Reliability Project L13

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

MAY 2011

In this project a feasibility study was conducted to assess the technical, economic, and business aspects of developing, operating, and maintaining an archive for data from SHRP 2 Reliability and related research projects. The archive would make this data accessible to researchers and practitioners for up to 50 years. In the course of the feasibility study, the research team developed three alternative solutions for the data archive. The actual archive would be created and implemented in Project L13A: Design and Implement a System for Archiving and Disseminating Data from SHRP 2 Reliabilities and Related Studies/Assistance to Contractors to Archive their Data for Reliability Projects. This document summarizes the final report of the feasibility study, Project L13. The final report is available on the SHRP 2 website: www.TRB.org/SHRP2. Zongwei Tao, Jeffrey Spotts, and Elizabeth Hess of Weris, Inc. comprised the project team. David Plazak, the responsible program officer, can be contacted at dplazak@nas.edu.

The Need for an Archive

The reliability of travel times on transportation facilities can only be assessed in the context of a statistical distribution of travel times. A number of factors impact 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 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. Reliability research in SHRP 2 will generate a great deal of such data that will be useful in the development of models, tools, and strategies for improving reliability, but only if the data are both preserved and accessible to researchers. To address this need, Project L13 conducted a study to examine the feasibility of creating such an archive and maintaining it for up to 50 years.

The data archive should meet three goals: (1) preserve all valuable digital assets collected and produced by SHRP 2 Reliability and related research projects for up to 50 years; (2) provide transportation researchers and practitioners with a way to discover and access these digital assets in standard, open formats; and (3) 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, provide data integration or mash-up services, and create a collaborative community.

A conventional relationship database system was determined to be inadequate to achieve these goals. The researchers proposed that the archive follow the structure of a digital library or museum. Libraries and museums focus on preserving information, maintaining its provenance, and putting it into context. Libraries and museums also have growing

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

TRANSPORTATION RESEARCH BOARD
OF THE NATIONAL ACADEMIES

bodies of standards, software tools, and best practices that are gaining worldwide adoption.

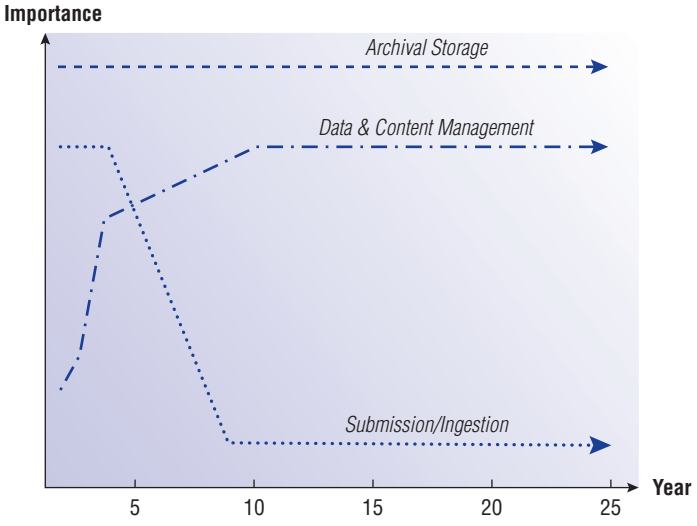
Users

The primary purpose of the Reliability data archive system would be to allow users to validate the research results from relevant SHRP 2 projects and to refine and build on research results into the future. The archive would be designed for leaders of transportation agencies, technical staff of transportation agencies, non-transportation professionals with some relationship to transportation operations (such as law enforcement officers, fire fighters, and special event venue managers), researchers, and analysts. Access to the archive is expected to be free of restrictions. In general, there is no perceived negative impact with respect to data rights as the majority of raw data used by the Reliability projects comes from the public sector and has few confidentiality or security issues.

Data

Reliability projects include a diverse array of file types and formats to embody the intellectual products of the research. For example, products will include structured datasets on roadway incidents; volume, occupancy, and speed; roadway characteristics; and extraordinary events, each in significantly or subtly different formats. The data archive must then accommodate both raw and aggregated datasets, for which relationships and linkages to conclusions must be maintained; derived data; models; tools; code; and written reports. Additionally, metadata and a data dictionary will be critical components of the archive.

FIGURE 1 ***Relative Importance of Functionality over Time***



Principal Implementation Agent

The research team recommends that a principal implementation agent be responsible for migrating the Reliability data archive to a production environment once its development is completed under SHRP 2. The implementation agent would be responsible for long-term stewardship of the data archive—including system administration, maintenance, and upgrade—and communicating activities of the archive to the user community. The agent would also establish relevant policies and procedures for using the archival system and maintain coordination with stakeholders at both strategic and technical levels.

Functionality

Determining a solution for the data archive requires identifying which software and hardware technologies might address the functionality and operational requirements of the archival system and how the technology could be acquired and implemented. The options include commercial off-the-shelf technology, open source software, in-house software development, hosting, cloud storage, and software-and-storage-as-a-service. An additional challenge is that the importance of each functionality requirement is not static; they will vary during the life of the archive system. The relative importance of functionality over time is illustrated in Figure 1.

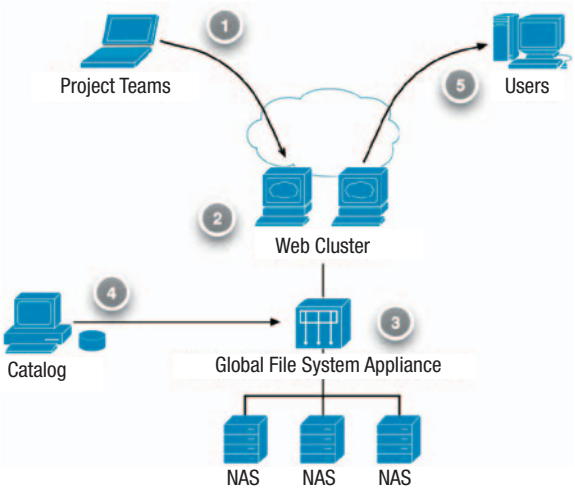
The Three Alternatives

In the course of this project, three alternative solutions for the archive were developed; each is briefly described in the following sections. Two critical issues that will influence the selection of potential alternatives are the relative importance of system functionality over time and the estimated total data volume to be preserved. Alternatives were evaluated for ability to meet the requirements, conformity with conceptual design, initial and recurring costs, benefits to stakeholders, risk mitigation, and schedule.

Alternative Number 1

This is the bare minimum alternative. It is simple and straightforward to implement, and it meets the minimum essential requirements to be considered a viable solution. 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 overall concept is depicted in Figure 2.

FIGURE 2 ***Concept of Alternative Number 1***
(NAS is an acronym for network attached storage)



Alternative Number 2

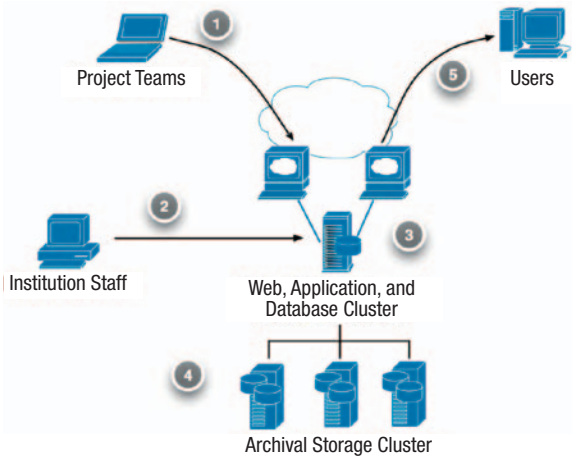
This alternative is based on digital object repository management software designed for universities, libraries, museums, archives, and information centers. This concept, depicted in Figure 3, was selected because the functionality provided by these software suites maps very closely to the functional requirements and conceptual design of the envisioned archival system.

Alternative Number 3

Alternative Number 3 is based on the same class of off-the-shelf software as Alternative 2, but instead of archiving data in self-hosted storage, archived data is preserved in a cloud-storage service. Cloud storage services store and retrieve files via a simple web service 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. Figure 4 illustrates this concept.

FIGURE 3 ***Concept of Alternative Number 2***



Benefits

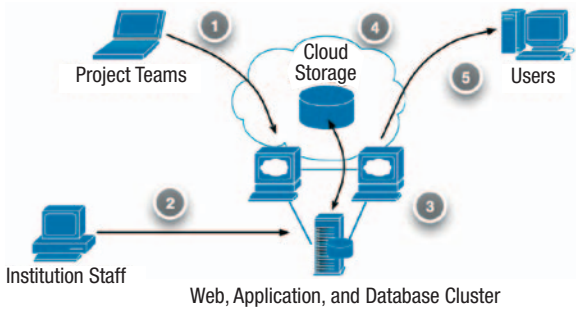
Project L13 determined that the entire SHRP 2 program could benefit from the implementation of the Reliability data archive. The benefits can be assessed with respect to long-term data preservation, sharing of system capabilities across projects and programs, and how the alternatives are best positioned to support the implementation of the SHRP 2 program results.

The Reliability archive could benefit a wide range of users. The final report includes assessments of user benefits—such as business functionality; follow-on research, testing, and evaluation; and advanced user accessibility. The report also assesses how each alternative enables system performance and reliability, as well as their relative complexity to manage over time. The table shows the research team's assessment of the relative benefits of the three alternatives.

Conclusion

Project L13 determined that it is highly feasible for the SHRP 2 program to cost-effectively deploy a data archival system that meets all the goals and objectives envisioned by its major stakeholders. The research team recommended

FIGURE 4 ***Concept of Alternative Number 3***



Benefit Targets	Benefit Aspects	Alternative 1	Alternative 2	Alternative 3
SHRP 2 Program				
	Initial investment under $1.2 M budget	No	No	Yes
	Can be implemented in 18 months	Possible	Possible	Lowest Risk
	Long-term preservation	Yes	Yes	Yes
	Sustainable (avoids 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

that SHRP 2 proceed with Project L13A: Design and Implement a System for Archiving and Disseminating Data from SHRP 2 Reliability and Related Studies/Assistance to Contractors to Archive their Data for Reliability Projects. An RFP for Project L13A was released in July 2010 and a contract is pending award approval.