and included them in the analysis. The team also identified an OSS tool that could be used to meet a short-term need and could play a role in a simple and straightforward alternative. The team did not consider alternatives that had any long-term dependence on community-supported OSS for the application or storage tiers.

With respect to the archival storage tier, the research team generally found that the application-level software abstracted

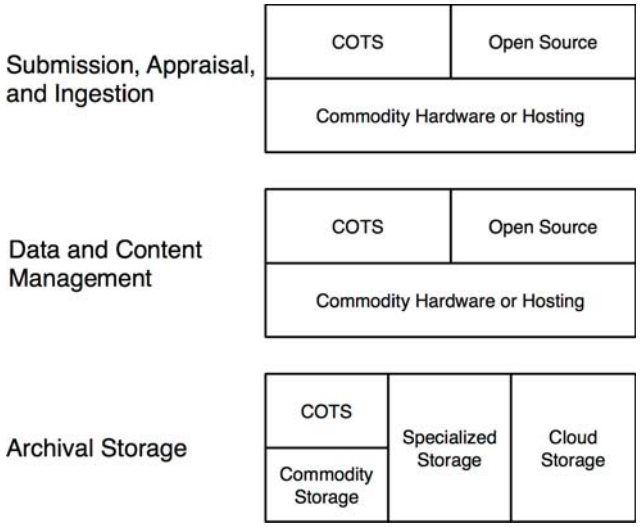


Figure 3.22. Solution framework.

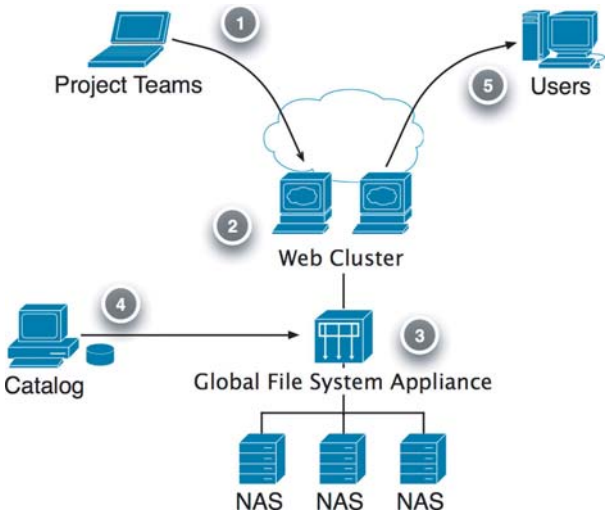


Figure 3.23. Alternative 1 concept.

the interface to the storage tier, thus allowing certain freedom of choice for the archival storage technology. As noted, the expected data volume in the archive will impose practical constraints on the storage options that can be considered.

While the archival storage tier may use commodity *components* such as SATA drives, the particular requirements of an archival storage *system* dictate specialized capabilities. With hosted storage, this specialization comes in the form of the embedded software that runs in the filer or storage controller that virtualizes the underlying generic storage devices, makes these subsystems largely self-managing and highly reliable, and facilitates managing physical migrations. In addition, the emergence of robust cloud storage services provides a viable option for archival storage. Using this framework, the research team proposed a number of alternative system solutions, which are described in the following section.

Alternative 1

The research team strove to find an alternative that might be described as the bare minimum, meaning that it would be simple and straightforward to implement and meet the minimum, essential requirements to be considered a viable solution (see Figure 3.23). This alternative is based on the use of a hierarchical file system to organize the files from each research project. A directory structure that follows basic naming conventions would establish an implied taxonomic hierarchy.

The system is based on simple building blocks and manual processes. The major elements of the system and the workflow through it are as follows:

1. Research project teams would be given log-in credentials and access to specific directories in the archive mapped to their respective projects. For example, the L03 team might have access to the directory \\root\reliability\L03. They would follow prescriptive guidelines to organize their content locally, and then transfer the files to the appropriate subdirectories in the archive using readily available FTP (file transfer protocol) client software.
2. The "web cluster" in this system is simply two commodity servers (for redundancy) that provide an FTP service to accept submissions and an HTTP service to support user access, which is discussed in point 5.
3. Archival storage in this alternative would be provided by self-hosted network-attached storage (NAS). NAS uses a special-purpose computer, sometimes called a filer, to provide file-based disk services on a network. The filer's file volumes are made visible as network shares.

   The maximum size of a disk volume that a modern, general-purpose NAS can export to the network is 16 TB. (This is the current maximum volume size of NAS market leader Network Appliance (NetApp), which is representative of this class of storage device.) Presenting the archival storage space as multiple volumes is undesirable from both a manageability and user-access point-of-view. Presenting it as a single namespace requires the use of a more specialized class of NAS or the insertion of Global Namespace technology, or both, usually in the form of an appliance, between the web server and the NAS filers. With a Global Namespace, users access a virtualized file system namespace where the files exist in multiple volumes but appear to be part of a single namespace.

   There are numerous options available from vendors such as EMC, Hitachi Data Systems, HP, Network Appliance, Sun, and others that can address this storage challenge. The point of the preceding discussion is simply to frame the issue for the purposes of the current analytical task. The storage requirements inform the class of storage

Name	Size	Kind
Capacity	--	Folder
Reliability	--	Folder
Focus Area Overview	--	Folder
ReliabilityBrief.pdf	116 KB	Portable Document Format (PDF)
ReliabilityResearchPlan2008.pdf	132 KB	Portable Document Format (PDF)
Project L03	--	Folder
Aggregated Data	--	Folder
Background Info	--	Folder
Final Report	--	Folder
Interim Reports	--	Folder
Raw Data	--	Folder
Research Plans & Methodology	--	Folder
Research Products	--	Folder
Project L13	--	Folder
Renewal	--	Folder
Safety	--	Folder

Figure 3.24. Browsing a hypothetical directory hierarchy.

that will be required and allow us to estimate storage acquisition, maintenance, and operations costs commensurate with this class of storage.

4. Institutional staff would use an OSS tool such as the Archivist Toolkit (AT) to catalog the files deposited into archival storage. A tool such as AT basically provides a form-based system to catalog descriptive metadata (sometimes called writing a finding aid) and then export it in various standard formats, such as METS or EAD (Electronic Archival Description). These exported files could be transformed by style sheets into static HTML pages to provide a simple, structured way to browse the file system. The concept is to manually publish fixed, top-level maps of the contents of the various subdirectories that users might browse, as described in the next point.
5. Access to this system would be based mainly upon directory browsing, a capability supported by all web servers, whereby a user types a URL into their browser and is permitted to view and navigate a list of files and directories instead of viewing an HTML page. It is very much like using Windows Explorer or Mac's Finder to browse local disks and network shares; it is simply done using a browser and accessing the archival system's directory structure over the Internet.

The primary user interface experience might look something like Figure 3.24.

Alternative 2

The second alternative is based on digital object repository management software designed for universities, libraries, museums, archives, and information centers (see Figure 3.25). This alternative was selected because the functionality provided by these software suites maps very closely to the functional requirements and conceptual design of the archival system as presented in Chapter 2.

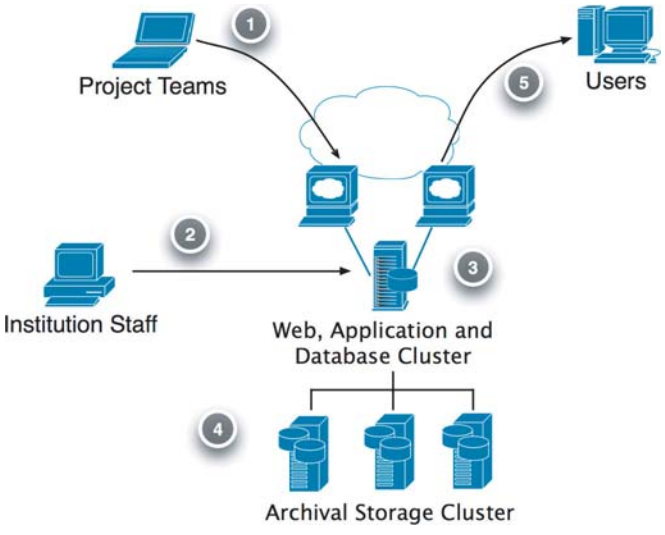


Figure 3.25. Alternative 2 concept.

The following systems enable institutions to manage digital entities end-to-end, from submission through access, while ensuring their integrity over time through continuous preservation actions:

1. Research project teams would submit content into the repository through a web-based interface. These systems generally employ configurable, form-based templates that allow publishers to upload files, enter metadata, and define access restrictions.
2. Review stages involving configurable automatic, semi-automatic, and manual workflows can be integrated to ensure that institutional staff has the ability to edit, delete, or approve submitted content prior to ingestion into the repository.
3. These applications are designed to manage any content type and typically have a very flexible metadata schema. Metadata is encapsulated with its associated content, usually in standard format such as METS, thus constituting a self-contained and self-describing package that is maintained in archival storage (see 4).

   A relational database (RDBMS) such as Oracle is typically used as an operational or runtime database to cache metadata and support web-based publishing and access processes. A key consideration from a sustainability and long-term preservation standpoint is that the runtime database can be rebuilt from metadata embedded in digital objects.

   The web, application, and database cluster is a small number of self-hosted commodity servers that run the application suite; that is, the processes related to submission, appraisal, ingestion, and data and content management.
4. Digital objects themselves are stored in self-hosted archival class storage under a write-once, read-only policy with object replication to ensure their security and integrity over time.
5. Researchers and practitioners would access the repository from a public access portal functionality that is built into these products. Web publishing is automatic and dynamically driven from the repository's metadata. The look-and-feel of the interface would be customized via HTML, CSS, and XML/XSL, and the user experience would be more akin to that depicted in the mock-ups.

   Users would be able to navigate the repository content through fixed and dynamic classification paths (menus), as well as perform full-text and faceted searches.

These systems support user self-registration and various authentication schemes and enforce access control restrictions that are encoded in the administrative metadata.

Alternative 3

This alternative is based on the same class of COTS software as Alternative 2; however, the system implementation differs substantially, as depicted in Figure 3.26.

The system functionality and topology, in the first three of the following points, are identical to Alternative 2:

1. Research project teams would submit content into the repository through a web-based interface. These systems generally employ configurable, form-based templates that allow publishers to upload files, enter metadata, and define access restrictions.
2. Review stages involving configurable automatic, semi-automatic, and manual workflows can be integrated to ensure that institutional staff has the ability to edit, delete, or approve submitted content before to ingestion into the repository.
3. These applications are designed to manage any content type and typically have a very flexible metadata schema. Metadata is encapsulated with its associated content, usually in standard format such as METS, thus constituting a self-contained and self-describing package that is maintained in archival storage (see 4).

   A relational database (RDBMS) such as Oracle is typically used as an operational or runtime database to cache metadata and support web-based publishing and access processes. A key consideration from a sustainability and long-term preservation standpoint is that the runtime database can be rebuilt from metadata embedded in digital objects.
4. The web, application, and database cluster is a small number of self-hosted commodity servers that run the application suite; that is, the processes related to submission, appraisal, ingestion, and data and content management.

   In this alternative, instead of residing in self-hosted storage, the archived data is preserved using a cloud storage service.

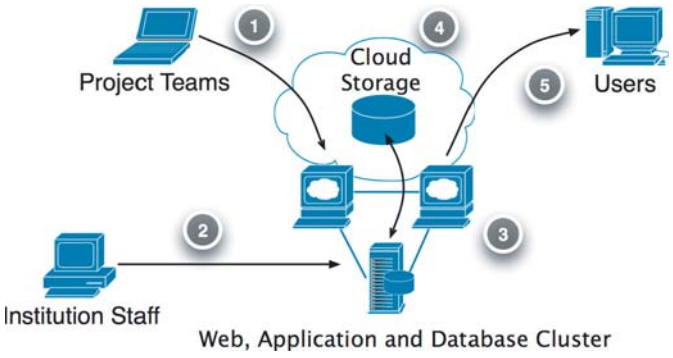


Figure 3.26. Alternative 3 concept.

5. Once submitted data have been appraised and approved for ingestion, the metadata-wrapped digital object is written to a cloud storage service. While it is beyond the scope of this document to describe cloud storage services in detail, a few highlights are worth noting.

   Cloud storage services do not operate like file systems or network-attached storage (NAS), which are mounted or mapped as either physical or virtual disks. Instead, they store and retrieve files via a simple web service (ReST: Representational State Transfer) interface, in essence, providing an object-based storage service. An object is stored and retrieved using a persistent identifier over encrypted communications in conjunction with a session authentication token. Each stored object is replicated within the storage cloud for high availability and fault tolerance (three ephemeral copies of an object is typical of these services). At many levels, the model maps well to archival storage requirements.

User access to the system is exactly as described for Alternative 2, except that the digital object repository management software, in its role as trusted intermediary to archived data, retrieves the requested object(s) from a cloud storage service instead of from a self-hosted storage.

Other Alternatives Considered

The research team considered an alternative solution based upon a category of COTS application software called Enterprise Content Management (ECM). AIIM (Association for Information and Image Management) defines ECM (*8*) as "the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists."

ECM systems provide a range of functions, which typically encompass at least the following areas:

- Document management: Organize documents into hierarchies of files and folders or compound documents; classify documents by adding metadata; manage document check-in, check-out, and versioning; manage change request, review, and approval workflows;
- Records management: Manage document retention and disposition through system-enforced rules;
- Digital asset management: Manage digital media and related metadata to support workflows around image, audio, and video file types; and
- Image management: Provide paper, fax and e-mail capture, recognition, and routing.

Representative products of this class of software include Documentum from EMC, ECM Suite from Open Text, Filenet from IBM, and Oracle UCM (Universal Content Management).

The research team considered this alternative because it is a well understood and proven means of managing electronic content within certain functions of some organizations.

Upon cursory examination, ECM systems seem to map well to the functional requirements of the Reliability Archive. On deeper examination, a number of key differences or optimizations become apparent when compared to digital object repository management software. In general, the following characteristics are typical of ECM systems:

- They are optimized for integration into the workflow of existing operational systems instead of being built for stand-alone use.
- They have more fixed metadata schemas.
- They are primarily document-centric but can manage other content types.
- They are typically deployed internally—i.e., behind a firewall on a corporate Intranet. Web publishing for public access involves add-on products, more hardware, and additional workflows.
- They are more complex and cumbersome to implement and maintain, and impose much more application software dependency.

An ECM system represents conventional thinking about content-centric applications typical of major technology vendors, many of whom have ECM software in their product offerings. These vendors have natural incentives to steer customers toward solutions they control and that drag substantial service revenue. They naturally would propose an ECM solution if given the opportunity.

An ECM system would have a virtually identical storage requirement to a digital object repository management system. Software acquisition costs would be significantly higher, as would system integration costs.

For these and other reasons the research team decided not to recommend this alternative or analyze its life-cycle costs and benefits.

Life-Cycle Costs Analysis

This chapter includes 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.

Assumptions

The following assumptions are used to support the cost–benefit analysis on the three selected alternatives:

1. The length of life cycle: This is the time from the beginning of a system's implementation project to the retire-

ment and replacement of that system. It includes the time during which the system will be operational as well as the time needed to develop and implement the system. Using the requirements from the L13 RFP, the research team calculated the life-cycle cost for a period of 25 years.
2. Cost distribution: Costs were estimated for each of the first 5 years (including initial acquisition costs) and then for 5-year increments for the next 20 years.
3. Base year: Following the current SHRP 2 Reliability program schedule, the research team used 2010 as the beginning of the system's life cycle.
4. Initial period: Following the current Reliability program plan, the research team assumed that the Reliability Archive would be implemented over a 2-year time frame from 2010 to 2011 and that the system would be in production in 2012.
5. A discount rate was used to relate present and future dollars. It is expressed as a percentage and used to reduce the value of future dollars in relation to present dollars. A discount rate of 5% was used for the analysis.
6. The defined alternatives represent types of solutions rather than specific products. Therefore, the cost of future selected products may differ from the estimated costs. However, such differences are not expected to have significant impact on the relevance or comparability of the alternatives.
7. The life-cycle cost considered in this analysis includes costs associated with initial acquisition, operations, and maintenance as well as periodic or occasional upgrades to accommodate technology advances and obsolescence.
8. The cost for the Reliability contractors to enter their project data into the archival system is not included in this analysis. Reliability Project L16 covers that effort.

Data Sources

The research team used cost information from a wide variety of sources, including the following:

- Vendors' websites and other sources in the public domain, including online configuration tools and price lists;
- Informal contacts with vendors; and
- The team members' experience and knowledge.

Cost Elements

The life-cycle cost considered in this analysis includes the costs associated with initial acquisition, operations, and maintenance as well as periodic or occasional upgrades to accommodate technology advances and obsolescence.

Figure 3.27 shows the cost breakdown structure the research team developed for the proposed alternative solutions. It takes into consideration the technical characteristics of these alternatives as analyzed earlier, as well as the current SHRP 2 Reliability program plan.

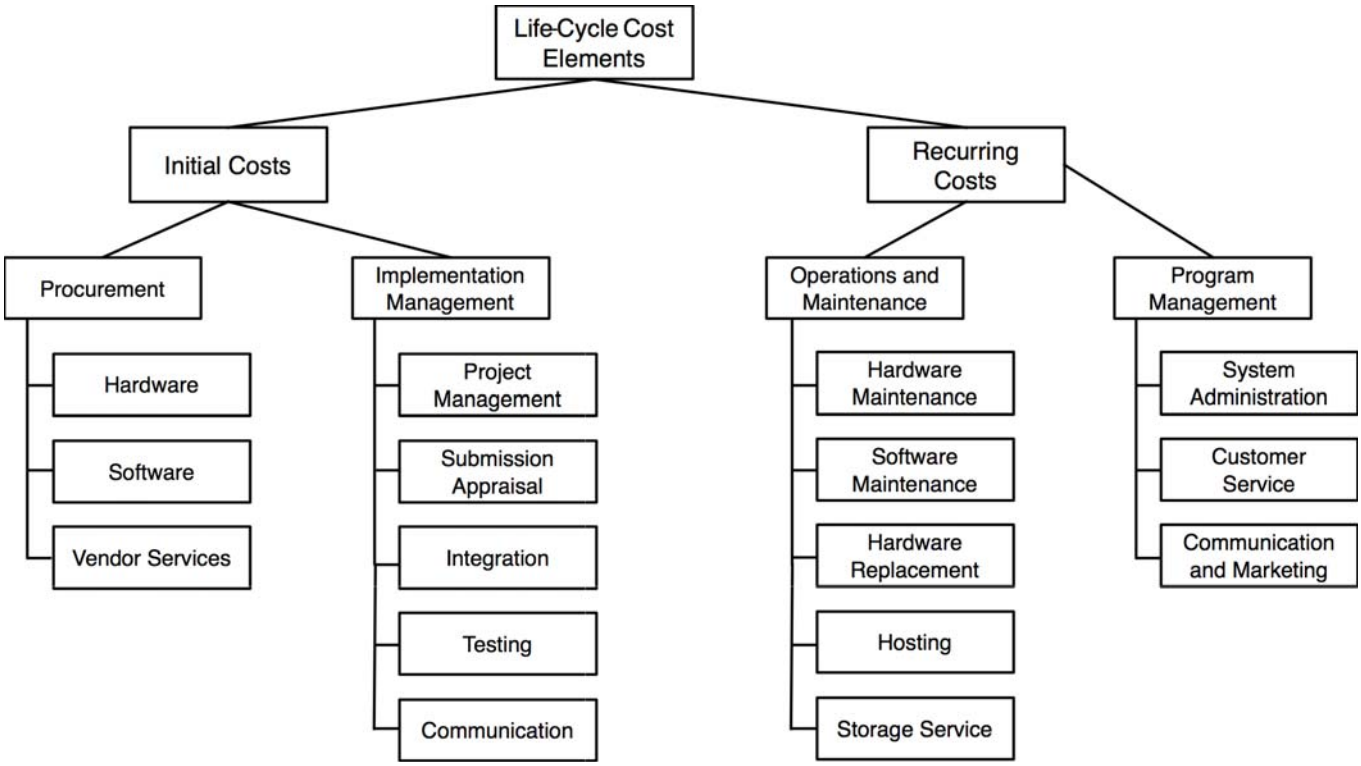


Figure 3.27. Life-cycle cost elements of Reliability Archive solution alternatives.

Initial Costs or Nonrecurring Costs

The initial costs incur during the first two years on a one-time basis. These nonrecurring costs represent the capital investment from the SHRP 2 Reliability program and should be closely tied to the budget of Reliability Project L13A.

The initial costs of the L13A Archival system can be grouped into procurement costs and program management costs.

Procurement Costs

The procurement costs for the Reliability Archive solution may include the following items, depending upon the alternative:

- Hardware: The cost to procure necessary hardware—i.e., servers, workstations, networking, and storage;
- Software: The cost to license COTS software; and
- Vendor services: The cost for the selected vendors to work with the SHRP 2 Reliability program staff to implement their solutions. It is anticipated that their services will include installation, customization, testing, and deployment.

Program Management Costs

The program management costs represent the effort of overseeing the entire Reliability Archive implementation and working with the selected vendors to ensure that their services and products are properly implemented to fully satisfy the Reliability Archive requirements. The program management team will represent the SHRP 2 Reliability program and ensure that the SHRP 2 program interests are best protected and realized.

The following are the components of the Reliability Archive program management costs:

- Project management: The cost to manage the implementation of the Reliability Archive, including schedule monitoring, task execution, and working with the vendors on a day-to-day basis.
- Submission appraisal: The cost to evaluate and appraise the submissions from individual Reliability project teams so that the information can be properly archived and the metadata can be encoded.
- System integration: A key part of the program management team's efforts is to ensure that all components of the solution—i.e., submission, storage, metadata management, and content management—are properly integrated. This is the cost of these efforts.
- System testing: The cost of performing acceptance tests on the solutions implemented by the vendors to ensure that all requirements are fully satisfied.
- Marketing and communication: The cost of communication with user communities on services provided by the Reliability Archive. The efforts will include newsletters, project website, and conference presentations.

Estimation of the program management cost was based on a basic project management team that includes roles such as project manager, technology specialists, archivists, analysts/developers, and quality assurance. The level of effort may vary from one alternative to another, depending on complexity.

Recurring Costs

Recurring costs are the continuing costs associated with the management and operation of the archival system. Recurring costs apply over a period of time throughout the system's life. In this analysis most of the recurring costs are incurred over a period of 23 years from 2012 to 2035. Those that also apply during the initial period follow.

System Operations and Maintenance Costs

The recurring system operations and maintenance costs for the L13 archival solution include the following cost items:

- Hardware maintenance: The cost to troubleshoot, replace, or repair hardware. This cost typically begins to accrue 90 days after hardware installation, so it must be accounted for during the initial period as well as over the operational life of the archival system.
- Hardware upgrade and replacement: The cost to regularly upgrade or replace acquired hardware to accommodate obsolescence, advances in technology, and growth in number of users. This cost is assumed to be incurred every 5 years.
- Software upgrade: The cost associated with software upgrades and replacement. This cost is also assumed to be incurred every 5 years.
- Software maintenance: The cost of obtaining product support and access to software fixes and updates from the vendor. This cost typically begins to accrue from 90 days to one year after software installation, so it must be accounted for during the initial period as well as over the operational life of the archival system.
- Hosting: The cost to house, power, cool, and physically maintain any archival system hardware, whether a fee from a commercial service or a chargeback from an implementation agent. These costs accrue once any hardware is installed, so they have to be accounted for during the

initial period as well as over the operational life of the archival system.
- Storage service: The usage cost of a commercial cloud storage service, such as Amazon S3. The cost is based on data storage capacity used, plus the amount of data transfer in and out of the service. The estimates of this cost are based upon the expected data capacity growth over time. Initial data transfer expense will relate to ingestion into the archive and later to data downloaded by users.

Program Management Costs

The success of the Reliability Archive implementation will depend on continued program support. This warrants a small-scale focus team dedicated to the support task. The following, then, are the estimated cost items for program management:

- System administration: The cost of administering, managing, and monitoring the operations of the archiving system on a daily basis.
- Customer service: The cost of providing services to address the needs or issues users encounter in using the system.
- Marketing and communication: The cost of promoting the services of the archival system. The typical efforts will include newsletters, conference presentations, and coordination with other programs.

Life-Cycle Costs of the Alternatives

The life-cycle costs of the three alternatives are summarized below. Worksheets that provide supporting details behind the initial and recurring costs can be found in Appendix B.

The tables in this section show the cost breakdown for a 25-year life cycle. Costs are shown on an annual basis for the first 5 years and thereafter in 5-year intervals. Shaded areas represent cost items that are not applicable to the specific life-cycle periods.

Alternative 1

This bare minimum alternative (see Table 3.3) focuses on preserving the data and providing a minimally acceptable level of user access. The level of manual effort involved in system implementation accounts for these costs being the highest among the three alternatives.

Alternative 2

The second alternative (see Table 3.4) has essentially the same storage-related costs as Alternative 1 but adds licensing costs for COTS application software and system software (e.g., RDBMS) that deliver considerably more functionality than Alternative 1. Estimated system implementation costs are lower because much of the effort will involve configuring out-of-the-box functionality. Other hardware costs are marginally higher because more servers are required to run the application functionality that is not present in Alternative 1.

Alternative 3

The final alternative (see Table 3.5) would offer the same functionality as Alternative 2, but with no cost over the life cycle of the system for procurement, installation, maintenance, and replacement of storage hardware. Estimated system administration costs are also lower because there is no storage hardware to manage.

Table 3.3. Alternative 1 Life-Cycle Cost Summary

Discount Factor 5%	2010 Year 1	2011 Year 2	2012 Year 3	2013 Year 4	2014 Year 5	2015–2019 Years 6–10	2020–2024 Years 11–15	2025–2029 Years 16–20	2030–2035 Years 21–24
Initial Cost ($)	**$1,002,600**								
Hardware	$515,000								
Software	$0								
Implementation	$243,800	$243,800							
Recurring Cost ($)	**$108,000**	**$278,000**	**$448,000**	**$448,000**	**$958,000**	**$2,750,000**	**$2,750,000**	**$2,750,000**	**$2,750,000**
Annual Costs									
System Administration	$—	$170,000	$170,000	$170,000	$170,000	$850,000	$850,000	$850,000	$850,000
System Maintenance	$102,000	$102,000	$102,000	$102,000	$102,000	$510,000	$510,000	$510,000	$510,000
Marketing and Customer Services	$—	$—	$170,000	$170,000	$170,000	$850,000	$850,000	$850,000	$850,000
Hosting	$6,000	$6,000	$6,000	$6,000	$6,000	$30,000	$30,000	$30,000	$30,000
Storage Service (Capacity and Access)	$—	$—	$—	$—	$—	$—	$—	$—	$—
Periodic Costs									
Software Upgrade					$—	$—	$—	$—	$—
Hardware Upgrade					$510,000	$510,000	$510,000	$510,000	$510,000
Summary									
Total Cost ($)	$1,110,600	$278,000	$448,000	$448,000	$958,000	$2,750,000	$2,750,000	$2,750,000	$2,750,000
Number of Periods (Years)	1	2	3	4	5	6	11	16	21
Total Cost ($, Present Value)	$1,057,714	$252,154	$386,999	$368,571	$750,618	$2,052,092	$1,607,868	$1,259,807	$987,092
Total Initial Cost (Present Value, first two years 2010 and 2011)	**$1,309,868**								
Total Life Cycle Cost (Present Value, 23 years from 2012 to 2035)	**$7,413,047**								

Table 3.4. Alternative 2 Life-Cycle Cost Summary

Discount Factor 5%	2010 Year 1	2011 Year 2	2012 Year 3	2013 Year 4	2014 Year 5	2015–2019 Years 6–10	2020–2024 Years 11–15	2025–2029 Years 16–20	2030–2034 Years 21–25
Initial Cost ($)	**$1,041,850**								
Hardware	$525,000								
Software	$170,000								
Implementation	$173,425	$173,425							
Recurring Cost ($)	**$140,000**	**$310,000**	**$480,000**	**$480,000**	**$1,000,000**	**$2,920,000**	**$2,920,000**	**$2,920,000**	**$2,920,000**
Annual Costs									
System Administration	$—	$170,000	$170,000	$170,000	$170,000	$850,000	$850,000	$850,000	$850,000
System Maintenance	$134,000	$134,000	$134,000	$134,000	$134,000	$670,000	$670,000	$670,000	$670,000
Marketing and Customer Services	$—	$—	$170,000	$170,000	$170,000	$850,000	$850,000	$850,000	$850,000
Hosting	$6,000	$6,000	$6,000	$6,000	$6,000	$30,000	$30,000	$30,000	$30,000
Storage Service (Capacity and Access)	$—	$—	$—	$—	$—	$—	$—	$—	$—
Periodic Costs									
Software Upgrade					$—	$—	$—	$—	$—
Hardware Upgrade					$520,000	$520,000	$520,000	$520,000	$520,000
Summary									
Total Cost ($)	$1,181,850	$310,000	$480,000	$480,000	$1,000,000	$2,920,000	$2,920,000	$2,920,000	$2,920,000
Number of Periods (Years)	1	2	3	4	5	6	11	16	21
Total Cost ($, Present Value)	$1,125,571	$281,179	$414,642	$394,897	$783,526	$2,178,949	$1,707,264	$1,337,686	$1,048,112
Total Initial Cost (Present Value, first two years 2010 and 2011)	**$1,406,751**								
Total Life Cycle Cost (Present Value, 23 years from 2012 to 2035)	**$7,865,075**								

Table 3.5. Alternative 3 Life-Cycle Cost Summary

Discount Factor 5%	2010 Year 1	2011 Year 2	2012 Year 3	2013 Year 4	2014 Year 5	2015–2019 Years 6–10	2020–2024 Years 11–15	2025–2029 Years 16–20	2030–2034 Years 21–25
Initial Cost ($)	**$541,850**								
Hardware	$25,000								
Software	$170,000								
Implementation	$173,425	$173,425							
Recurring Cost ($)	**$71,200**	**$192,200**	**$414,200**	**$414,200**	**$434,200**	**$2,091,000**	**$2,091,000**	**$2,091,000**	**$2,091,000**
Annual Costs									
System Administration	$—	$85,000	$85,000	$85,000	$85,000	$425,000	$425,000	$425,000	$425,000
System Maintenance	$34,000	$34,000	$34,000	$34,000	$34,000	$170,000	$170,000	$170,000	$170,000
Marketing and Customer Services	$—	$—	$170,000	$170,000	$170,000	$850,000	$850,000	$850,000	$850,000
Hosting	$1,200	$1,200	$1,200	$1,200	$1,200	$6,000	$6,000	$6,000	$6,000
Storage Service (Capacity and Access)	$36,000	$72,000	$124,000	$124,000	$124,000	$620,000	$620,000	$620,000	$620,000
Periodic Costs									
Software Upgrade					$—	$—	$—	$—	$—
Hardware Upgrade					$20,000	$20,000	$20,000	$20,000	$20,000
Summary									
Total Cost ($)	$613,050	$192,200	$414,200	$414,200	$434,200	$2,091,000	$2,091,000	$2,091,000	$2,091,000
Number of Periods (Years)	1	2	3	4	5	6	11	16	21
Total Cost ($, Present Value)	$583,857	$174,331	$357,802	$340,763	$340,207	$1,560,336	$1,222,564	$957,911	$750,548
Total Initial Cost (Present Value, first two years 2010 and 2011)	**$758,188**								
Total Life Cycle Cost (Present Value, 23 years from 2012 to 2035)	**$5,530,132**								

References

1. Ganz, J. F., D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz. *The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010.* IDC, Framingham, Mass., March 2007. www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf. Accessed March 2, 2011.
2. Peterson, M., G. Zasman, P. Mojica, and J. Porter. *100 Year Archive Requirements Survey.* Storage Network Industry Association, San Francisco, Calif., Jan. 2007. www.snia.org/forums/dmf/programs/ltacsi/forums/dmf/programs/ltacsi/100_year/100YrATF_Archive-Requirements-Survey_20070619.pdf. Accessed March 2, 2011.
3. *Reference Model for an Open Archival Information System.* Recommendation for Space Data System Standards (OAIS). Consultative Committee for Space Data Systems 650.0-B-1. Blue Book. Issue 1. Washington, D.C., Jan. 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf. Accessed March 2, 2011.
4. Library of Congress. Metadata Encoding & Transmission Standard. www.loc.gov/standards/mets. Accessed March 2, 2011.
5. *Special Report 296: Implementing the Results of the Second Strategic Highway Research Program—Saving Lives, Reducing Congestion, Improving Quality of Life.* Transportation Research Board of the National Academies, Washington, D.C., 2009. http://onlinepubs.trb.org/Onlinepubs/sr/sr296.pdf. Accessed March 2, 2011.
6. Battelle. *Traffic Data Quality Measurement.* Final Report, BAT 03-007, Federal Highway Administration, Sept. 15, 2004. http://isddc.dot.gov/OLPFiles/FHWA/013402.pdf. Accessed March 2, 2011.
7. Codd, E. F. A Relational Model of Data for Large Shared Data Banks, *Communications of the ACM,* Vol. 13, No. 6, June 1970, pp. 377–387.
8. Association for Information and Image Management. What is Enterprise Content Management (ECM)? www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx. Accessed March 3, 2011.

CHAPTER 4

Conclusions

Final Recommendations

The research team recommends that, when evaluated against the criteria of requirements met, conformity with the conceptual design, initial and life-cycle costs, benefits to stakeholders, risk mitigation, and schedule, the SHRP 2 program proceed with the L13A Reliability Archive project based on the solution Alternative 3 approach. The following sections justify this recommendation based on each of these criteria.

Justification Based on Requirements

The research team used a simple 3-point scoring method (2 = meets/exceeds requirement; 1 = minimally meets requirement; 0 = does not meet requirement) to evaluate how well each alternative met the system's requirements. The results, summarized in Table 4.1, show that Alternative 3 best meets the system requirements. The detailed scoring worksheet can be found in Appendix C. With the exception of the general "Systemwide" category, the requirements are categorized based on functions of an archival system described by the OAIS model.

Because many of the requirements are based on functionality provided by application software, Alternative 1 scored the lowest. Alternatives 2 and 3 are more comparable because both use the same digital repository management software. Alternative 3 ultimately scored the highest because it meets many technical requirements while obviating the need to manage technology complexity.

In cases where it was difficult to evaluate the alternatives without knowing the exact product being assessed, the alternatives were scored identically. The team believes these detailed requirements will remain useful when making specific product selections should the SHRP 2 program decide to move ahead with Reliability Project L13A.

Justification Based on Conceptual Design

Assessing how well the alternatives conform to the conceptual design is more subjective than assessing requirements on a function-by-function basis. OAIS describes the roles of producers and consumers, along with the functionality users in these roles will expect from an archival information system. Producers will require ways to organize, package, submit, assess, and classify information, whereas consumers will require ways to find and access information to which they had been granted access.

Together, the six OAIS core archival information system functions of ingest, data management, archival storage, access, administration, and preservation planning are responsible collectively for preserving the collection of digital artifacts, monitoring and ensuring their integrity through physical migrations and format transformations, maintaining their physical security, facilitating information discovery, and enforcing access control.

The availability of COTS software that is built on OAIS principles and concepts clearly influenced the research team's thinking about solution alternatives. Because this class of digital repository management software provides broad, out-of-the-box coverage of the needs of producers, consumers, as well as the core archival information system functions, the team judged Alternatives 2 and 3 to be clearly superior to Alternative 1 in terms of conformity to the conceptual design.

Because archival storage is a central and indispensable part of an open archival information system, it follows that, so long as the system can meet its data preservation mandate, the ultimate value of the system is best measured in terms of the service delivered to users. In this respect, the research team believes that Alternative 3 has the advantage.

Cloud storage is part of a larger and rapidly growing trend toward a style of utility computing where services are provided by and accessed through the Internet. While the tangible benefits manifest themselves most clearly in terms of costs and simplified management, as will be discussed next, the research team believes that there is a significant long-term advantage to a deployment strategy that focuses primarily on service delivery and less on technology management. The spirit of the conceptual design is that preserving and curating

Table 4.1. Requirements Scoring Summary

	Alternative Scores		
Category	**1**	**2**	**3**
Producers	0	6	6
Ingestion	4	10	11
Archival storage	20	23	32
Data management	1	14	14
Preservation planning	26	31	37
Administration	12	15	15
Access	3	7	7
Consumers	4	18	18
Systemwide	16	16	28
Totals:	86	140	168

the collection so that users benefit from it is what matters most. This is why the team gives Alternative 3 the ultimate edge in this category.

Justification Based on Cost

Table 4.2 summarizes the life-cycle cost analysis of the three alternatives, details on costs for which can be found in Appendix B. The table illustrates that the life-cycle cost analysis strongly favors Alternative 3.

Lowest Initial Cost

Alternative 3 is about 37% below the $1.2 million budget threshold that SHRP 2 has stipulated for the development and implementation of the L13A Archive. Both Alternative 1 and Alternative 2 are about 15% above this budget limit. This is primarily because Alternative 3 is designed to use the commercial cloud storage services. Because of this advantage, Alternative 3 is able to avoid large, up-front capital investment costs. At the same time, Alternative 3 also incurs lower system maintenance cost during the first 2 years. As a result, Alternative 3 poses minimum risk to the SHRP 2 program.

Lowest Recurring Cost

Not only does it exhibit the lowest initial cost, but Alternative 3 also has the lowest life-cycle cost among the alternatives. Both Alternative 1 and Alternative 2 will incur much higher system maintenance costs and periodic hardware upgrade costs during the system's life. The largest portion of Alternative 3's recurring costs is related to the use of cloud storage services. The research team's estimate is based on the current pricing structure of Amazon's S3 cloud storage service. Since cloud computing is a disruptive trend in information technology, the team expects that the cloud storage price will become even more competitive in the future, thus leading to lower life-cycle costs than the current estimate for Alternative 3.

In summary, based on its lowest initial and life-cycle costs, Alternative 3 is the most cost-effective solution.

Justification Based on Benefits

In a benefit analysis, benefits are usually defined as either quantitative and tangible or qualitative and intangible improvements expected or resulting from a system investment. Tangible benefits are defined as benefits that can be expressed in terms of monetary value. They typically represent direct revenue to be received during the life cycle of the investment. The intangible benefits are those that are qualitative in nature and cannot be ascribed monetary value directly.

Similar to many information technology investments, the Reliability Archive project faces the following typical challenges in assessing benefits:

- Difficulty in identifying benefits that do not have an obvious market value or price; and
- Difficulty in quantifying the value of benefits that do not directly accrue to the investment in the project.

To address these challenges, the research team assessed the relative benefits of the alternatives from the perspective of

Table 4.2. Life-Cycle Costs Summary

Cost	Description	Alternative 1	Alternative 2	Alternative 3
Total initial cost	Present value: First 2 years, 2010 and 2011	$1,309,868	$1,406,751	$758,188
Total life-cycle cost	Present value: 23 years from 2012 to 2035	$7,413,047	$7,865,075	$5,530,132
Average annual life-cycle cost	Present value: 23 years from 2012 to 2035	$322,306	$341,960	$240,441

the parties who will benefit from the implementation of the Reliability Archive, as well as how well these alternatives can support the implementation strategies recommended in the recent SHRP 2 implementation report that was prepared for Congress (*1*).

Of course, the relative benefits of each alternative can also be assessed with respect to cost.

Benefits to the SHRP 2 Program

The entire SHRP 2 program will benefit from the implementation of the L13A Archive. The benefits in this category can be assessed with respect to long-term data preservation, sharing of system capabilities across projects and programs, and how the alternatives are best positioned to support the implementation of the SHRP 2 program results.

Initial Cost. From the initial investment perspective, Alternative 3 provides SHRP 2 with a huge benefit. The initial cost of Alternative 3 is significantly below the $1.2 million budget constraint, while Alternative 1 and Alternative 2 are both above it. Alternative 3 clearly helps SHRP 2 avoid a large investment made early in the project prior to system build-out, and well before the business benefits are realized.

Schedule. The research team estimated the duration of development and implementation for each of the three alternatives in the subsequent section. Overall, Alternative 3 provides the highest likelihood of implementing the L13A Archive within the 18-month period that SHRP 2 specifies because it significantly reduces the amount of time to procure and install the necessary IT environment in order for the L13A Archive to operate. The cloud storage services required for Alternative 3 are readily available and require much less time for configuration.

Long-Term Data Preservation. SHRP 2 requires the Reliability Archive to be available for 20 to 50 years. All of the proposed alternatives are capable of meeting this goal, assuming they are properly managed over the life cycle of the system. Because Alternative 3 shifts a major portion of this management responsibility to a service provider with domain expertise in this area, the research team judges it to be superior to the other alternatives in this regard.

Sustainability. Ensuring that obsolescence is avoided wherever possible and that technology transitions are well managed are keys to the sustainability of the archival system over time. Alternatives 2 and 3 are superior to Alternative 1 in this respect because the application software suites they use provide automated strategies and tools, including support for multiple versions (formats) of a digital object. Alternative 3 has a further advantage by way of virtualization of the most complex technical aspect of the solution, which is the archival storage tier. Offloading responsibility for this to a service provider eliminates the need to periodically refresh storage hardware and manage physical migrations.

Potential Leverage for Other SHRP 2 Programs and Projects. Quite a few projects from other SHRP 2 programs such as Renewal and Safety also collect extensive amounts of data and may eventually require an archive system to preserve their data. Since there is no practical limit to cloud storage capacity, Alternative 3 can easily meet this need, whereas additional hardware would have to be procured, managed, and periodically upgraded for the other two alternatives.

Capacity for Greater Information Sharing. Cloud storage is part of a broader web services platform that is constantly evolving and expanding to offer additional functions, applications, and capacities that enable delivery of a wider array of capabilities to users. As a result, compared to the other two alternatives, Alternative 3 will provide SHRP 2 with more agility and flexibility to increase its data sharing and collaboration capability with other national and regional programs.

Support for Possible Future Institutional Structures and Governance Models. The recent SHRP 2 implementation report suggests several approaches and ideas for building the long-term implementation agent. However, no decision has been made regarding future institutional structures and governance models. The research team expects that, in the interim, the National Academies or TRB will take the responsibility of maintaining the L13A Archive. Alternatives 1 and 2 require the National Academies or TRB to build an extensive IT environment that may not be easy to transition to a future governance structure. By using cloud storage, Alternative 3 has the smallest in-house "footprint" and is relatively neutral to current and future governance models.

Packaging, Branding, and Enabling Research Results to Products. As suggested in the SHRP 2 implementation report, the benefits of some SHRP 2 projects may be optimized if the project results are combined with those of other related research projects into a unified package with unique branding. From an information structure perspective, Alternatives 2 and 3 provide the most capabilities to enable SHRP 2 to achieve this objective because they have the most extensive metadata management functionality. Alternative 3 has a further advantage because under this option the archived data will reside in cloud storage, which, as noted earlier, is part of a broader services platform. This offers the future possibility of easily adding new functionality without incurring capital expense. Some of the on-demand capability available today includes processing and analysis of large data sets, and data integration (e.g., mashup) services.

Benefits to the User Community

As described, the Reliability Archive system will serve a broad range of users. The following are the research team's assessments of the relative user benefits of the three alternatives.

Business Functionality. Alternative 2 and Alternative 3 should provide the same or similar business functions to the future users in terms of using the L13A Archive. This is because both approaches are based on the same type of digital object repository management software. Alternative 1 will provide directory browsing as the only means to access the data, thus its functionality is much more limited.

Follow-On Research, Testing, and Evaluation. The primary purpose of the Reliability Archive is to enable future researchers to test, evaluate, and validate the research results and even to build new research on the existing data. Although the primary function of the L13A Archive is to facilitate access and not to provide tools to conduct this research, it has already been noted how Alternative 3 offers more potential because of the possibility to leverage other capabilities that are part of the broader web services platform, of which cloud storage is a part.

Advanced User Accessibility. For advanced user access such as downloading large amounts of data, Alternative 3 eliminates the need to engineer, procure, and manage infrastructure sufficient to meet peak and occasional demands.

Benefits to Long-Term Implementation Agent

As described, a long-term implementation agent should be established to lead and support the implementation of the Reliability Archive system. Its responsibilities include implementing and deploying the archive system to a production environment and providing long-term stewardship of the system. Benefits for the implementation agent can be assessed on the basis of the relative capacity of the three alternatives to enable better system performance and reliability, and the relative complexity to manage these alternatives over time (see Table 4.3).

Table 4.3. Summary of Benefits by Stakeholder

Benefit Targets	Benefit Aspects	Alternative 1	Alternative 2	Alternative 3
SHRP 2 Program				
	Initial investment under $1.2 million budget	No	No	Yes
	Can be implemented in 18 months	Possible	Possible	Lowest Risk
	Long-term preservation	Yes	Yes	Yes
	Sustainability (avoiding obsolescence, migration management)	Yes, but with highest effort	Yes	Yes, lowest effort
	Potential leverage for other SHRP 2 programs and projects	Minimal	Good	Best
	Capacity for greater information sharing	Minimal	Good	Best
	Support for possible future institutional structures and governance models	Least flexible	More flexible	Most flexible
	Support of program implementation strategy	Minimal	Good	Best
User Community				
	Basic data access and functionality	Minimal	Yes	Yes
	Follow-on research, testing, and evaluation	No	Good	Best
	Advanced user accessibility	No	Good	Best
Long-Term Implementation Agent				
	System administration burden	Highest	Moderate	Lowest
	System maintenance burden	Moderate	Moderate	Lowest
	Recurring cost	Higher	Higher	Lowest
	Internal expertise required	Higher	Higher	Lowest
	Long-term stewardship	Acceptable use of resources	Better use of resources	Best use of resources

System Administration Burden. Alternative 3 greatly reduces the burden on the long-term implementation agent to provide system administration support for the L13A Archive because this alternative manages data storage via cloud storage services.

System Maintenance Burden. By the same token, the burden of system maintenance on the long-term implementation agent will be minimized under Alternative 3 because it eliminates the efforts required for hardware and software upkeep and migration.

Recurring Costs. As analyzed in the previous section, Alternative 3 is expected to incur lower recurring costs than the other two alternatives. Cloud storage services used by Alternative 3 adopt the pay-as-you-go model that requires a low initial investment and additional investments that are usage-based. This will help the long-term implementation agent avoid surge-type costs due to periodic hardware upgrade or replacement and data migration, thereby allowing better budgetary planning based on system usage over time.

Internal Expertise Required. Because of minimized system administration and system maintenance efforts associated with Alternative 3, the long-term implementation agent will be able to significantly reduce its dependency on specialized IT resources while gaining more flexibility in structuring its customer support team.

Long-Term Stewardship. Maintaining an IT operational environment is unlikely to be a part of the core competency of the future long-term implementation agent. By alleviating this IT burden on the agent, Alternative 3 allows the agent to focus on the primary goals of the L13A Archive, which are to preserve SHRP 2 research data and to make them accessible to the users via better program management, innovation, and collaboration with other transportation programs. An added benefit is that part or all of the cost savings in technology-based capital expenditure may be redirected to better serve the primary mission of the L13A Archive as a long-term data preservation and dissemination tool.

Justification Based on Risk Mitigation

Risks for the Reliability Archive exist from both technical and business perspectives. This section enumerates these risks by category and compares the ways the three alternatives mitigate these risks.

Technical Risk

With any technical solution come a variety of risks and dependencies that must be recognized and managed. The following are among the risks associated with the Reliability Archive.

Data Protection. One of the primary goals of the Reliability Archive is the long-term preservation (20 to 50 years) of project-level research information and the accumulated knowledge that accrues from it in the form of research products. All of the alternatives can deliver the necessary level of data protection if managed correctly over the life of the system. Alternative 3 has the lowest risk, however, because it follows best practices for data protection and does not require the National Academies, TRB, and the implementation agent to acquire and maintain sufficient domain expertise in this area to ensure the same level of risk mitigation.

Technology Obsolescence and Migration. It is reasonable to assume that technology will continue to evolve at the current rates, necessitating hardware replacement at 3- to 5-year intervals over the life of the Reliability Archive. This portends a minimum of three hardware migrations over a 20-year service life. Again, Alternative 3 has the lowest risk because its virtualized, network-based storage has intrinsic capabilities to migrate data.

Format Obsolescence and Migration. Obsolescence of file formats over time presents a risk for data loss. Alternatives 2 and 3 are superior to Alternative 1 in that their OAIS-influenced application models support (1) the identification of file formats in the archive that are at risk, and (2) multiple versions of a digital object, allowing for transformation of a soon-to-be-obsolete format to one that is machine- or human-readable.

Security. The requirement to maintain physical security as well as prevent unauthorized electronic access is the same for all three alternatives. The differences among the alternatives are simply where the security controls must be applied and by whom; therefore, the three alternatives present similar risk profiles in this respect.

Privacy. The requirement to tightly control access to certain sensitive information is also the same for all three alternatives. Alternatives 2 and 3 pose a lower risk because their application environments support user-level or role-based access control.

Vendor Viability. The long-term viability of any technology vendor, regardless of size, cannot be predicted. Several factors mitigate this risk. Technologies and technology products that have reasonable market adoption generally continue to be sold and supported by successor companies long after a merger or acquisition. This applies to all the alternatives. Next, data can be insulated from application-level dependencies if they are managed in a self-describing, standards-based packaging format. Alternatives 2 and 3 have lower risk

because they employ this strategy. Furthermore, data can be insulated from specific storage dependencies if they are network-accessible over standard protocols. Alternative 3 presents the lowest risk in that it makes migrating data from one application environment to another or from one storage provider to another relatively simple, should this become necessary.

Cost. The cost of replacement technology and of managing it is another technical risk that must be considered. It is safe to assume that the pace of technology innovation will continue at the current rate for the foreseeable future, making technology less costly over time (i.e., the cost per unit of storage, network bandwidth, and so forth will decline). Therefore, in the research team's judgment, none of the alternatives poses a significant risk for unacceptable cost escalation. Because the cloud computing trend will continue to grow, the economy of scale it leverages makes Alternative 3 the lowest risk.

Flexibility. Alternative 3 also poses the lowest risk should the archive's scope increase dramatically (e.g., to support other programs and house more data) or if it becomes necessary to discontinue its operation. In either case, the pay-as-you-go model is an inherent advantage of Alternative 3.

Business Risk

In addition to understanding the technology risks, it is also important to understand the business implications of the three alternatives as they may greatly impact the future long-term implementation agent and its roles and responsibilities. On the basis of the research team's experience with similar projects, the following risks are quite real during the Reliability Archive's life cycle:

- Potential loss of institutional support for the continuation of the archive's critical activities and for the maintenance and operation of the system in the post-L13 era;
- Lack of a reliable source of continued funding into the uncertain future; and
- Lack of backup, by the implementation agent's staff or contractor personnel, for ongoing functional and technology operations.

A recent letter from the TRB Long-Term Pavement Performance (LTPP) Committee addressed to the executive directors of FHWA and AASHTO is a case in point where such business risks are becoming a real threat to the continued development and implementation of projects such as LTPP (*2*).

For the Reliability Archive, these business risks may occur in the near future or years after the long-term implementation agent takes over the responsibility for maintaining and operating the archive. It is prudent to take such risks into consideration during the decision-making stage in determining the optimal solution alternative. Even though Alternative 3 cannot and will not completely eliminate these potential business risks, it has the highest probability of reducing the risk factors because of its significantly lower recurring costs, its minimal dependency on specialized IT expertise, and its capacity to allow the implementation agent to make better use of available resources to enhance program management, communications, and collaborations.

Justification Based on Schedule

One of the key feasibility requirements specified by SHRP 2 is that the implementation of the Reliability Archive be completed within an 18-month time period. Typically, a project schedule is dictated by the technical approach and other factors such as project management, availability of resources, and quality of work.

It is not practical at this stage to develop a prescribed project timeline for each alternative that would accurately specify the duration of each activity. However, in order to compare the three alternatives and draw conclusions on the likelihood of their being completed within the 18-month time period, it is imperative to provide estimates on implementation timeline. Table 4.4 lists the major implementation steps of the three alternatives and the estimate of the duration of each step. Note that some steps are not applicable to every alternative. This table is built on the following assumptions:

- All three alternatives will be implemented in accordance with standard systems development life cycle phases;
- Variations on specific implementation steps may arise because of the unique approaches of each alternative; and
- Although some activities could be performed in parallel depending on how the project is planned and managed, it is conservatively assumed that all these activities will be performed in a sequential manner.

The estimated durations shown in Table 4.4 are based on the research team's understanding of the nature of these steps within each alternative as well as the team's knowledge of the standard software development life cycle methodology. In actuality, the duration of a step can be affected by many technical, business, and political factors. Its impact on the overall project schedule could be quite substantial depending on the criticality of that step. Table 4.5 assesses the criticality of these activities to the project schedule under each alternative, as well as the likelihood of schedule overrun.

From the comparison shown in Table 4.5, the research team expects that the implementation will start with project

Table 4.4. Estimated Implementation Duration

	Average Duration (in months)		
Major Implementation Steps	**Alternative 1**	**Alternative 2**	**Alternative 3**
Project planning	1	1	1
Finalize requirements	1	1	1
Finalize system architecture	1	1	1
Finalize user interface design	1	1	1
Procure hardware	4	4	2
Set up data archive infrastructure	4	4	n/a
In-house development	5	n/a	n/a
COTS installation	n/a	1	1
COTS configuration	n/a	2	2
Cloud computing service setup	n/a	n/a	3
Acceptance testing	3	3	3
Estimated total project duration	**20**	**18**	**15**

planning and the finalization of requirements and design. These steps will largely involve reviewing and refining the outcome of this project (L13). The deliverables from L13 should provide a jumpstart to these tasks.

Both Alternative 1 and Alternative 2 are highly dependent on hardware procurement and installation as well as on data storage environment setup. These steps are on the critical path of the project schedule. They require adherence to National Academies and TRB computer hardware procurement procedures, in addition to extensive logistical coordination with the National Academies. This is expected to be a lengthy process.

Both Alternative 2 and Alternative 3 will need to install and configure the selected COTS software. The RFP process is expected to lead to the selection of a mature software product. Thus, while this is a critical step, it is not expected to impose significant risks to the project schedule.

Although Alternative 1 is viewed as the bare minimum solution, it requires a great deal of effort in project scoping, in-house

Table 4.5. Assessment of Implementation Schedule Risk

		Likelihood of Schedule Overrun		
Major Implementation Steps	**Critical to Schedule**	**Alternative 1**	**Alternative 2**	**Alternative 3**
Project planning	Average	Low	Low	Low
Finalize requirements	Average	Low	Low	Low
Finalize system architecture	Average	Low	Low	Low
Finalize user interface design	Average	Low	Low	Low
Procure hardware	High	High	High	Low
Set up data archive infrastructure	High	High	High	n/a
In-house development	High	High	n/a	n/a
COTS installation	High	n/a	Low	Low
COTS configuration	High	n/a	Low	Low
Cloud computing service set up	High	n/a	n/a	Standard
Acceptance testing	Average	High	Medium	Medium

Table 4.6. Summary of Schedule Risk Analysis

	Alternative 1	Alternative 2	Alternative 3
Estimated duration	20 months	18 months	15 months
Likelihood of project schedule overrun	High	Medium	Low

software development, as well as in extensive system and user testing. Therefore, Alternative 1 is considered to be high risk.

The cloud storage services needed for Alternative 3 are critical to the overall project schedule; however, they are readily available and require minimal time for configuration. From a technical perspective, all three alternatives can be reasonably completed within an 18-month time period if they are properly managed (see Table 4.6). However, Alternative 1 and Alternative 2 include certain critical steps that could bring potential risks and impact the project schedule, while Alternative 3 does not have such risk factors. Thus, Alternative 3 provides the highest likelihood of completing the implementation of the L13A Archive on time.

Conclusions

The research team believes that it has established that it is highly feasible for the SHRP 2 program to cost-effectively deploy a data archival system that meets all of the goals and objectives envisioned by its major stakeholders.

Furthermore, the team believes that Alternative 3 represents the best path to success because of the following considerations:

- COTS digital repository management software offers the requisite functionality for producers, consumers, and the effective long-term management of a digital collection. A commercial solution that meets the functional requirements for the system should be less costly over its life cycle than a custom solution.
- Cloud storage is a cost-effective archival storage solution that obviates the need for long-term management of complex technology.
- Its lowest initial and recurring costs make it the most cost-effective approach.
- It offers the maximum set of benefits to the SHRP 2 program, user community, and the implementation agent.
- It carries the lowest risk both technically and businesswise.
- It has the highest likelihood of being deployed within the desired time frame.

For these reasons the research team recommends that the SHRP 2 program proceed with the L13A Reliability Archive project as planned, following the approach described in this report.

References

1. *Special Report 296: Implementing the Results of the Second Strategic Highway Research Program: Saving Lives, Reducing Congestion, Improving Quality of Life.* Transportation Research Board of the National Academies, Washington, D.C., 2009.
2. Letter from William H. Temple to Jeffery F. Paniati and John Horsley, 17 June 2009, Transportation Research Board. http://onlinepubs.trb.org/onlinepubs/sp/ltpp_letter_24.pdf. Accessed March 3, 2011.

APPENDIX A

Overview of Reliability and Other Projects

Table A.1. Overview of Reliability and Other Projects

Reliability and Relevant Projects	Status	Using Raw Data			Research Outcome					Reliance on Other Projects	Metadata
		Raw Data Needed	Data Sources	Proprietary Data Rights	Derived Data	Models	Tools	Codes	Reports		
L01 Integrating Business Processes to Improve Reliability	Active	No	n/a	n/a	n/a	✓	n/a	n/a	✓	n/a	No
L02 Establishing Monitoring Programs for Mobility and Travel Time Reliability	Active	Yes	State DOTs, private, mobile technologies	Yes	✓	✓	✓	✓	✓	n/a	TBD
L03 Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies	Active	Yes	State DOTs, Inrix	Yes	✓	✓	✓	✓	✓	n/a	Project has own data collection guidelines
L04 Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools	Active	Yes	Private vendors	Yes	✓	✓	✓	✓	✓	L02, L03	TBD
L05 Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes	Planned	Yes	SHRP 2 projects	No	✓	✓	✓	n/a	✓	L01, L03, L04, L07, L11	TBD
L06 Institutional Architectures to Advance Operational Strategies	Active	No	n/a	n/a	n/a	✓	n/a	n/a	✓	n/a	No
L07 Evaluation of Cost-Effectiveness of Highway Design Features	Active	Yes	TBD	Yes	✓	✓	n/a	n/a	✓	n/a	TBD
L08 Incorporating Nonrecurrent Congestion Factors into the Highway Capacity Manual Methods	Planned	Yes	TBD	Yes	✓	✓	✓	✓	✓	L01, L03, L07	TBD

(continued on next page)

Table A.1. Overview of Reliability and Other Projects (continued)

Reliability and Relevant Projects	Status	Using Raw Data			Research Outcome					Reliance on Other Projects	Metadata
		Raw Data Needed	Data Sources	Proprietary Data Rights	Derived Data	Models	Tools	Codes	Reports		
L09 Incorporation of Nonrecurrent Congestion Factors into the AASHTO Policy on Geometric Design	Planned	Yes	TBD	Yes	n/a	✓	n/a	n/a	✓	L07, L08	TBD
L10 Feasibility of Using In-Vehicle Video Data to Explore How to Modify Driver Behavior that Causes Non-recurring Congestion	Active	Yes	TBD	Yes	✓	✓	✓	✓	✓	TBD	TBD
L11 Evaluating Alternative Operations Strategies to Improve Travel Time Reliability	Active	Yes	n/a	n/a	✓	✓	n/a	n/a	✓	n/a	TBD
L12 Improving Traffic Incident Scene Management	Active	No	n/a	n/a	n/a	n/a	n/a	n/a	✓	n/a	n/a
L14 Effectiveness of Different Approaches to Disseminating Traveler Information on Travel Time Reliability	Active	Yes	TBD	Yes	✓	✓	✓	✓	✓	TBD	TBD
L15 Reliability Innovations Deserving Exploratory Analysis (IDEA)	Planned	Yes	TBD	Yes	✓	✓	✓	✓	✓	TBD	TBD
C04 Improving Our Understanding of How Highway Congestion and Pricing Affect Travel Demand	Active	Yes	State DOTs, survey data, previous studies, modeling	Yes	✓	✓	✓	✓	✓	Yes	Project has own data collection template and guidelines
C05 Understanding the Contribution of Operations, Technology, and Design to Meeting Highway Capacity Needs	Active	Yes	State DOTs, previous studies, modeling	Yes	✓	✓	✓	✓	✓	Yes	Project has own data collection template and guidelines

APPENDIX B

Life-Cycle Cost Worksheets

Table B.1. Alternative 1 Life-Cycle Cost Worksheet

		No. Hours per Resource						Hourly Rate per Resource					
Initial Cost (from 2010 to 2011)													
Hardware	$515,000												
Servers (Web, Application or Database)	$10,000												
Networking (Load balancers, switches, etc.)	$5,000												
Storage	$500,000												
Software	$0												
Application Software	$0												
System Software	$0												
Implementation Support/ Program Management	$487,600	PM	System Architect	Archival Expert	Analysts	QA/QC	Admin	PM	System Architect	Archival Expert	Analysts	QA/QC	Admin
Appraisal, Approval & Ingestion	$142,000	150	250	250	500	200	100	$130	$125	$125	$80	$75	$50
Metadata Development	$142,000	150	250	250	500	200	100	$130	$125	$125	$80	$75	$50
File System/Software Set up	$142,000	150	250	250	500	200	100	$130	$125	$125	$80	$75	$50
Marketing and Communication	$61,600	120	40	40	200	200	100	$130	$125	$125	$80	$75	$50
Recurring Cost													
Annual Costs		Hours	Rate										
System Administration	$170,000	2,000	$85										
System Maintenance	$102,000												
Hardware Maintenance (Break/Fix)	$102,000												
Software Maintenance (Support, Update)	$0	Hours	Rate										
Marketing and Customer Services	$170,000	2,000	$85										
Hosting (Floorspace, power and cooling)	$6,000												
5-Year Periodic Costs													
Software upgrade	$0												
Hardware upgrade	$510,000												

Notes: Storage hardware estimate based on 94 TB unformatted capacity Sun 7410 Unified Storage System as representative of class. Usable, protected capacity after RAID, hot spare allocation, and so forth would be in the 70-TB range. Two systems configured as replication pair for site-level disaster protection. Hosting cost is floor space (one rack), power and cooling chargeback estimated at $500 per month.

Table B.2. Alternative 2 Life-Cycle Cost Worksheet

		No. Hours per Resource						Hourly Rate per Resource					
Initial Cost (from 2010 to 2011)													
Hardware	$525,000												
Servers (Web, Application or Database)	$20,000												
Networking (Load balancers, switches, etc.)	$5,000												
Storage	$500,000												
Software	$170,000												
Application Software	$150,000												
System Software	$20,000												
Implementation Support/ Program Management	$346,850	PM	System Architect	Archival Expert	Analysts	QA/QC	Admin	PM	System Architect	Archival Expert	Analysts	QA/QC	Admin
Appraisal, Approval & Ingestion	$115,750	150	40	250	500	200	100	$130	$125	$125	$80	$75	$50
Metadata Development	$98,500	75	60	250	500	100	50	$130	$125	$125	$80	$75	$50
System Configuration & Customization	$71,000	75	125	125	250	100	50	$130	$125	$125	$80	$75	$50
Marketing and Communication	$61,600	120	40	40	200	200	100	$130	$125	$125	$80	$75	$50
Recurring Cost													
Annual Costs		Hours	Rate										
System Administration	$170,000	2,000	$85										
System Maintenance	$134,000												
Hardware Maintenance (Break/Fix)	$100,000												
Software Maintenance (Support, Update)	$34,000	Hours	Rate										
Marketing and Customer Services	$170,000	2,000	$85										
Hosting (Floorspace, power and cooling)	$6,000												
5-Year Periodic Costs													
Software upgrade	$0												
Hardware upgrade	$520,000												

Notes: Four servers at $5,000 each. Same storage and hosting costs as Alternative 1. Licensing fees for COTS application software and system software included.

Table B.3. Alternative 3 Life-Cycle Cost Worksheet

Initial Cost (from 2010 to 2011)													
Hardware	$25,000												
Servers (Web, Application or Database)	$20,000												
Networking (Load balancers, switches, etc.)	$5,000												
Storage	$0												
Software	$170,000												
Application Software	$150,000			No. Hours per Resource						Hourly Rate per Resource					
System Software	$20,000												
Implementation Support/ Program Management	$346,800	PM	System Architect	Archival Expert	Analysts	QA/QC	Admin	PM	System Architect	Archival Expert	Analysts	QA/QC	Admin
Appraisal, Approval & Ingestion	$115,750	150	40	250	500	200	100	$130	$125	$125	$80	$75	$50
Metadata Development	$98,500	75	60	250	500	100	50	$130	$125	$125	$80	$75	$50
System Configuration & Customization	$71,000	75	125	125	250	100	50	$130	$125	$125	$80	$75	$50
Marketing and Communication	$61,600	120	40	40	200	200	100	$130	$125	$125	$80	$75	$50
Recurring Cost													
Annual Costs		Hours	Rate										
System Administration	$85,000	1,000	$85										
System Maintenance	$34,000												
Hardware Maintenance (Break/Fix)	$4,000												
Software Maintenance (Support, Update)	$30,000	Hours	Rate										
Marketing and Customer Services	$170,000	2,000	$85										
Hosting (Floorspace, power and cooling)	$1,200												
Cloud Storage Service	$124,000												
5-Year Periodic Costs													
Software upgrade	$0												
Hardware upgrade	$20,000												

Notes: Same software costs as Alternative 2. Hosting cost estimated at $100 per month for four 1U type servers.

The usage charges, estimated using the Amazon S3 pricing model, are as follows:

- Storage
  - $0.150 per GB: First 50 TB/month of storage used
  - $0.140 per GB: Next 50 TB/month of storage used
  - $0.130 per GB: Next 400 TB/month of storage used
  - $0.120 per GB: Storage used/month over 500 TB
- Data Transfer
  - $0.170 per GB: First 10 TB/month data transfer out
  - $0.130 per GB: Next 40 TB/month data transfer out
  - $0.110 per GB: Next 100 TB/month data transfer out
  - $0.100 per GB: Data transfer out/month over 150 TB
- Requests
  - $0.01 per 1,000 PUT, COPY, POST, or LIST requests
  - $0.01 per 10,000 GET and all other requests

The model assumes a 2-year ramp in capacity, as projected data is ingested into the archive, and a steady state in year 3 and thereafter. Provision is made in Year 1 and 2 cloud storage costs for significant data downloaded (transfer out) for testing purposes.

APPENDIX C

Requirements and Scoring

Table C.1. Requirements and Scoring

Category	Number	Description	Priority	Alternative Scores 1	2	3
Producers	A1	The system should provide tools that enforce best practices for preparing digital objects for submission.	Critical	0	2	2
	A2	The system should support a deposit workflow initiated by the creator of content in order to submit a digital work or collection to the Accessioning Workbench function.	Critical	0	2	2
	A3	The system must provide an Accessioning Workbench function where digital objects can be staged for appraisal, classification, and packaging prior to ingestion into permanent storage for safekeeping.	Critical	0	2	2
			Category Total:	**0**	**6**	**6**
Ingestion	B1	The system should support API, HTTP, NFS v3 and v4, and CIFS file system-based ingestion methods.	Desirable	1	1	1
	B2	HTTP, NFS, and CIFS interfaces should not require external appliances.	Desirable	1	1	1
	B3	The system must return a positive or negative confirmation of ingestion back to the archiving application. A positive confirmation can only be returned after all the preservation policies are being enforced—i.e., retention period is set, multiple instances are created or replicated, and so forth.	Critical	0	2	2
	B4	The system must raise a failed ingestion alert to the standard reporting interfaces. Data reported with a failed ingestion event should include, but not be limited to, reason for failure, time stamp, source data item, and target.	Critical	0	2	2
	B5	The system should have the optional ability to log positive confirmation of ingestion to the standard reporting interfaces.	Desirable	0	2	2
	B6	The system should support multiple active paths and interfaces to allow parallel writing.	Desirable	1	1	2
	B7	It should be possible to optionally dedicate an interface to ingesting data from a specific application on a shared-application device.	Desirable	1	1	1
			Category Total:	**4**	**10**	**11**

(continued on next page)

Table C.1. Requirements and Scoring (continued)

Category	Number	Description	Priority	Alternative Scores 1	2	3
Archival storage	C1	The archival storage function must guarantee against data loss, through hardware redundancy and/or by maintaining multiple internal instances of a digital object.	Critical	1	1	1
	C2	Data in archival storage must be guaranteed to be immutable. The system must actively monitor the integrity of digital objects—e.g., by recomputing a digital signature and comparing it against a value in the preservation metadata.	Critical	1	1	1
	C3	Archival storage must appear to the overall archival information system as a single namespace.	Critical	1	1	2
	C4	It must be possible to address data objects within the archival storage function without addressing boundaries.	Critical	1	1	2
	C5	The maximum capacity of a single, manageable instance of archival storage (measured in terms of user files and objects and not counting any internal replicas) must be at least 50% larger than the amount of data expected to be submitted initially to the archive.	Critical	1	1	2
	C6	The maximum size of a single object that can be managed by archival storage must be at least 50% larger than the largest object expected to be submitted to the archive.	Critical	1	1	1
	C7	The maximum number of user objects capable of being managed within the single storage namespace must be at least 50% greater than the number of user objects expected to be submitted initially to the archive.	Critical	1	1	2
	C8	It must be possible to add storage capacity without impact to the user community or the need for significant operations staff involvement.	Critical	1	1	2
	C9	It must be possible to add capacity in a zero-disruption manner. This should include addition of disks in existing shelves, new disk shelves, and new disk connectivity such that the addressable capacity grows seamlessly.	Critical	1	1	2
	C10	The system must automatically take advantage of the additional capacity without configuration change or administrative action.	Critical	1	1	2
	C11	It must be possible to perform data-in-place upgrades—e.g., replacement of storage controllers and heads while retaining existing storage to allow step-change upgrades.	Critical	1	1	2
	C12	It must be possible to move data off a particular storage subsystem component, such as a shelf or disk, as a background activity to allow upgrade or replacement once the internal data migration is complete.	Critical	1	1	2
	C13	It must be possible to segment the storage namespace—e.g., into different directories and folders—in order to logically segregate data and simplify management.	Critical	1	1	1
	C14	It must be possible to direct ingested data to a particular segment.	Critical	1	1	1
	C15	It must be possible to set a default data retention period for a segment.	Critical	1	1	1
	C16	It should be possible to configure multiple tiers of storage—i.e., with different performance, costs, or other characteristics—into a single, manageable archival storage instance.	Desirable	1	2	2
	C17	It must be possible to migrate data between storage tiers such that data access works in an unchanged manner during and after migration.	Critical	1	2	2
	C18	It should be possible to automatically migrate data from a low-performance tier to a high-performance tier based on access patterns.	Desirable	1	2	2

(continued on next page)

Table C.1. Requirements and Scoring (continued)

Category	Number	Description	Priority	Alternative Scores 1	2	3
	C19	The system should have the ability to use de-duplication, compression, and/or similarity-reduction processes on ingested data. Any such technique must guarantee immutability and ensure against data loss in any circumstance.	Desirable	1	1	1
	C20	In the event of power degradation or failure, archival storage must protect from data loss or corruption.	Critical	1	1	1
			Category Total:	**20**	**23**	**32**
Data management	D1	The system must be able to manage digital objects made up of multiple components, versions (generations), and files.	Critical	0	2	2
	D2	Relationships between these elements must be maintained to ensure that the files constituting an object can be delivered in the right order.	Critical	0	2	2
	D3	The system must maintain the reference, provenance, context, and fixity information needed for preservation management purposes.	Critical	0	2	2
	D4	The system must have an open, extensible, and standard way of packaging metadata for digital objects.	Critical	0	2	2
	D5	A digital object should be packaged for submission or dissemination in a way that is completely self-describing.	Critical	1	2	2
	D6	Each component of a digital object should contain its own self-describing descriptive metadata.	Critical	0	2	2
	D7	It must be possible to add new versions of descriptive metadata alongside older versions of metadata.	Critical	0	2	2
			Category Total:	**1**	**14**	**14**
Preservation planning	E1	The system must prevent a digital object from being deleted until its specified retention period has expired.	Critical	1	1	1
	E2	It must be possible to extend the retention periods for individual digital objects or groups of objects.	Critical	1	1	1
	E3	It should be possible to specify an infinite retention period.	Desirable	1	1	1
	E4	It should be possible to specify an indefinite retention period—i.e., the retention period is infinite until explicitly set at a later time.	Desirable	1	1	1
	E5	Data deletion must be an auditable event.	Critical	1	1	1
	E6	All instances of a data object should be deleted simultaneously at the end of its life cycle—e.g., replica instances.	Critical	1	1	1
	E7	It should be possible to perform a privileged, audited delete on a data object prior to the end of its retention life cycle.	Desirable	1	1	1
	E8	Deletion of a data object must be transparent to any de-duplication mechanism.	Critical	1	1	1
	E9	The system should support the ability to delete data securely (shred, wipe, or clean) to industry-accepted standards—e.g., US DoD 5520-M.	Desirable	1	1	1
	E10	It must be possible for the Archival Storage tier of the Archival System to be replicated to one or more additional storage instances to provide resilience to site-level disasters.	Critical	1	1	2
	E11	It must be possible to compress replication traffic.	Critical	1	1	2
	E12	It should be possible to throttle replication traffic to avoid swamping network links.	Critical	1	1	2

(continued on next page)

Table C.1. Requirements and Scoring (continued)

Category	Number	Description	Priority	Alternative Scores 1	2	3
	E13	It must be possible to separate ingestion and retrieval traffic from replication traffic to use different networks and interfaces.	Critical	1	1	2
	E14	If replication occurs asynchronously, it should be possible to adjust the frequency of replication on demand (manual, scripted, and based on thresholds) to meet service needs.	Critical	1	1	2
	E15	It should be possible to flip-flop source and target relationships and resynchronize simply. Full resynchronization is acceptable in very limited scenarios.	Critical	1	1	2
	E16	It must be possible to establish a one-to-one replication relationship.	Critical	1	1	1
	E17	It should be possible to establish one-to-many, many-to-one and chain replication relationships.	Desirable	1	1	0
	E18	It must be possible to encrypt in-flight replication data without making the customer responsible for any key management methodology employed with such a solution.	Critical	1	1	1
	E19	Replication failures must be reported to the management and reporting interfaces.	Critical	1	1	1
	E20	Tools must be available to debug and resolve replication failures.	Critical	1	1	2
	E21	The system must provide the ability to back up the data in the system to an NDMP-compliant target.	Critical	1	1	1
	E22	The system must provide the ability to select a subset of data to back up based on timescale (e.g., data ingested between two dates), pool, and file name.	Critical	1	1	1
	E23	The system must provide the ability to save system configuration and system metadata to an external device or system in order to facilitate system recovery and reconstruction.	Critical	1	2	2
	E24	The format of the backup package must be documented and based upon industry standards to ensure that backed up data can be recovered independent of any backup hardware or software.	Critical	1	1	1
	E25	It should be possible to encrypt the contents of a backup operation.	Critical	1	1	1
	E26	It should be possible to compress the contents of a backup operation.	Desirable	1	1	1
	E27	The system must support the logical migration of digital objects from one format to another—e.g., from a proprietary document format to PDF—while preserving all previous generations of the object.	Critical	0	2	2
	E28	Logical migration must be an auditable event.	Critical	0	2	2
			Category Total:	**26**	**31**	**37**
Administration	F1	The system must run continuously with no scheduled downtime.	Critical	1	1	1
	F2	Administration of the system must be possible over encrypted protocols.	Critical	1	1	1
	F3	Where a secure protocol is available, it must be possible to disable the insecure administration access method.	Critical	1	1	1
	F4	Changes must be fully logged and traceable to a specific administrator.	Critical	1	1	1
	F5	All factory admin. or backdoor support users must be specified. It must be possible for the customer to secure such access and enable it on an as-needed basis.	Critical	1	1	1
	F6	Change execution functionality must protect the administrator from being able to make unsupported or incorrect changes wherever possible.	Critical	1	1	1
	F7	The system must report and raise alerts in all failure and degraded service scenarios.	Critical	1	1	1

(continued on next page)

Table C.1. Requirements and Scoring (continued)

Category	Number	Description	Priority	Alternative Scores 1	2	3
	F8	The system should report performance trends and bottlenecks using historical data.	Desirable	1	1	1
	F9	The system must provide real-time performance and usage reporting.	Critical	1	2	2
	F10	Reporting functionality should include the ability to collect metrics at all appropriate configuration points and pass them on to external management reporting systems.	Desirable	1	2	2
	F11	It should be possible to integrate the Archival System into alerting platforms via SNMP.	Desirable	1	2	2
	F12	Ports and protocols used for management and reporting should be documented in case access through firewalls is required.	Desirable	1	1	1
			Category Total:	**12**	**15**	**15**
Access	G1	The system should support multiple active paths and interfaces to allow parallel retrieval.	Desirable	1	1	1
	G2	The system must enforce access control to digital objects based on rights and permissions encoded in administrative metadata.	Critical	0	2	2
	G3	The system must provide content and metadata indexing and search function.	Critical	1	2	2
	G4	The content and metadata indexing and search function should be replaceable as technology advances are made in this area.	Desirable	1	2	2
			Category Total:	**3**	**7**	**7**
Consumers	H1	Users accessing data in the Archival System must be named and mapped to roles (groups) reflecting their access rights.	Critical	1	2	2
	H2	The system must provide a self-service user management function, including, but not limited to, requesting a user name and resetting one's password.	Critical	0	2	2
	H3	The system must provide a workflow to manage the approval of user registration requests.	Critical	0	2	2
	H4	The system must provide an access portal function that allows named, authenticated, and authorized users to find and access information according to their role(s).	Critical	1	2	2
	H5	The system must provide a structured way of navigating through content in the archival system.	Critical	1	2	2
	H6	Structured navigation (e.g., menus) should be dynamic—e.g., driven by a flexible content management system—and not hard coded.	Desirable	0	2	2
	H7	The system must provide a flexible, search-driven navigation and filtering.	Critical	0	2	2
	H8	The system should provide a way for users to integrate external data with data accessible via the Archival System.	Desirable	1	1	1
	H9	The system should provide a capability for users to select subsets of data that are extracted and exported for download.	Desirable	0	1	1
	H10	The system should facilitate communication and collaboration among its user community.	Desirable	0	2	2
			Category Total:	**4**	**18**	**18**

(continued on next page)

Table C.1. Requirements and Scoring (continued)

Category	Number	Description	Priority	Alternative Scores 1	2	3
Systemwide	I1	The system must provide a configurable audit function capable of logging actions, including, but not limited to, ingestion, policy setting and enforcement, access, and administrative actions.	Critical	1	1	1
	I2	The system should provide a reporting interface that can serve management, capacity, performance, and other similar information to an external reporting system.	Desirable	1	1	1
	I3	All system components must allow lights-out management on a 24/7 continuous operational basis.	Critical	1	1	2
	I4	Performance should be predictable during exhaustion of any finite resources. Hard limits on performance must be handled gracefully.	Critical	1	1	1
	I5	The system should provide quality of service management to mitigate performance impacts from resource-intensive ingestion or access operations.	Critical	1	1	2
	I6	Failover between clustered or paired components must be seamless to ingest and access functions.	Critical	1	1	2
	I7	The system should be resilient and tolerate failure of external component services, such as management servers and DNS and NIS servers.	Critical	1	1	2
	I8	Automated failure detection and secure remote vendor support options must be present, including call-home to inform vendors of an issue and dial-in to assist with diagnosis and resolution.	Critical	1	1	2
	I9	Call-home diagnostics should include all data relevant to the failure. For example, in the case of a storage device containing different types of disk drives, any call-home should include the specific drive type for replacement to avoid incorrect parts being sent to site.	Critical	1	1	2
	I10	It should be possible to upgrade system software and firmware without service impact.	Critical	1	1	2
	I11	All hardware components should be configured with at least N+1 redundancy with full mitigation of component failure.	Critical	1	1	2
	I12	Failed components should be swappable on a break-fix basis.	Critical	1	1	2
	I13	The number of nonswappable components must be minimal and clearly enumerated.	Critical	1	1	1
	I14	All hardware deployed as part of the Archival System should be compliant with then-current “e-waste” directives.	Desirable	1	1	2
	I15	All hardware deployed as part of the Archival System should be compliant with then-current RoHS (Reduction of Hazardous Substances) directives.	Desirable	1	1	2
	I16	Data center environmental requirements (including floor space, power, and cooling) should be computable on a per-usable-unit-of-storage basis.	Desirable	1	1	2
			Category Total:	**16**	**16**	**28**

APPENDIX D

Relevant Systems Reviewed

Table D.1. Relevant Systems Reviewed

No.	Name	Description	Strengths/Useful Features	Limitations
1	LTPP Product Online (www.ltpp-products.com)	The site includes all available Long-Term Pavement Performance (LTPP) data and products maintained by FHWA.	The registration screen follows good practices by automating the front-end process of user management. Registration requires the user to agree to various terms and conditions before being granted access. The Community section signals an intention to allow users to connect, but the discussion forums and mailing lists pages are both empty.	The navigation is highly structured and hierarchical. No search capability is provided. The data access tools require explicit understanding of the database structure and the system's data model. In addition, the SQL Export tool requires knowledge of SQL.
2	LTPP Standard Data Distribution (SDR) Offline (The DVD can be requested from the FHWA LTPP office)	LTPP program information and data can also be accessed offline by obtaining the program's SDR on DVD.	The SDR consists of two DVDs. It allows the user to explore the contents of the disk and contains installable programs that mirror some of the tools available online—e.g., the Table Navigator. The database itself is contained on the second DVD of the SDR. Using it entails unzipping 12 compressed files containing approximately 36 GB of data in Microsoft Access 2000 format, and then accessing the data using the Microsoft Access program.	While these data DVDs are readable on almost any system, in practice, they are only useful on a Windows system.
3	NORC Data Enclave (www.norc.org/DataEnclave)	The NORC data enclave provides a confidential, protected environment within which authorized researchers can access sensitive microdata remotely from their offices or onsite at NORC.	It places a strong emphasis on metadata. The tool chain used by data producers, and the process followed to package data with metadata, has strong parallels to the processes researchers might follow when preparing their data for submission to the proposed archival system.	The Enclave goes to what might be considered extreme lengths to protect data privacy and prevent misuse of data. While it is web accessible, it also requires installation of client software.

(continued on next page)

Table D.1. Relevant Systems Reviewed (continued)

No.	Name	Description	Strengths/Useful Features	Limitations
4	Google Book Search (http://books.google.com)	Google is working with several major libraries to include their collections in Google Book Search and, like a card catalog, show users information about the book.	It provides a contemporary reference model for navigating information in a digital library. The full, limited, and snippet views can serve as meaningful references for how L13 might provide different approaches for users to access and view the archived data. This approach can also be applicable to L13 in dealing with different types of restricted data rights expected to be imposed on certain data. The My Library capability is a model of a community/collaboration function that might be valuable to the L13 data archive.	Their collection is limited to books.
5	Amazon.com (www.amazon.com)	The navigation facets on the left side of its screen allow the user to "drill down," or filter their results by department, by products available via certain shipping methods, or by format.	This is essentially metadata that is derived by entity extraction analysis of product descriptions and other text and presented as navigation facets. This technique allows Amazon to derive a wide range of possible connections among diverse items.	No significant limits. Outside of Google, perhaps no website shapes user's expectations about what web applications should deliver more than does Amazon.com.
6	SHRP 2 Capacity Program: Collaborative Decision Making Framework (CDMF) (www.trb.org/shrp2)	The CDMF is a systems-based, transparent, well-defined framework for consistently reaching collaborative decisions on transportation capacity enhancements.	The entry-level view is a series of portals in each phase of the transportation process where one or more decision points may occur. Increasingly detailed information can be retrieved for each key decision point.	*The framework is under development.* From its conceptual design, the framework will be primarily driven by processes. It is not clear if and how relevant data sets will be retrieved.
7	Transportation Research Information Services (TRIS) Database (Now part of TRID: The TRIS and ITRD Database, http://trid.trb.org)	TRIS is a bibliographic database funded by sponsors of the Transportation Research Board (TRB), primarily state departments of transportation and selected federal transportation agencies. The International Transport Research Documentation (ITRD) Database is produced by ITRD member countries under the Joint Transport Research Centre (JTRC) at the Organisation for Economic Cooperation and Development (OECD).	TRID contains more than 900,000 records covering reports, monographs, journals, and research in progress. TRID includes both simple and advanced query screens and offers browsing of recent publications by mode. When available, links are provided to the full text of documents or to direct ordering information. TRID allows users to print, download, directly e-mail, or share search results. It also includes "Hot Topic" searches on subjects of current interest. TRID offers users the ability to subscribe to RSS feeds to get the latest publications on a specific topic. The Transportation Research Thesaurus (http://trt.trb.org) is available from the TRID website and may be used to search TRID.	TRID is a bibliographic database only. Links to full text are not available for all documents.
8	TransXML (www.transxml.org)	This site provides the information about NCHRP Project 20-64, XML Schemas for Exchange of Transportation Data. The project developed a set of Extensible Markup Language (XML) schema for transportation applications in a framework called TransXML.	Primarily to disseminate project information. It also has a community feature through which the hosting party can post news and update project progresses. Users can also respond to certain items with their comments.	Project reports and products such as the TransXML schemas are the only available items for users to download.

TRB OVERSIGHT COMMITTEE FOR THE STRATEGIC HIGHWAY RESEARCH PROGRAM 2*

CHAIR: **Kirk T. Steudle,** *Director, Michigan Department of Transportation*

MEMBERS

H. Norman Abramson, *Executive Vice President (Retired), Southwest Research Institute*
Anne P. Canby, *President, Surface Transportation Policy Partnership*
Alan C. Clark, *MPO Director, Houston-Galveston Area Council*
Frank L. Danchetz, *Vice President, ARCADIS G&M, Inc.*
Dan Flowers, *Director, Arkansas State Highway and Transportation Department*
Stanley Gee, *Acting Commissioner, New York State Department of Transportation*
Michael P. Lewis, *Director, Rhode Island Department of Transportation*
Susan Martinovich, *Director, Nevada Department of Transportation*
John R. Njord, *Executive Director, Utah Department of Transportation*
Charles F. Potts, *Chief Executive Officer, Heritage Construction and Materials*
Gerald Ross, *Chief Engineer, Georgia Department of Transportation*
George E. Schoener, *Executive Director, I-95 Corridor Coalition*
Kumares C. Sinha, *Olson Distinguished Professor of Civil Engineering, Purdue University*

EX OFFICIO

Victor M. Mendez, *Administrator, Federal Highway Administration*
Ron Medford, *Acting Administrator, National Highway Transportation Safety Administration*
John C. Horsley, *Executive Director, American Association of State Highway and Transportation Officials*

LIAISONS

Tony Kane, *Director, Engineering and Technical Services, American Association of State Highway and Transportation Officials*
Jeffrey F. Paniati, *Executive Director, Federal Highway Administration*
John Pearson, *Program Director, Council of Deputy Ministers Responsible for Transportation and Highway Safety, Canada*
Margie Sheriff, *Director, SHRP 2 Implementation Team, Office of Corporate Research, Technology, and Innovation Management, Federal Highway Administration*
Michael F. Trentacoste, *Associate Administrator, Research, Development, and Technology, Federal Highway Administration*

RELIABILITY TECHNICAL COORDINATING COMMITTEE*

CHAIR: **R. Scott Rawlins,** *Deputy Director/Chief Engineer, Nevada Department of Transportation*
VICE CHAIR: **John F. Conrad,** *Director, Highway/Bridge Market Segment, Transportation Business Group, CH2M HILL*

MEMBERS

Malcolm E. Baird, *Consultant*
Kevin W. Burch, *President, Jet Express, Inc.*
John Corbin, *State Traffic Engineer, Wisconsin Department of Transportation*
Henry de Vries, *Captain, New York State Police*
Leslie S. Fowler, *ITS Program Manager, Intelligent Transportation Systems, Bureau of Transportation Safety and Technology, Kansas Department of Transportation*
Steven Gayle, *Consultant, Gayle Consult, LLC*
Bruce R. Hellinga, *Associate Professor, Department of Civil and Environmental Engineering, University of Waterloo, Ontario, Canada*
Lap Thong Hoang, *President, Lap Thong Hoang, LLC*
Patricia S. Hu, *Director, Bureau of Transportation Statistics U.S. Department of Transportation*
Sarath C. Joshua, *ITS and Safety Program Manager, Maricopa Association of Governments*
Mark F. Muriello, *Assistant Director, Tunnels, Bridges and Terminals, The Port Authority of New York and New Jersey*
Richard J. Nelson, *Assistant Director, Operations, Nevada Department of Transportation*
Richard Phillips, *Director, Administrative Services, Washington State Department of Transportation*
Constance S. Sorrell, *Chief of Systems Operations, Virginia Department of Transportation*
L. Scott Stokes, *Deputy Director, Idaho Department of Transportation*
Jan van der Waard, *Program Manager, Mobility and Accessibility, Netherlands Institute for Transport Policy Analysis*
John P. Wolf, *Assistant Division Chief, Traffic Operations, California Department of Transportation (Caltrans)*
Margot Yapp, *Vice President, Nichols Consulting Engineers, Chtd.*

FHWA LIAISONS

Robert Arnold, *Director, Transportation Management, Office of Operations, Federal Highway Administration*
Margie Sheriff, *SHRP 2 Implementation Director, Office of Corporate Research, Technology, and Innovation Management, Federal Highway Administration*
David Yang, *Highway Research Engineer, Office of Operations Research and Development, Federal Highway Administration*

CANADA LIAISON

Andrew Beal, *Manager, Traffic Office, Highway Standards Branch, Ontario Ministry of Transportation*

*Membership as of March 2011.

Related SHRP 2 Research

Establishing Monitoring Programs for Mobility and Travel Time Reliability (L02)

Analytic Procedures for Determining the Impacts of Reliability Mitigation Strategies (L03)

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 (L13A)

A Framework for Improving Travel Time Reliability (L17)

WWW.TRB.ORG/SHRP2

THE NATIONAL ACADEMIES™
Advisers to the Nation on Science, Engineering, and Medicine

The nation turns to the National Academies—National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council—for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org

ISBN 978-0-309-12900-8
90000
9 780309 129008