Technical Aspects of Sustainable Digital Archives

Cheng-Hung Li[1], *Wei-Chun Chung*[1], *Shiang-An Wang*[1], *Chia-Hao Lee*[1]
Chi-Wen Fann[1], *Lee-Hom Lin*[1], *Jan-Ming Ho*[1,2]
[1]Research Center for Information Technology Innovation
[2]Institute of Information Science
Academia Sinica
Taipei, Taiwan
{chli, wcchung, sawang, chahao, fann, monica, hoho}@iis.sinica.edu.tw

***Abstract*—In Taiwan, the National Digital Archives Program was initiated in 2002. It was subsequently integrated with the National Science and Technology Program for E-Learning into the Taiwan e-Learning and Digital Archives Program (TELDAP) on January 1, 2008. The program has created hundreds of professional digital archives, websites, and databases to store and manage these valuable national cultural and ecological artifacts. These websites and databases are open to the public. TELDAP will celebrate its 10-year anniversary and recess at the end of 2012. In this paper, we will present efforts made by the Digital Archive Architecture Laboratory, Academia Sinica, for preserving and maintaining accessibility of contents through previously developed systems. We will primarily focus on the reliability of the archive system and the promotion of its diverse contents. Our long-term preservation procedure includes both data and software. In the area of preservation, we recommend digitization guidelines for data formats so that these valuable artifacts can be digitized at a level that is equivalent to state-of-the-art display quality. We will also provide format conversion services to convert digital objects to the latest common open format. The digital archive developed and proposed 9 standard operating procedures for each stage in the development of the digital archive system. In practice of software maintainability, we will preserve the design documents at each stage of development. Currently, we are developing a management framework on the basis of distributed virtualization technology to host digital archive systems so that management cost can be minimized while systems can remain operational.**

Keywords-digital archive; digital preservation;long-term preservation.

I. INTRODUCTION

Preservation of human knowledge for long-term accessibility is one of the missions of an institution of cultural and ecological heritage. For thousands of years, paper has been the preferred medium for storage of text and images. Currently, due to the rapid development of information technology, numerous objects (e.g., books, cultural relics, paintings, calligraphy, etc.) have been archived in digital form for their dissemination, backup, and reuse [1][2]. However, whether institutions decide to use either analog or digital approaches to archiving, they will need to employ comprehensive preservation strategies for sustainability so that archived objects will remain readable [3][4].

The TELDAP [5] is a joint project between scholars in the fields of humanities and technology. It has built a platform known as the portal of TELDAP [6] to preserve the artifacts of Taiwan's rich cultural heritage in digital form. This program functions through the collaborative efforts of 19 organizations and government agencies. It has archived over 5 million digital objects, along with metadata annotated by domain experts. In addition, it has created hundreds of professional digital archives, websites, and databases to store and manage these valuable national cultural and ecological artifacts. We must address three types of problems that affect the preservation of these systems: physical deterioration, technology obsolescence, and improper management. The first two problems are similar to problems that occur during traditional digital preservation. Both processes involve difficulties related to the preservation of format materials. In addition, inadequate management of existing systems can cause gradual system loss. In the past, when archiving institutions outsourced their systems, they did not insist on comprehensive analysis and planning of system requirements. Most of these systems lacked architectural flexibility. Therefore, expansion and data integration of these systems were difficult. Some systems had to be closed due to a lack of standard maintenance and development processes. We hope to improve the reliability, overall representation, scalability, quick response, and user-friendly, value-added environment for these valuable archives through the development of a complete suite of standard operating procedures (SOPs) that can be used to fully integrate and preserve these websites, systems, and databases.

In this paper, we propose a plan that details preservation strategies for digital archives. Through the standard process of formulation, an archive system can be developed through standardization. Thus, the system can be preserved completely through short-term system migration or long-term preservation.

This paper is organized in the following manner. In Section II, we review literature related to our plan. We also review relevant strategies for archive system preservation. In Section III, we provide detailed explanations of architecture of the sustainable digital archives we have developed. Finally, in Section IV, we present our conclusions.

ISBN: 978-1-61208-228-8

II. Related work

In this section, we introduce several issues and techniques that relate to the preservation of digital information. Digital Preservation involves a set of processes, activities, and management of digital information that is conducted over time to ensure the long-term retention and accessibility of archived information. Hence, it is not just the product of a program. It is an ongoing process. Functions within this process include management of object names and locations, updates of the storage media, and documentation of content and tracking of hardware and software changes to ensure availability and comprehensibility of objects [7]. The Research Libraries Group [8] defines Digital Preservation as a series of activities conducted to ensure that digital data can be maintained and queried continuously. The American Library Association (ALA) briefly defined Digital Preservation as a series of policies, strategies, and actions employed to ensure access to the digital contents over time, despite the challenges presented by media failure and technological change. Although there has been progress in the field of Digital Preservation, many tasks remain undone: Professionals must decide which types of file formats should be preserved and agree on the level of preservation needed. Further, in order to promote responsible stewardship of digital information, professionals must also agree to comply with standards. Presently, data can be preserved for as little as five years or for as long as more than ten years [10]. In fact, some scholars believe that the preservation period should last more than a lifetime.

Therefore, it is evident that the accessibility of digital information is extremely vulnerable in today’s rapidly evolving technological environment. A majority of the selected research stated that the field of digital preservation faces three major obstacles: physical deterioration, technology obsolescence, and the rapid growth of information. The problems inherent in the preservation of archival objects in both digital and traditional print form are rather distinct. Therefore, preservation strategies must be studied and evaluated; further planning is necessary [11]. The National Archives of UK suggested a number of criteria for selection of file formats for long-term preservation: ubiquity, support, disclosure, documentation quality, stability, ease of identification and validation, intellectual property rights, metadata support, complexity interoperability, viability, and re-usability [12]. In 1995, in order to address the issue of Long-term Digital Preservation, the International Standard Organization (ISO) established The Consultative Committee for Space Data Systems (CCSDS). In 2003, this committee formulated the Open Archival Information System (OAIS, ISO14721:2003) to provide a complete reference model of digital preservation architecture [13].

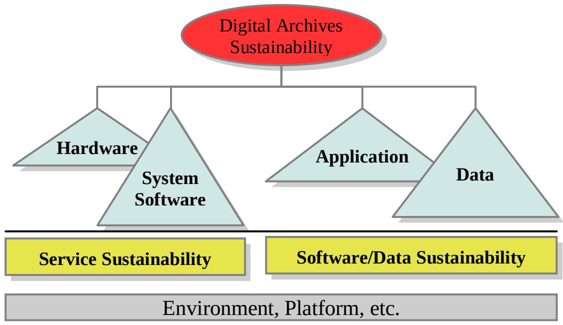
Knowledge storage institutions face a common dilemma: How can digital information sustainability be preserved? In the fight against the loss of digital information, several technical approaches have been proposed: refreshing; migration; emulation; standardization; system (technology) preservation; encapsulation (data description); replication (redundancy); and conversion to paper or analog media. Several researchers have examined migration, emulation and system preservation as strategies for digital preservation, [14][15][16][17]. Each of the aforementioned traditional approaches or strategies suggest that preservation of information can be achieved by the use of technology.

In addition, the aim of the research of the Software Sustainability Institute was to raise awareness and build capacity throughout the Further and Higher Education (FE/HE) sector through engagement in preservation issues as part of the process of software development [18]. They provided a breakdown of the different approaches used in the achievement of software sustainability: technical preservation (techno-centric), emulation (data-centric), migration (functionality-centric), cultivation (process-centric) hibernation (knowledge-centric), deprecation, and procrastination. The success of each of these approaches depends on many factors: the importance of the software, the maturity of the software, the size of its community, and the availability of resources needed to achieve sustainability. The suggested framework can help groups of developers to understand and gauge the benefits of ensuring that preservation measures are built into software development processes and the benefits of active preservation of legacy software.

III. Our architecture

The purpose of TELDAP is to build a flexible architecture that emphasizes interoperability, continued maintenance, and continuous development. Most of TELDAP’s digital archives systems are software-based projects that include four elements: hardware, system software, application, and data. Based on the properties of these elements, projects can be divided into different sustainable levels for each element (e.g., service sustainability, software sustainability, and data sustainability). In Figure 1, the hardware and system software are categorized under service sustainability and the application and data are categorized under software sustainability and data sustainability, respectively.

In order to ensure digital archives sustainability, we composed a set of rules to manage the processes of a given software-based project. We formulated complete Standard Operating Procedures (SOPs) for digital archives sustainability for different levels and tasks: Digitization (DP); Development of Software Project (DSPP); Metadata Interoperability (MIP); Intellectual Property Inventory (IPIP); User Account and Single Sign-on (UASSP); Website Identification (WIP); Website Traffic Observation (WTOP); Catalog Index and Data Exchange (CIDEP); and Long-term Service Preservation (LSPP). As we can see in Table 1, each task among the different levels of digital archives sustainability needs different SOPs to maintain its integrity and interoperability. The service sustainability level includes hardware and system software. Therefore, the procedures of LSPP, WTOP, and UASSP must be considered when this project needs to host, validate, or migrate. The main procedure in software sustainability level is DSPP, which helps the content provider control the development flow of software. For data sustainability, the MIP and CIDEP are major procedures used in validation and migration tasks. Table 2 shows the combination of the framework of the Software Sustainability Institute and OAIS with our proposed architecture. In this paper, we will discuss DSPP, CIDEP, and LSPP in software sustainability, service sustainability, and data sustainability, respectively.


Digital Archives Sustainability
Hardware
System Software
Application
Data
Service Sustainability
Software/Data Sustainability
Environment, Platform, etc.

* Shadow means the state we may need backup procedure

Figure 1. The Architecture of Sustainable Digital Archives

TABLE I. THE SOPS WITH THE DIGITAL ARCHIVES SUSTAINABILITY

Levels / Task	Service Sustainability	Software Sustainability	Data Sustainability
Hosting	LSPP WTOP UASSP	Documents (Application)	Documents (Data format) IPIP, WIP
Validation			
Migration		DSPP	MIP, CIDEP

TABLE II. THE SOPS WITH THE FRAMEWORK OF SSI AND OAIS

Subjects / Task	Service Sustainability	Software Sustainability	Data Sustainability
Hosting	Emulation	Hibernation	
Validation		N/A	OAIS
Migration		Migration	

A. *Development of Software Project Procedure (DSPP)*

In this section, we aim at discussing the improvement of the quality and performance of the software development process and ensure flexibility, scalability, and completeness for the digital archive system. As shown in Figure 2, we applied the Verification and Validation in process area of CMMI—level 3 and ISO/IEC 15504 in order to integrate the development flow of software project with the participant and related output development documents. This procedure defined the development flow and documents to help the content providers (Digital Archives Systems) control the processes in each step and confirm the results when completed by the plan and development groups.

B. *Catalog Index and Data Exchange (CIDEP)*

TELDAP continuously archived over one million digital objects and approximately 100 websites with databases. In the process, TELDAP digitalized a large amount of data that contained information on a variety of cultural heritage, historical files, and archaeological artifacts. These are distributed as heterologous website systems. The task of integration, preservation, and popularization of these cultural objects from different websites that were constructed by different institutes is both important and enormous. As shown in Figure 3, we implemented an integrated platform for the cross-directories' knowledge retrieval platform on the basis of an intelligent crawler that used the behavior analysis of the digital archive users.

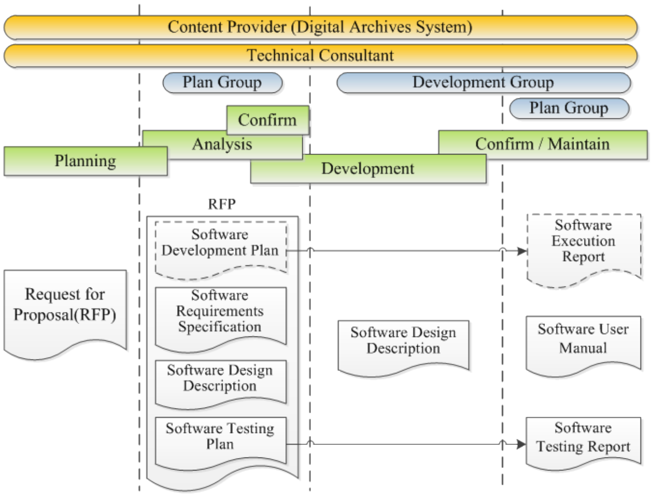

Content Provider (Digital Archives System)
Technical Consultant
Plan Group
Development Group
Plan Group
Confirm
Analysis
Confirm / Maintain
Planning
Development
RFP
Request for Proposal(RFP)
Software Development Plan
Software Requirements Specification
Software Design Description
Software Testing Plan
Software Design Description
Software Execution Report
Software User Manual
Software Testing Report

Figure 2. The development flow of software project

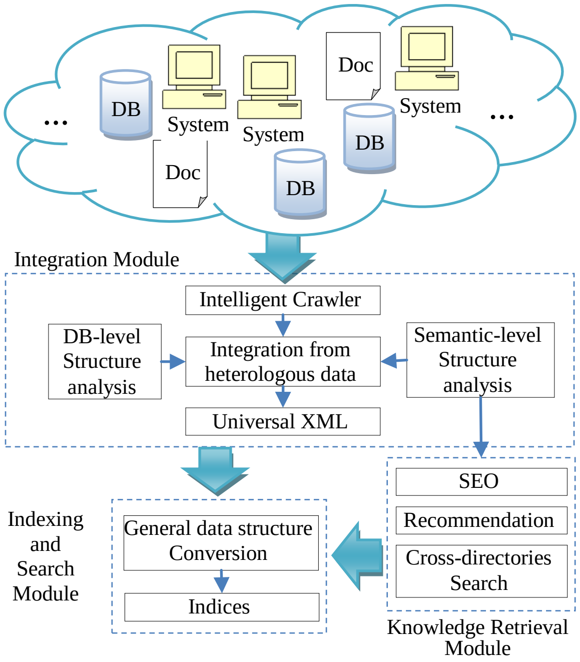

...
DB
System
System
Doc
System
...
DB
Doc
DB
Integration Module
Intelligent Crawler
DB-level Structure analysis
Integration from heterologous data
Semantic-level Structure analysis
Universal XML
Indexing and Search Module
General data structure Conversion
Indices
SEO
Recommendation
Cross-directories Search
Knowledge Retrieval Module

Figure 3. The architecture of the catalog index and data exchange

This framework used an intelligent crawler and the integration of heterologous data, indexing and search, and knowledge retrieval. In addition, it presented a user interface that offers search objects to the general public and provides opportunities for the public to study the information in these archives. Based on the achievements of this framework, individuals can easily enjoy the treasures of 5000 years of Chinese civilization. They can also learn about Taiwan's ecology and history.

C. *Long-term Service Preservation (LSPP)*

In system preservation, we must consider website migration, duplication, and emulation. Migration is used to ensure usability between generations of media. Migration also insures that a content object can be sustained in new generation media. Duplication is usually employed for the fault-tolerance of contents. Creation of an off-site backup can improve error recovery and ensure the availability of content. We believe that duplication is a basic element of preservation. Through emulation, a system can be maintained in a stable state. There are two types of emulation: static and dynamic. For a website that contains static pages, preservation by *snapshot* is recommended. When a crawler is employed, the contents of the site can be retrieved and saved by using a particular format.

For the dynamic-type system (e.g., page content within a database), we recommend the use of emulation or virtualization. An emulator is used to simulate hardware devices on the software platform. It causes the guest Operating System (OS) to run as a Personal Computer (PC). Virtualization, a state-of-the-art computer science strategy, is used to allow a guest OS to access (and/or share) hardware with a monitor (a hypervisor) on a host's OS. The choice of strategy involves trade-offs between performances, flexibilities, scalabilities, security, etc. Combinations of migration, duplication, and emulation can be used to fulfill the many requirements of digital preservation. As Table 3 indicates, these combinations can simultaneously meet the needs of a variety of websites and their availabilities.

TABLE III. STRATEGIES OF PRESERVATION

Method			**Description**
Duplication	**Emulation**	**Migration**	
V			The foundation of digital preservation
V	V		Snapshot: For static-typed website Virtualization: For dynamic-typed website
V	V	V	Ensure contents can be read in the future

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed strategies for the improvement of the reliability, overall representation, scalability, quick response, and friendly value-added environment for valuable archival objects. We hope to assist institutions involved in digital archiving to formulate a standard process for system analysis, planning, development, testing, and acceptance. In addition, we aim to provide technical consultation to these institutions to help them provide specifications and advice to outsourced vendors on the development of a standard digital archive system. We also provide guidance on the best way to draft a document to outline standard digital archiving procedures. Overall, we have provided a number of strategies that can help digital archiving institutions develop, manage, and preserve digital archive system sustainability

However, there are several areas for improvement. Until now, digital archiving institutions have faced many limitations in the areas of operation and innovative techniques. We need to develop new preservation strategies to address these new technologies. We must also provide ways for institutions to improve their operations so that they synchronize with standard processes. In the future, we hope to provide more developed plans for the improvement of the interoperability and integration of the wide range of available archiving systems. In doing so, we hope to preserve the integrity of current information for use in the future.

ACKNOWLEDGMENT

This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC101-2631-H-001-005, NSC101-2631-H-001-006 and Academia Sinica.

REFERENCES

[1] M. Hedstrom and S. Montgomery, Digital Preservation Needs and Requirement in RLG Member Institutions, Report, A Study Commissioned by the Research Libraries Group, Dec. 1998.

[2] F. Moore, “Long term data preservation,” Computer Technology Review, Third Quarter 1999, pp. 32-33.

[3] D. A. Kranch, “Preserving electronic documents,” Proceedings of the Third ACM Conference on Digital Libraries, May 1998, pp. 295-296.

[4] E. Cooly, and N. Chip, “Integrating solutions: examining the collection management process using OCLC's firstsearch electronic collections online,” Library Acquisitions: Practice & Theory, Vol. 22, No. 1, 1998, pp. 97-102.

[5] Taiwan e-Learning and Digital Archives Program, *http://teldap.tw*.

[6] TELDAP Resources Portal, *http://digitalarchives.tw*. 08.15.2012

[7] DigitalPreservationEurope, *http://www.digitalpreservationeurope.eu*. 08.15.2012

[8] RLG/OCLC, "Trusted digital repositories: attributes and responsibilities: an RLG-OCLC report," RLG/OCLC Working Group on Digital Archive Attributes, May. 2002, [Online]. Available: http://www.rlg.org/en/pdfs/repositories.pdf. 08.15.2012

[9] American Library Association, Definitions of Digital Preservation. Definitions of Digital Preservation, 2007, [Online]. Available: http://www.ala.org/alcts/resources/preserv/defdigpres0408. 08.15.2012

[10] I. Verheul, Networking for Digital Preservation. Current Practice in 15 National Libraries, IFLA Publication Series, [Online]. Available: http://www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf. 08.15.2012

[11] TG. Beamsley, “Securing digital image assets in museums and libraries: a risk management approach,” Library Trends, Vol.48, No. 2, Fall 1999, pp. 359-378.

[12] The National Archives, Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation, 2008, [Online]. Available:http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf. 08.15.2012

[13] Wikipedia, ”Open Archival Information System,” *http://en.wikipedia.org/wiki/OAIS*. 08.15.2012

[14] A. Waugh, R. Wilkinson, B. Hills, and J. Dell'oro, “Preserving digital information forever,” Proc. of the Fifth ACM Conference on ACM 2000 Digital Libraries, June, 2000, pp. 175-184.

[15] Andrew K. Pace, “Digital preservation: everything new is old again,” Computer in Libraries, Vol. 20, Issue 2, Feb. 2000, pp. 55-57.

[16] Wiggins, Richard. “Digital Preservation Paradox & Promise,” Library Journal, Vol. 126 Issue 7, Spring 2001, pp. 12-15.

[17] A. Muir, “Legal deposit of digital publications: a review of research and development activity,” Proc. of the First ACM/IEEE-CS Joint Conference on Digital Libraries, Jun. 2001, pp. 165-173.

[18] Software Sustainability Institute, *http://software.ac.uk*. 08.15.2012