Long-Term Preservation of Large-Scale Multimedia Collections: a Digital Preservation Workflow Approach

Richard Marciano[1], Chien-Yi Hou[1], Lynn Burnstan[2], Harry Kreisler[3], Reagan Moore[1], Arcot Rajasekar[1]
San Diego Supercomputer Center (SDSC), [2] University of California Television (UCTV), [3] University of California Berkeley – Institute of International Studies (IIS)

Abstract

The ability to incorporate and embed reusable preservation procedures within complex multimedia environments while minimizing the impact on production represents a substantial challenge.

The San Diego Supercomputer Center has been researching and prototyping preservation cyberinfrastructure for over a decade. In this latest work, we explore a digital archiving framework which focuses on the design, development, automation and reuse of preservation processes (including accessioning, description, arrangement, storage, and preservation) which can be embedded into existing digital production environments. This current work is funded in part by a joint Library of Congress and National Science Foundation program called Digarch, "Digital Archiving and Long-Term Preservation."

We prototype this approach with a collection called "Conversations with History" which is produced in a distributed workflow environment (University of California Berkeley with Harry Kreisler and University of California TV with Lynn Burnstan). The collection includes video, audio, images, text, transcripts, web-based material, databases of administrative and descriptive metadata, and contains diverse types of data, created at multiple stages within the content production workflow. Additional partners include the University of California San Diego Libraries. The test collection is 3.5TB in size (7TB when replicated across two different storage devices).

Introduction

For nearly 25 years now, Harry Kreisler at the Institute of International Studies at UC Berkeley, has been conducting interviews as part of the "Conversations with History" series [1]. Over 350 guests have been interviewed and 239 interviews are in the UCTV archive. Interviewees include diplomats, statesmen, soldiers, economists, political analysts, scientists, historians, writers, foreign correspondents, activists, and artists. These interviews are one-hour video-taped conversations. This significant "at risk" collection includes video, audio, text transcripts, web-based material, databases of administrative and descriptive metadata and contains diverse types of data, created at multiple stages within the content production workflow and stored partly offline, at UCTV and at UC Berkeley.

Lynn Burnstan at UCSD-TV [2] who runs the only broadcast television station operated by the University of California also launched UCTV [3], a 24-hour station offered on EchoStar Satellite's Dish network, sending programming from all UC campuses to millions of viewers across North America. UCTV broadcasts and houses the Conversations with History programs.

In the digital preservation project of the "Conversations with History" collection, we plan on demonstrating accessioning, description, arrangement, storage, and preservation processes, across both institutions and into a repository located at the San Diego Supercomputer Center. Our plan is to archive the 239 videos currently in the UCTV archives for long-term preservation at SDSC. The digital video formats include:

- Digital master files in DV format (.mov files) of typical size 12GB (compressed)
- UCTV broadcasting file in MPEG format of typical size 2GB
- Web archive files in Real Player format of typical size 200 MB

This makes the video content roughly 15GB per show or 239 x 15GB = 3.6TB for the complete collection. When preserving this content, we replicate the collection in at least two locations, making this a 7.2TB persistently archived collection.

__Figure 1__. "Conversations with History" Interview Series. A sample set of guests. Top row: Zhores Alferov, Lakhdar Brahimi, James Fallows, Natan Sharansky, Brian Urquhart, Steven Chu. Middle row: Amartya Sen, Amy Chua, Howard Zinn, Kenneth Waltz, Noam Chomsky, Massimo D'Alema. Bottom row: Leon Panetta, Ron Dellums, Robert McNamara, John Galbraith, Evan Hoffman, Hanan Ashrawi.

Related Work

In the UCTV/SDSC video repository, we store the master video content. As described in [4], Howard Besser proposes a very useful characterization: "In a digital environment, the concept of a "master" is likely to be more useful than the concept of an "original". Unlike the word "original", "master" does not necessarily convey physical embodiment. But it does convey the idea that this is the definitive version of a work, though not in such strong terms that it prevents the possibility of multiple variant forms."

The work pursued in our project relates to larger digital TV preservation efforts such as the NDIIPP-funded PBS/WGBH [5], where the long-term goal is to design a digital TV preservation repository. While the PBS/WGBH project's aims are to establish an inventory of at-risk materials in preparation for selection, to review the best practices in the field of video archiving, to help set standards and policies, and to ingest sample materials to test the repository, our objectives are complementary. We propose to develop a preservation workflow framework on a well-defined and self-contained "at-risk" video collection.

Other work such as the NDIIPP-funded Open Video Repository project at the University of Carolina at Chapel Hill [6], also looks at video preservation repositories, but with a special focus on context and interactivity in video browsing.

In that respect, our repository-building project is complementary to these projects. In summary, the overall aims of our project are:

- To make the preservation system as a (re) configurable workflow system that can be adapted to other multimedia production systems.
- To minimally disrupt the multimedia production system while integrating with the digital preservation workflow.
- To conform to standardized practices in digital archival preservation - eg, SIP-AIP package framework, metadata standards such as METS.
- To show that we can reproduce the user experience in the future in obtaining all the information about the production process.
- To capture all legacy information into the archival framework adapting the digital preservation workflow.

Digital Preservation Workflow Tools

In this project, we experiment with integration of preservation workflows into existing production video workflows. We explore two types of integration. The first is a passive program that can be invoked by the production workflow process whenever it reaches a particular point at which the preservation system needs to upload content or metadata. The second method is an active system that will monitor the "work spaces" of the production system and trigger content and metadata transfer when a constraint is satisfied.

This reusable preservation framework is based on two main components: the SDSC **Storage Resource Broker** (SRB) and the **Kepler** scientific workflow environment.

The SRB [7] data grid framework includes hierarchical collection management, interoperability among heterogeneous systems, access control supporting intellectual property rights, persistent naming through a logical name space, content and systemic metadata management, and aggregation methods for ingest, access, and metadata entry; deals with hardware and software obsolescence through technology migration; and supports workflow processing systems. SRB is currently being used as a preservation environment for multiple federally funded projects, including the NARA prototype persistent archive, the NHPRC Persistent Archives Testbed (PAT) [8], and the NSDL. SRB data grid technologies provide the preservation mechanisms needed to control and track the integrity of the archived data. This includes consistent and persistent management of both administrative metadata about each file (location, access controls, checksums, usage audit trails, ownership, versions, creation time) and descriptive metadata that may be specific to a collection. The SRB supports a standard set of access operations for file and metadata manipulation that can be used to support new digital library interfaces, such as DSPACE and FEDORA. The SRB supports organization of digital entities into collection hierarchies, making it possible to manage each preserved collection as an independent collection.

Kepler [9] is a scientific data workflow environment that allows a user to build workflows by dragging and dropping modules or components. These preservation workflows are made up of archival components (often implemented as web services) chained together using Kepler workflow diagrams. The modules used in this project include generic modules as well as grid-enabled (SRB) modules [10,11], which can be "stitched" together in reusable workflow descriptions. We will leverage current work implementing actors (processes) in Kepler that have been developed through efforts in building workflow systems for scientific applications. The actors will be used to provide interfaces to SRB and relational database systems.

Legacy Video File Preservation Workflow

While we develop workflows to handle new video programs incrementally added to the archival collection, our first test case is the handling of the legacy video files, with over 230 historical programs whose master file needs to be converted from tape to digital format. This is the main workflow we choose to demonstrate in this paper.

This workflow is called "legacy load workflow". Its function is to retrieve, transmit, and remove data. In order to run it as a routine, we need an additional workflow which we call *"alarm clock workflow"* to trigger the *"legacy load workflow"* every day or every month. With these two workflows, we can automate the preservation process without interrupting the UCTV production process.

Errors tend to happen when we upload or replicate data. We need to develop methodologies to detect errors and recover from them. We describe some of the sub-workflow concepts:

- **Calculating MD5 checksums:** In our workflow, we calculate the checksums of the files before and after each transfer operation. These values will be use to determine if the process is successful or not.
- **Feedback loops--persistent archive:** These feedback loops are used to restart the preservation process in case of failure.
- **Detecting failures and automatic fixes:** We compare the checksums to see if the files have been uploaded successfully. In case of failure, we use the feedback loop mentioned in the previous sub-workflow to restart the process. In this case, we can guarantee the data is persistent before going to the next step.

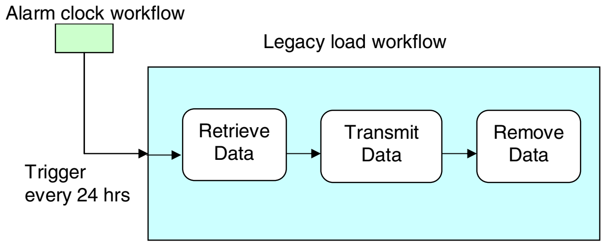
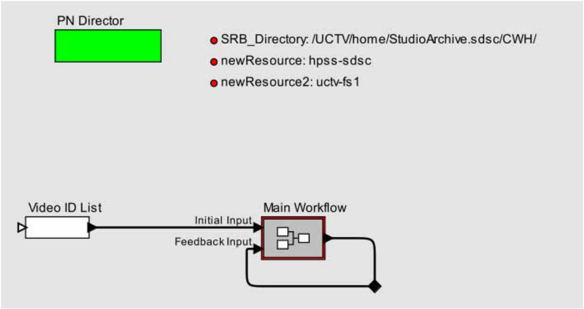


***Figure 2.** Conceptual legacy load workflow*

- **Cleaning up sub-workflow:** For both the UCTV local machine and the SRB staging machine, we need to clean up the disk spaces after successful transmittal and replication.



SRB_Directory: /UCTV/home/StudioArchive.sdsc/CWH/
newResource: hpss-sdsc
newResource2: uctv-fs1

***Figure 3.** Legacy Load Workflow*

UCTV manages a video ID list which contains all the video IDs of the "Conversations with History" program. Our first step is to read the list and check if all three types of files are ready (.mov, .rm, and .mpeg). Due to limited disk space, it is impossible to store all the files on the disk at the same time. The total size of these files is around 3.5TB. In Figure 3, we read one ready video ID from the list per round. If there are still ready video IDs in the list, we continue until all IDs are exhausted.

Figure 4 shows the most important part in our workflow, illustrating the consistency requirement as separate modules. From left to right, we use **Triplet Monitor** to check the availability of IDs, **Video File Retrieval** to get the files from some specific directories, *Safe_Sput* to put the files to SRB, *Safe_Sreplicate* to replicate files to some other resources, *Safe_Clean* to remove the files from the staging machine, and **Update ID List** to remove the IDs after they have been uploaded properly.

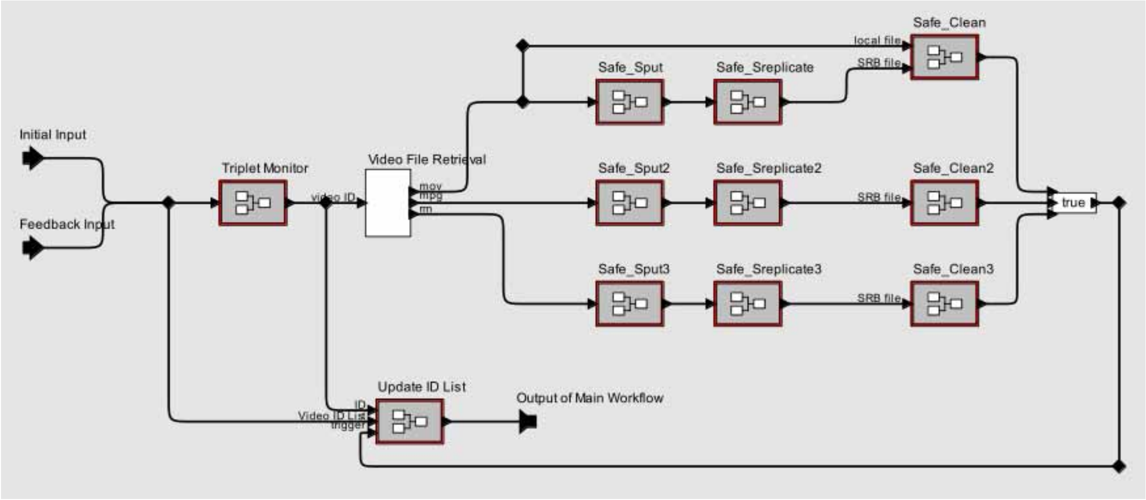


***Figure 4**. Main workflow*

Finally, we discuss the use of *Safe_Sput, Safe_Sreplicate*, and *Safe_Clean* in greater detail.

- **Safe_Sput:** Sput is an SRB actor used to upload files to an SRB-enabled storage resource. No transmission is guaranteed to be perfect, so we compute the MD5 checksum on the file before and after uploading. If the checksum values are the same, then we can proceed to the next step. If not, we use a feedback loop to re-do the Sput operation.
- **Safe_Sreplicate:** As in the case of *Sput*, errors can also occur when we try to replicate the files from one resource to another, especially when the file size is large. We use MD5 checksum again to ensure that the files are replicated perfectly. We also record these checksum values as metadata in the SRB MCAT catalog for further reference.
- **Safe Clean:** Safe clean means that we can purge the files from the staging machines safely. Ensuring the correctness of file removal is easier than executing *Sput* and *Sreplicate*. We simply need to verify that the MD5 checksums all match before removing the file.

Future Work

Future work involves the integration of these Kepler/SRB workflows with digital library services developed by Chris Frymann at UCSD Libraries. The services used are the UCSD Libraries XDRE (eXtensible Digital Resources Environment – pronounced "extra"). XDRE is a content management system implemented in Java and RDF-based.

Acknowledgements

This research was partly sponsored by the National Science Foundation and the Library of Congress through the NSF Digarch program. It also builds on the NSF-NPACI-ACI-9619020 (NARA Supplement) research and NHPRC Persistent Archives Testbed (PAT) project, which have both explored scientific workflows for preservation environments. Views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of the Government, or any person or agency connected with them.

This work is a collaboration between SDSC, UCTV, UC Berkeley and UCSD Libraries. SDSC partners also include Francine Berman (co-PI), Ilkay Altintas, Efrat Jaeger, and Lucas Gilbert. UCTV partners also include Steve Anderson, Matt Gaffney, Melissa Weber, and Bee Bornheimer. UCSD Libraries include Brian Schottlaender (co-PI), Luc Declerck, Brad Westbrook, Arwen Hutt, Ardys Kozbial, Chris Frymann and Vivian Chu.

References

[1] Harry Kreisler, "Conversations with History", http://globetrotter.berkeley.edu/conversations/

[2] Lynn Burnstan, About UCSD-TV, http://www.ucsd.tv/about.shtml

[3] Lynn Burnstan, About UCTV, http://www.uctv.tv/about.shtml

[4] Howard Besser, "Digital Preservation of Moving Image Material?" The Moving Image 1 #2, Fall 2001: pg. 39-56.

[5] NDIIPP-funded: "Preserving Digital Public Television", http://www.ptvdigitalarchive.org

[6] Gary Marchionini, "Preserving Video Objects and Context", Open Video repository project, http://www.open-video.org/project_info.php

[7] SRB, Storage Resource Broker, Version 3.1, http://www.sdsc.edu/dice/srb, 2004.

[8] Persistent Archives Testbed (PAT), http://www.sdsc.edu/PAT

[9] Kepler: A System for Scientific Workflows, http://kepler-project.org/

[10] Arcot Rajasekar, Michael Wan, Reagan Moore, George Kremenek, and Tom Guptill, Data Grids, Collections and Grid Bricks (MSST2003, 20th IEEE/ 11th NASA Goddard Conference on Mass Storage Systems & Technologies San Diego, California, April 7-10, 2003).

[11] Tim Wong, Developing Data Grid Workflows using Storage Resource Broker and Kepler (UC Davis), http://www.thwong.com/documents/SRB_Paper.doc

Author Biographies

Richard Marciano is Director of the Sustainable Archives and Library Technologies (SALT) laboratory at SDSC and Lead Scientist in the Data Intensive Computing Environments group. Richard holds degrees in Avionics, Electrical Engineering, and Computer Science (M.S. and Ph.D) and worked as a Postdoc in Computational Geography.

Chien-Yi Hou is graduate student in the Computer Science Department at UCSD and member of the DICE group. He holds a B.S. in Computer Science from the National Chiao Tung University(NCTU) in Hsinchu, Taiwan. His research interests include temporal querying in XML, query reformulation and schema evolution.

Lynn Burnstan, a documentary filmmaker by training, is the Managing Director of UCSD-TV and UCTV. She manages the entire staff and budget of the stations. Previously, she produced and edited award-winning films in Chicago, New York, Los Angeles, and San Diego for television, film festivals, corporations and commercials. She has a B.A. from UCSD and received her MFA from the School of the Art Institute of Chicago

Harry Kreisler is Executive Director of the Institute of International Studies at the University of California at Berkeley. He is also creator, executive producer and host of "Conversations with History", an interview program, broadcast nationally every Friday evening on Echostar Satellite's DISH Network, channel 9412, and on cable throughout California.

Reagan Moore is Director of Data Intensive Computing Environnments at the San Diego Supercomputer Center. He coordinates research efforts in development of massive data analysis systems, data grids, and persistent archives. Reagan has been at SDSC since its inception, initially being responsible for operating system development. He has a Ph.D. in plasma physics from the University of California San Diego and a B.S. in physics from the California Institute of Technology.

Arcot Rajasekar is the Director of the Data Grid Technologies Group at SDSC. He is a key architect of the SDSC Storage Resource Broker, an intelligent data grid integrating distributed data archives, file repositories and digital collections. Arcot received his Ph.D. from the University of Maryland and has published extensively in artificial intelligence, databases and data grid systems