Digital Preservation Plan

for the Texas Legacy Project

May 2005

Contents

About this Document

This document was created by Thomas P. Kiehne between January and May of 2005. Inquiries about this document should be directed to kiehnetp@mail.utexas.edu.

Revision History:

5/9/2005	Compilation of needs analysis, research, and draft plan; Final draft of complete plan
5/14/2005	Revised final draft

Acknowledgments

My thanks go out to the following people, without which, this plan would not have been possible:

David Todd, CHAT, for all information about and guidance from CHAT and the Texas Legacy Project.

Quinn Stewart, School of Information, University of Texas at Austin, for technical expertise in digitization and technology.

Pat Galloway, School of Information, University of Texas at Austin, for guidance and instruction in digital preservation and metadata.

Brenda Gunn, Center for American History, for information about the CAH.

Michael Lauter, GWH&A/Kinotonik, Boulder, CO, for a brief discussion from a producer's perspective that resulted in a critical insight about video formats.

Needs Analysis

The purpose of the needs analysis is to document information management procedures and resources available to the Conservation History Association of Texas (CHAT). In addition to this information, requirements and constraints regarding the creation of a preservation plan for CHAT's Texas Legacy Project information assets are included. The information contained herein was collected through in-person and email interviews with staff from CHAT, the Center for American History (CAH), and the School of Information (iSchool) at the University of Texas during the month of January 2005.

This section is organized into three general areas: a description of CHAT's structure and information assets, an overview of resources and constraints applicable to the preservation plan, and specific preservation and access requirements.

Conservation History Association of Texas

Organization, Purpose, and Goals

CHAT is a non-profit organization headquartered in Austin, TX and is composed of an executive board, advisory board, and operating staff. CHAT's Texas Legacy Project is an oral history project that uses narrative audio and video interviews to document a wide variety of environmental issues within the state of Texas. Interviews have been conducted with hundreds of legislators, scientists, educators, and individual citizens.

The information gathered for the Texas Legacy Project is oriented on six main themes:

- Personal profiles
- Conservation groups, sanctuaries, and policies
- Ecology and restoration
- Conservation history
- Conservation awareness in education
- Conservation awareness in media

In addition to these major themes, the Texas Legacy Project operates under the following explicit goals:

- Chronicle Texas environmental history
- Document ecological baselines
- Recognize individual conservationists
- Encourage a new generation of stewards

Information Assets

The primary information assets gathered and maintained for Texas Legacy consist of over 340 hours of audio and video interviews of typically one to two hours in length. From these, a number of information products are derived. At least one copy of the original digital tape and several analog and compressed copies are created in order to prevent catastrophic physical loss of the data. Additionally, segments of the full interviews are extracted, documented, then combined to produce longer, thematic videos.

The second largest body of information assets is the Texas Legacy Web site (www.texaslegacy.org). The Web site contains information about the project and descriptive information about the interviews. As such, the Web site is the primary public user interface for access to the project's materials. Compressed versions of the interviews, shortened segments, and thematic productions are available for streaming and download. Additionally, transcripts of many of the interviews may be viewed online.

Numerous analog and digital records document the interview material and project activities. 35mm photographic negatives are the source of still imagery for the Web site and other published materials. Microsoft Excel spreadsheets function as inventory logs and record metadata for the various information assets, some of which are also represented in a Web database. Text transcripts exist for many of the interviews, some of which are stored as Microsoft Word files on CHAT workstations, and the rest as HTML on the Web site. Some of these transcripts are also printed and stored at the CAH. A Filemaker Pro database manages contact data for narrators and other people involved with CHAT. Finally, paper records documenting correspondence, organizational activities, and project notes are maintained in the CHAT offices.

Current storage requirements for all digital assets are estimated at approximately 5 terabytes. A detailed inventory of analog and digital assets is shown in Appendix B.

Information Management

Numerous entities are involved in the creation and management of CHAT's information assets. Appendix C diagrams the relationships and exchanges between information generation, conversion, and storage processes. Each node (rectangle) represents an information process. The edges (lines with arrows) represent the transfer of an information object, either physical or electronic, and are labeled with the general type of information transferred. This diagram and the following description do not describe a comprehensive information audit for CHAT, but represent the predominant information activities.

Source videos are recorded at various locations by a team that includes a videographer (David Weisman, CHAT), the project coordinator/interviewer (David Todd, CHAT), a sound technician, and a lighting technician (node #1 on diagram). A mini-DV camera is currently used to record the interview, however Hi-8 format, and in one case VHS, was used for older interviews. As soon as possible after creation the mini-DV tapes are copied (#2). One copy is retained by David Weisman and stored in California. The other copy is stored at the CHAT office in Austin, TX. Other materials such as still photographs or audio recordings may also be produced during the interview.

Interviews are collected serially during regional tours of Texas. After the tour has concluded the

raw interviews are processed (#3). The video is transcribed at a later time into a Microsoft Word document, which is also stored on CHAT workstations and as HTML on the Web site. Robin Johnson, an independent contractor operating as "Keystroke Gurus" in Cedar Glen, California, transcribes the majority of the interviews. Transcriptions are also done by Judy Holloway at the CAH or David Todd at CHAT. Each interview is evaluated based on the quality of the production, informative value, and interest and is assigned a rank on an arbitrary scale of one through four (four being the highest). This evaluation, as well as descriptive data, keywords, and a unique index number (reel number) are recorded in a Microsoft Excel spreadsheet which is stored on workstations at the CHAT office. The same information is uploaded using a Web form into a MySQL database on the Texas Legacy Web server (http://www.texaslegacy.org/txdb/chatlog.php) which is maintained by the iSchool. Additionally, a separate database (http://www.texaslegacy.org/txdb/timeline.php) is also updated which contains a chronology of Texas environmental events, as learned from the interviewee and other sources.

The original mini-DV tape is dubbed to VHS video tapes and distributed one each to David Weisman in California, the CHAT office, the interview narrator, and the CAH in Austin (#4). CHAT materials are donated to the CAH under a Letter of Gift from May of 1999 and are made available for viewing by patrons on-site. Creation of VHS copies ensures that the interviews are available in what is currently a lowest common denominator format. Unfortunately, it also represents a reduction in generational quality. The original mini-DV tape and its first-generation copy remain as the only authoritative, high-quality records of the interview.

At some point after the interview is processed, a subject Web page is created using the metadata generated during processing (#5). The Web page typically displays a Web version of a still photo taken during the interview, a short description of the subject and some identifying and descriptive information pertaining to the source interview. As derivative products are created, they are linked to the subject Web pages and other index pages. These Web pages are at present manually coded by David Todd at CHAT, but could in the future be dynamically generated using information from the project's MySQL database.

On an irregular basis a copy of the mini-DV tape is loaned to Quinn Stewart at the iSchool for conversion into Web streaming format (#6). The raw Digital Video (DV) data is dumped to a workstation and saved as full DV quality Audio Video Interleave (AVI) files, from which Real Media compressed files are created. Two separate versions are encoded, one optimized for low bandwidth (lowband) streaming and another for high bandwidth (broadband) streaming. After conversion, basic metadata such as location and date are added to the Real Media file. Next, the encoded file is copied to the Web server and manually linked to the subject and index pages on the Web site by David Todd. Up until now, none of the raw AVI files of full interviews have been kept due to lack of drive space. Quinn has expressed the need to keep these files in the future since they not only represent a first generation copy of the originals, but keeping them will reduce the time needed to migrate or create derivative products in the future while reducing wear on the mini-DV tapes.

The iSchool is experimenting with a presentation method named Glifos (www.glifos.com) that has been used for one of the encoded interviews thus far and is expected to be applied to others in the near future. Glifos generates a Web-based presentation within the browser that links a

compressed video file to a corresponding transcript and topic index. A user of this product can control the video by clicking on timecode-marked links in the transcript or index, thus creating a novel method for searching and viewing online videos. Preparing a video for Glifos presentation involves creating additional HTML files that use the existing encoded presentation. Glifos can also generate XML files and Windows Media Player files viewable in Windows Internet Explorer browsers, or SMIL files with Real Media streaming video on a diverse array of browsers and platforms. Glifos files do not re-encode or contain the encoded videos, hence, they do not add significantly to the required digital storage. The Glifos developer environment is proprietary software, part of a complete library and digital library automation software system marketed by Infolib S.A. of Guatemala, but the output is human-readable XML and HTML which should be usable without the authoring software.

Full-length interviews are edited to create shorter segments for re-use in theme-based video productions (#7). These segments are created from the original mini-DV tapes by David Weisman in California. Segments of two to three minutes in length are culled from the complete interviews, additional identifying information is added, then the presentation is exported to mini-DV tape. Additionally, David Weisman encodes the segments in both Real Media and Quicktime format and copies the compressed versions to CD-R. The mini-DV tape is copied and sent to the Austin CHAT office along with a copy of the CD containing the compressed versions. The mini-DV tape is subsequently compressed into Real Media format at the iSchool and exchanged in the same manner as above. Quinn has recently begun to save the AVI originals used in converting these shorter segments and stores them on an external drive separate from the Web server.

The shorter clips are assembled into longer topic segments (#8). Each topic segment is based on a central theme or geographical region of the state and is about 30 to 60 minutes in length. The segment is assembled according to a script, then narration and a descriptive “leader” clip are added. Once production of the topic segment is complete, it is copied, compressed, and exchanged with the iSchool as above. VHS copies are also created and kept by David Weisman and distributed to David Todd and the CAH in Austin. Furthermore, several of the longer productions are encoded by David Weisman in Quicktime format, copied to DVD-R, and stored in both California and Austin. Additionally, a transcript for the topic segment is created in Microsoft Word and maintained on CHAT workstations in Austin.

The CHAT office and workstations, Texas Legacy Web site and iSchool drives, David Weisman's office and workstations, the CAH, and individual narrators comprise a distributed information repository for all of the project's information in its various forms (indicated by bold outlines on diagram).

Preservation Project Resources and Constraints

Current Information Technology Resources

CHAT

The CHAT office in Austin, TX, has several office-grade workstations connected to the Internet

via a residential, asynchronous DSL line. Specifically, there is an Apple Macintosh G3 (350 MHz with 2.8 gigabytes of free storage), a generic Windows PC (1.1 GHz with 53 gigabytes of free storage), an Apple Macintosh G4 (450 MHz with 15 gigabytes of free storage), and an external Firewire hard drive with 129 gigabytes of free storage. A DVD/CD burner and 100 Mb Zip drive are also available. These machines store correspondence and administrative project files.

David Weisman maintains an Apple Macintosh based Media 100 nonlinear video editing workstation at his office in Morro Bay, California. All derivative video products are produced there.

iSchool

The iSchool maintains a stand-alone server for the Texas Legacy project. The server is a Linux OS-based machine running Apache Web server, MySQL database, and PHP middleware. For compressed video streaming the server uses Real's Helix Server Basic, a free version which is restricted to a limited number of concurrent streams. There are currently two 250 GB Maxtor USB/Firewire external hard drives connected to the server which contain the Real Media files served by Helix. These drives are configured in a low-level RAID configuration for redundancy. A third, non-networked 250 GB Maxtor drive is used to store raw DV quality AVI files. The server is connected to the University of Texas (UT) network which can readily handle the bandwidth required to stream or download high quality video within the UT network or any comparable network connected to it.

CAH

The CAH provides access to VHS copies of the project's interviews, so the technology made available to the center's patrons is of particular importance to the project. Tapes may be played on VHS video cassette recorders on site and audio equipment is also available for replay of audio copies. The center currently has no DVD players for patron use, but one is available in the center's conference room and may be used by special request. Consideration has been given to procuring DVD players to provide patron access as more DVDs are acquired by CAH. The center maintains four network connected computer workstations for patron use, but these workstations do not have the ability to view CD-ROMs or other disc media. Some discussion has also been given to remedying this situation. The CAH currently has no digital preservation or access programs in place beyond that of making certain types of media available to its patrons.

Budget

The budget for the project is largely dependent on the needs determined in the course of developing the preservation plan. Since CHAT is a non-profit organization that relies on donations, it is preferred that the project costs be kept as low as possible, especially in avoiding excessive computer hardware costs. It is possible, however, that funding may be available through technology grants for digitization or digital preservation. Additionally, funding sources may be found in conservation or environmental grants focused on supporting oral history or documentation projects. Further research is required to discover such funding opportunities.

Directed Constraints

In addition to the above resources, a number of constraints have been suggested:

- Time: The project has been generating materials since 1997. Many of the interviews are with individuals who are no longer living or otherwise unreachable which increases the importance of such items. As a result, CHAT wishes to commence preservation activities as soon as possible to prevent loss of data.
- Longevity: CHAT recognizes that digital media are constantly changing and prone to obsolescence. Whenever possible, proprietary file formats, codecs, and software should be avoided in deference to open or common standards.
- Simplicity: Any new procedures introduced into CHAT's information management process should not require specialized skills that cannot be readily taught to existing staff. Automated processes are preferred as long as they reduce both effort and costs.
- Retention of rights: Prior to conducting the interviews, narrators sign a release for CHAT that grants full rights to use the material without further consultation with the interviewee. In turn, CHAT grants the narrator rights to use the material so long as attribution to CHAT is maintained. The transfer of CHAT materials to CAH, however, does not include the transfer of copyrights. The CAH, its patrons, and the University of Texas must seek clearance from CHAT to reuse materials, just as any other member of the general public. At the same time, CHAT would like to make these materials widely available without burdening access and re-use with clearance restrictions. Such a situation lends itself to the implementation of alternative licensing regimes such as the Creative Commons license. There is some trepidation, however, that controversial materials could be used against the interviewee or the activities of CHAT. Any release of source material under a rights or access scheme must permit CHAT to discourage the misuse of its materials while simultaneously encouraging as many constructive uses as possible.

Preservation and Access Requirements

Project Vision

The purposes for which the Texas Legacy Project exist are fundamentally forward looking and rely on the notion that the videos and supporting materials produced by the project will be available to future generations. A longitudinal comparison of current and future environmental issues will not otherwise be possible. Therefore, the Texas Legacy project requires an explicit preservation plan to achieve secure and accessible storage for all of the project's source materials and derivative products. The plan must secure against threats both in terms of physical destruction of source materials and digital threats such as obsolescence, degradation, and accidental loss. The preservation plan must also address contingencies for recovery from loss and ensure that the plan is periodically reviewed and updated in the future. Additionally, changes must be implemented in the near term to improve current workflow procedures while preparing for long-term preservation objectives.

User Definition

The primary user group for CHAT materials is defined by research information needs. These users include those from education, government, scientific research, and media backgrounds. Research needs are currently met by providing access to compressed videos and full text transcripts through the Texas Legacy Web site or alternately using the materials donated to the CAH.

A secondary user group may be defined by those producing original materials using the interviews and derivative products. Such users are characterized as news media (e.g.: radio and television) or documentary producers. It may be assumed that these users require access to high quality copies of source and derivative materials, however, no such requests have been made to date. For example, the Texas Parks and Wildlife radio program, "Passport to Texas," has used interview segments for their broadcasts, but used audio extracted from the compressed versions available on the Web site. Such may not be the case in the future and requests for high quality material should be expected.

A third user group may be characterized as using the CHAT archives as an experimental testbed for video digitization and access instruction. The large volume of original, high-quality video material collected by CHAT can be used as a corpus from which students in digital preservation and archives studies may draw from in the course of their studies. Knowledge gained from these instructional activities may serve to modify or suggest future preservation and access programs for CHAT materials.

Near-Term Objectives

As described earlier, CHAT ensures that multiple copies of source materials are made and physically dispersed in order to avoid catastrophic loss. There is still much that can be done to improve the survivability of, and improve access to these materials. Additionally, these improvements can be achieved with relatively low cost in the near-term and will help prepare for long-term improvements. Near-term objectives include:

- Increase copy quality: The workflow process previously described shows several points in the process of creating derivative materials where the copies lose one or more generations of quality as they are reproduced. The original and first generation mini-DV cassette tapes are the only remaining full quality records of the source material. In the unfortunate event of the loss of both mini-DV tapes, the source material can only be regenerated using inferior VHS copies or lossy, compressed Real Media files. Although the semantic content would be saved, the regenerated copies will have lost the fidelity of the original. To mitigate this possibility, the workflow process must be modified to improve the quality of the incidental copies that are made to a newer, higher-fidelity format.
- Increase first generation copies: In addition to improving the quality of second generation copies, the first generation digital data must be stored in a format that is less physically vulnerable than mini-DV cassette tapes. In the process of encoding these records for the Web, the data are transferred in perfect digital form to a hard drive. Until recently, most of these files have been discarded after the encoding is complete in order to conserve disk space. Plans must be made now to ensure that these full quality copies are maintained in

some form, not only to retain another first generation copy, but to reduce wear, and thus risk, on the original tapes.

- Improve metadata generation: Much descriptive data, often referred to as metadata, is already collected and indexed by CHAT. Although current practice may be sufficient for routine operations, consideration must be given to describing these materials more thoroughly for future reference. A long-term preservation system should ensure that these data are collected and retained as appropriate, but steps should be taken now to ensure that existing materials are well described and understandable by those not directly involved in their production. For example, technical information such as that regarding the conversion, copying, and migration of data must be captured to assist future technical efforts. Current metadata practices in the UT libraries and archives should be investigated in order to inform the development of CHAT metadata standards.
- Documentation of technical procedures: In conjunction with the development of technical metadata, the procedures used by David Weisman and Quinn Stewart for encoding, copying, conversion, and metadata enrichment should be documented. This documentation should take a more verbose form than that of the technical metadata, providing procedural instructions that serve to encapsulate the best practices used in the process of creating the various digital materials. In addition to passing on domain knowledge in this area, the materials may assist future migration or "digital archeology" efforts.

Long-Term Objectives

CHAT anticipates having a total of up to 500 hours of interview footage within five years. The near-term objectives described above will ensure that new materials are retained in the highest possible quality and that improvements begin for existing materials. Long-term objectives include:

- Increase available digital storage: Source materials should be maintained in the highest digital quality possible within the budget and resources available to CHAT. Unfortunately, the digital storage requirements for 500 hours of digital video is on the order of 7000 gigabytes of disk space, or 7 terabytes (assuming a 13 GB/hour conversion ratio and a safety factor). Such space requirements currently present a problem given the cost, but such concerns may not exist within five years as storage costs decrease and new physical storage media are introduced. The long-term plan must provide contingencies for a staged acquisition of and migration to new storage media. Additionally, any storage solution must be redundant, regularly backed up, and secure against network threats.
- Improve user access and preservation of digital objects: The current Web presence should be evaluated in terms of the role it plays in providing access to Texas Legacy's user base. A digital repository or library solution could be implemented to improve the user experience, as well as streamline information workflow and metadata management. The compressed videos that currently reside on the Web site can readily be integrated into such a solution and compression quality may be increased as storage and bandwidth capabilities improve. Currently available software should be evaluated in concert with evaluation of

digital storage options. The evaluation of software should also be informed by the recommendations set forth in the Open Archives Information System (OAIS) model (CCSDS, 2002).

- Formalize hosting arrangements: As new storage and software solutions are devised as described above, maintenance and handling of the Web server(s) will increase proportionally. The CHAT Web server is currently hosted and maintained by the iSchool, but future requirements may exceed the resources available there. The long-term plan must anticipate the possibility that new hosting arrangements could be required and prescribe requirements to assist in negotiating these arrangements.
- Improve quality of derivative products: In addition to improving the storage of source materials, products meant for outside consumption must be improved as new compression codecs and streaming formats become available. Such improvements will be incidental to the use of newer encoding and conversion products for new material, but the long-term plan must include contingencies for improving existing compressed products. Having the source material already in a high quality, digital form will ease these conversions. Additionally, these high-quality digital products should be copied to physical media as soon as high-capacity storage options capable of holding the data become available.
- Emergency planning: Specific procedures and responsibilities must be described in the long-term plan in the event that source or derivative materials are destroyed for some reason. Additionally, plans must be in place in the event that a current repository or service is unable or unwilling to continue to hold CHAT's assets or in the event of dissolution of CHAT itself.

Literature & Research Review

Having established requirements for the Texas Legacy digital preservation project, potential courses of action must be devised and selected. This section summarizes research into existing preservation practices and technical issues within the scope defined in the needs analysis. The feasibility and complexity of potential options may be assessed within the context of the findings in this review.

This section is organized into four areas: archival perspectives, digital file formats, digital storage media, and repository systems.

Archival Perspectives

Having only come into existence within the last 40 years, digital information presents many challenges to archival practice. This section presents some of the most pertinent digital archives research with respect to long term preservation of digital objects.

In a white paper produced by the Dutch National Archives, digital preservation is defined as "ensuring that records which are created electronically using today's computer systems and applications, will remain available, usable, and authentic in ten to one hundred years time, when the applications and systems which were used to create and interpret the record will, more likely than not, no longer be available" (Digital Preservation Testbed Project, 2001, p. 4). One of the most pressing challenges in digital preservation is characterized by the frail and dynamic nature of digital information. The accelerated rate of loss of digital data through obsolescence and neglect forces us to reexamine some of the basic assumptions of traditional archives in a way that takes into account the critical role of the underlying technology. Whereas traditional archives seek to preserve an original object for hundreds of years for access in its original form, this is not possible with digital information. For example, an English language archives holding printed materials may presume that the body of knowledge necessary to understand its holdings (e.g.: English language literacy) will exist well into the future. Unfortunately, with electronic records the underlying knowledge for comprehension includes not only language, but the machine-assisted translation of what are essentially arbitrary strings of binary data, encoded on storage media that may only be accessed with technology that utilizes similarly arbitrary conventions.

Two Digital Preservation Frameworks

The metaphor shift does not end with the essential characteristics of digital records, but suggests a reassessment of the entire archival process. Models presented by two major research initiatives represent such a reassessment. The first is the reference model for an Open Archival Information System (OAIS) recommended by the Consultative Committee for Space Data Systems (CCSDS, 2002). The OAIS framework is the result of a multidisciplinary effort to define archival systems in the context of dynamic information environments. The model is abstracted to the extent that it is applicable to the design of systems that manage digital records, analog records, or both, and that intend to preserve and provide access to the records over the long term.

At the top level of this abstraction, the OAIS model defines the system environment in terms of

three primary roles: producer, consumer and management (CCSDS, 2002, p. 2-2). Producers provide the information to be preserved. With respect to the needs analysis, the producer for the Texas Legacy archive is the Texas Legacy project team that produces the interviews and other materials. Management is responsible for the broader policy decisions concerning the archive, which includes obtaining funding and resources, and, in this case, is analogous to the Conservation History Association of Texas. Consumers interact with the OAIS service to obtain the information that they seek. These are the users defined in the needs analysis.

While the OAIS model provides a necessary abstraction of the functions of archives, much of the detail about how electronic records are maintained for the long term are omitted. The U.S. team of the International Research on Permanent Authentic Records in Electronic Systems Project (InterPARES) released its findings in 2002 (US-InterPARES Project, 2002). The InterPARES report is concerned primarily with defining the requirements for producing authentic records within an archival system, but another important aspect of the team's report is the preservation task force (PTF) model for preserving electronic records (US-InterPARES Project, 2002, p. 27). The PTF model describes preservation strategy beyond that of the OAIS model in that it takes into account the technical procedures for maintaining digital objects within the repository over time. In particular, the obsolescence of digital objects is addressed by one of three processes in the PTF model: migration to current formats, conversion to standard formats, and conversion to hardware independent representations (US-InterPARES Project, 2002, p. 31). The Dutch National Archives report expands on these methods, outlining seven strategies that may be used for long term preservation of digital objects (Digital Preservation Testbed Project, 2003, pp. 6-8):

- Technology preservation – Preserve the technology required to access original records for as long as those records are required
- Printing to paper
- Emulation – Preserve not only the record, but also an emulator specification which contains enough details about the original environment for that environment to be recreated on a future computer when necessary
- Encapsulation – Retain the record in its original form, but encapsulate it with a set of instructions on how the original should be interpreted.
- Virtual machine software
- eXtensible Markup Language (XML)
- Migration and storage in standard formats

Despite the differences between the two frameworks, the OAIS and InterPARES PTF models agree in several key criteria:

- Dynamic digital environments require that archival considerations be addressed as close as possible to the source of information creation in order to support authenticity and access
- Archival processes must be technologically agnostic in that they are able to adapt to rapid changes in information technology without losing record integrity

- The generation and maintenance of metadata is crucial in the support of long term preservation
- An archival system is fundamentally user-oriented

The Role of Metadata

Metadata is commonly associated with descriptive data about a record or object, however, in order to support long term preservation efforts, description is only the beginning. Gilliland-Swetland (2000) defines five types of metadata for preservation:

- Administrative – Includes rights information, access restrictions, accession information, and locations of associated objects within the repository.
- Descriptive – Includes bibliographic information (author, title, etc.), identifying information, and relationships to other records.
- Preservation – Includes documentation of preservation actions such as migration and conversion.
- Technical – Includes documentation of required software and/or hardware, format information, and authentication information.
- Use – Includes user access data and information about derivative versions.

These data may be retained by a repository system in a variety of ways. Descriptive and technical metadata may be embedded in the digital object itself, the data may be stored in a separate record that refers to the digital object by a unique identifier, or it may be inherent in the system design (e.g.: Web server access logs; databases). Likewise, these data must be collected constantly throughout the lifetime of an object, from the point of its creation through its use in the repository. Taken together, these types of data comprise the information necessary to ensure that digital objects remain intact and accessible over time.

XML is the underlying technology for a number of metadata standards. XML metadata is typically extracted from databases or standalone records and used for exchange with other systems. Additionally, XML may form a wrapper around digital content in a repository, thus not only describing the record but providing a means of exchange. One of the most widely used but least stringent standards is Dublin Core. Dublin Core provides a flexible set of metadata elements that is primarily used for resource discovery and description (Wactlar & Christel, 2002, p. 4). Given the highly technical nature of digital video, a number of metadata sets have been developed particularly for audio and video. These include MPEG-7 and MPEG-21. These sets provide a richer array of technical and structural description than Dublin Core, and can accommodate the abstraction of audio-visual media description (Wactlar & Christel, 2002, p. 5). These standards were the most commonly used among the archives and repositories described later in this report.

Digital File Formats

Subsequent discussion of storage media and repository systems will be informed by having an

understanding of the basic structure of digital video information. According to the OAIS model, information is a combination of data and representation information (CCSDS, 2002, p. 4-19). This distinction is typically conflated for physical information objects such as books, but is significant for digital information where the data are represented by magnetic or optical impulses that undergo a series of translations to become usable by humans. Representation information is essentially the file format or formats in which the data are encoded by the software used to create the digital object.

For most types of digital data there is a one-to-one relationship between the data file and its file format. A Microsoft Word or Adobe Portable Document Format file requires only one set of instructions to convert the raw bitstream into a viewable representation (actually, there is quite a bit of simplification here in that several other transformations are performed by the hardware and operating system, but since these are common to all operations, the distinction may be omitted for this discussion). Multimedia information, such as the videos that make up the majority of the Texas Legacy project's information, requires several formats for interpretation.

Figure 1 diagrams the process of creating and transmitting digital video over a network. In the first stage of the process, the video and audio are recorded by a camera and encoded to tape. At this stage, the digital and audio data are encoded in one of the various broadcast formats. These formats include analog transmission standards such as NTSC or PAL, and digital formats such as DV, Beta-SP, and HDTV. At some point after the video is recorded, the data will be transferred by a direct connection to an encoding workstation. Software on the workstation will convert the broadcast format to video and audio codecs (shorthand for compression / decompression) that are preferred by the user or as a default working setting for the software. The converted audio and video data is edited then saved to disc. At the time the data is saved, the video and audio, in codec form, is encoded again, this time into a file format (not to be confused with the broadcast formats above) that packages the data into a form that the computer can manipulate. Essentially, a container, or "wrapper," format is used to package the data for exchange while the codecs define representations of the audio and video data contained within. Software on the client workstation will unpack the file format container in order to access and translate the encoded audio and video data within. It is these additional levels of structure that make long term preservation of digital audio and video more complex than that of discrete documents.

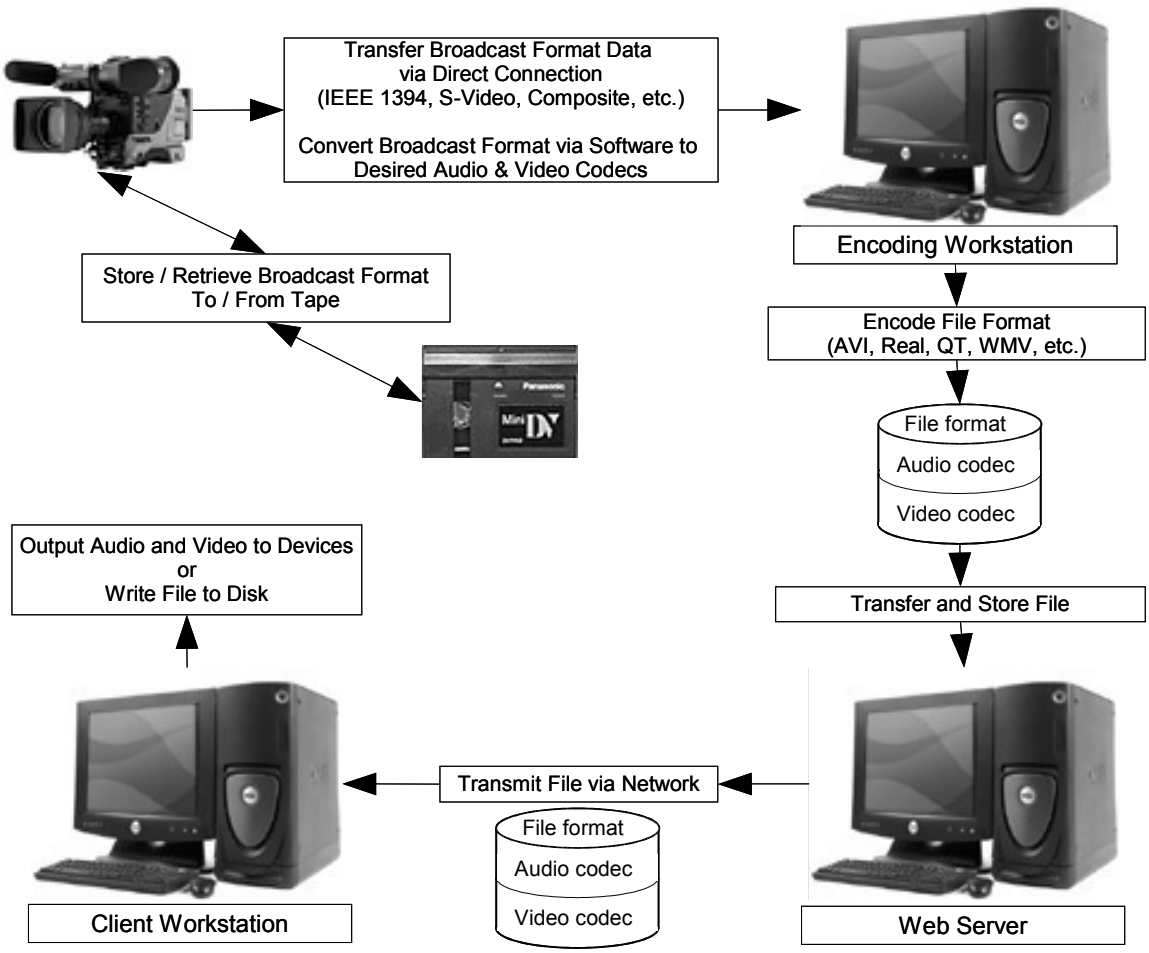


Figure 1: Digital video encoding and transmission process

Assuming that a migration strategy is employed to ensure access to Texas Legacy videos for the long term, there must be a thorough understanding of the container and codec formats currently in use and those formats that are being developed. Unfortunately, finding authoritative information about digital file formats can be difficult, requiring one to gather information from individual standards bodies, vendors, and third-parties (Lawrence, 2000, Access to Format Data). Centralized file format registries are being developed that will help ease this process, including the Global Digital Format Registry (GDFR, http://hul.harvard.edu/formatregistry) and PRONOM (http://www.nationalarchives.gov.uk/aboutapps/pronom), among others (Univ. of Leeds, 2004, pp. 40-41). Another notable source is the National Digital Information Infrastructure and Preservation Project (NDIIPP) of the Library of Congress which provides detailed information about file formats in a preservation context (NDIIPP, 2005). Each format is evaluated in terms of seven sustainability factors:

- Disclosure – The degree to which complete specifications and tools for validating technical integrity exist and are accessible
- Adoption – The degree to which the format is already used by the primary creators, disseminators, or users of information resources

- Transparency – The degree to which the digital representation is open to direct analysis with basic tools
- Self-documentation – Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata
- External dependencies – The degree to which a particular format depends on particular hardware, operating system, or software for rendering or use
- Impact of patents – The degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents
- Technical protection mechanisms – Implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository

Current Video File Formats

The digital video formats currently in use by Texas Legacy are summarized below with respect to responsible organization, quality, metadata, and protection schemes (NDIIPP, 2005, Format Descriptions).

Audio Video Interleaved (AVI)

AVI (file type extension: .avi) is a wrapper format that may contain a number of different codecs. The name of the format refers to the alternating of audio and video data segments within the bitstream. The format was jointly developed by Microsoft, Inc. and IBM at the beginning of the 1990s. Texas Legacy currently uses AVI as an intermediate format for editing and as a storage format for selected video clips using low-compression DV codecs for video. AVI is widely used for film and video production owing to its good resolution and clarity as well as its low compression rate. Limited technical and descriptive metadata may be encoded within the file header. AVI is likely to be a common format for video editing and storage for some time.

RealVideo

RealVideo (file type extensions: .rm, .rv, .rmvb, .ram) is a file container and codec developed by RealNetworks, Inc. Texas Legacy currently uses Real as the primary format for distribution to end users over the Internet. The format, currently at version 10, is a closed, proprietary format with little disclosure of technical documentation, although some licensing arrangements exist for open source and research development. RealVideo uses a "lossy" compression scheme optimized for Internet streaming applications. Audio and video quality is moderate to good, depending ultimately on network conditions while streaming. Basic descriptive and technical metadata may be embedded within the file. RealVideo may only be decoded by software distributed by RealNetworks due to authentication instructions embedded by the production software. Additionally, the creator of the file may activate copy protection that prevents the stream from being copied by the end user. RealVideo is one of three dominant, proprietary streaming formats, and as such, the long term viability of the format is largely dependent on market forces.

QuickTime

QuickTime (file type extensions: .mov, .qt) is a wrapper format developed by Apple Computer, Inc.. The format may contain any of a number of modified codecs including special versions of MPEG and MPEG-4. QuickTime is a proprietary but openly documented format. Video quality depends on the codec that is used. Limited descriptive metadata may be encoded in the file header. Files may be configured to require that a key be entered before playback. QuickTime is one of three dominant, proprietary streaming formats, and as such, the long term viability of the format is largely dependent on market forces.

Windows Media Video (WMV)

Although the Texas Legacy Project does not currently use WMV, the format is worth mentioning because it is currently available and fully implemented. WMV (file type extension: .wmv) is a wrapper format developed by Microsoft. Several codecs may be enclosed within the WMV format, all of which are also Microsoft codecs. WMV is a subset of the Advanced Systems Format (ASF), which is a proprietary format developed by Microsoft for which documentation is openly available. The format is a compressed format that offers good quality, but may be upgraded with "pro" codecs to achieve full HDTV resolution. Descriptive metadata is encoded in the file header and each file receives a unique ID number upon creation. Rights metadata and encryption may be added that require license validation before content may be played. WMV is one of three dominant, proprietary streaming formats, and as such, the long term viability of the format is largely dependent on market forces. However, the dominance of Microsoft in the personal computer market is likely to ensure a relatively long life for WMV and its successors.

Emerging Container Formats

The formats just described represent dominant standards for both video authoring and streaming. The NDIIPP also offers a glimpse of emerging formats whose features may prove to be better for long-term preservation of DV quality video (NDIIPP, 2005, Format Descriptions).

MPEG-4

MPEG-4 (file type extension: .mp4) is an ISO standard wrapper format developed by the Motion Picture Experts Group (Note: MPEG is an acronym used to note a variety of standards designated by the Motion Picture Experts Group; This includes standards for video, audio, and metadata). MPEG-4 was based initially on the QuickTime format, but greatly extends and improves upon it (Video Development Initiative, 1999, p.12). MPEG-4 represents a variety of digital objects in an object oriented, binary format. The wrapper format is an open standard with extensive documentation. The video codecs available present video in comparable quality to MPEG-2 (used in DVDs) up to HDTV resolutions, and audio quality that is considered superior to that of MP3 (MPEG-2 Layer 3 audio). Additionally, MPEG-4 supports embedded MPEG-7 metadata and an intellectual property protection interface. The format defines two tracks for additional data, one for text and one for metadata that allows indexing at the frame level (Video Development Initiative, 1999, p.11). Unfortunately, licenses for the audio and video codecs apply to hardware and software manufacturers, but may also levy fees for content providers

based on the number of users or extent of content delivered. These licensing constraints may slow the adoption of the format.

Advanced Authoring Format (AAF) and Material Exchange Format (MXF)

AAF (file type extension: .aaf) and MXF are two interrelated formats that are still in development. Both formats are interoperable and provide an object oriented container framework for video, audio, metadata and other bitstreams. MXF has been described as a generic implementation of AAF that may be intended for delivery of content over networks, while AAF is positioned as a storage format for production and archiving. Support for multiple codecs is expected, but has yet to be documented. Metadata support is still in development, but is expected to provide extensive structural description in addition to descriptive and technical information. MXF is an open, platform independent standard sponsored by the Society of Motion Picture and Television Engineers (SMPTE), while AAF is developed by the Advanced Authoring Format Association and is built partly on Microsoft technologies for object and structural definitions. As such, AAF could become encumbered by patent or license regimes or platform-specific dependencies.

Theora / Ogg Vorbis & VP3

Theora is an open source video format developed by Xiph.org that is meant to compete with MPEG-4 and other leading video formats. Theora is a superset of the VP3 video codec that will be combined with the Ogg Vorbis audio codec within an Ogg multimedia container format to form a completely non-proprietary video format. The format will be released under a BSD-style license. Although some portions of the VP3 codec are patented, Xiph.org has negotiated a perpetual, free license for use in Theora. The format is currently in development, but the bitstream specification is stable and products are beginning to support it (Theora.org, 2005, FAQ).

Digital Storage Media

Storage media are another component of digital archives that are susceptible to obsolescence and failure. In order to understand the risks involved, it is helpful to divide storage media into two categories: active storage and passive storage. Active storage refers to the property of the storage media to be readily accessible by a system or network. Active storage consists of hard disk drives, drive arrays, or other networked storage devices that provide virtually immediate access to stored data using a random access file system. Active storage is typically less susceptible to format obsolescence because the hardware standards, such as SCSI and IDE, and the file systems, such as FAT and NTFS, change less often and are usually well documented and broadly supported. Additionally, data stored on active media may be copied quickly once improved storage media becomes available (Wachtlar & Christel, 2002, p. 11). Unfortunately, active media are always operational and require constant monitoring. Additionally, guarding against failure through redundancy multiplies the cost of active storage as capacity is increased.

Passive media refers to the need for intervention to make the contained data accessible. Passive media includes any removable device, such as magnetic tapes and disks, optical disks, magneto-

optical cartridges, and so on. Although they have slower transfer speeds than active media, passive media typically have higher per unit storage capacities than active media devices (Baumann, 2002, p. 4). Because of the large capacities and slower access speeds, passive media are often used as backup storage for active media devices or for long-term, inactive storage of data. Removable media are more susceptible to degradation over time: among other things, tapes stretch and may break with use, discs may scratch or delaminate if stored improperly, and humidity may corrode data surfaces (Wachtlar & Christel, 2002, pp. 10-11; Van Bogart, 1995, pp. 4-8). Estimates for latent tape lifespans ranges from 10-30 years, depending on a number of factors including care & handling, storage environment, and data throughput (both read & write). Additionally, tapes used as regular backup media must be replaced much more frequently than those used for long-term retention of data. Lifespans for optical media are typically much longer than for tapes, but may be significantly reduced in the case of frequent re-use of rewritable optical media.

Migrating between media or converting to active media is a time consuming process. The capacities, compression schemes, and physical containers for such media are constantly changing and are susceptible to market failure (Lindner, 1996, ¶ 6 & 7). As a result, there is a litany of obsolete passive storage media formats that rivals that of obsolete digital file formats (Cornell University Library, 2003). Unfortunately, whereas obsolete file formats only require the proper conversion software to continue access, obsolete media requires that both hardware and software be maintained. In general, hardware obsolescence is likely to be more of a factor than media lifespans for determining passive media service life.

Given current storage media capacities and the amount of data represented by the Texas Legacy archive, storage cost is of primary concern. The InterPARES report suggests a Hierarchical Storage Management (HSM) scheme to balance the cost of storage against access concerns (US-InterPARES Project, 2002, p. 45; See also Figure 2). At the top of the hierarchy are active storage media, representing higher cost but faster access. According to archives policy, less critical or less accessed objects will be migrated downward in the hierarchy to less expensive, passive media. This system may also be seen in terms of quality of service, that is, providing immediate access to smaller, acceptable quality surrogates at the top of the hierarchy while retaining larger, full quality objects and backup copies on passive media towards the bottom.

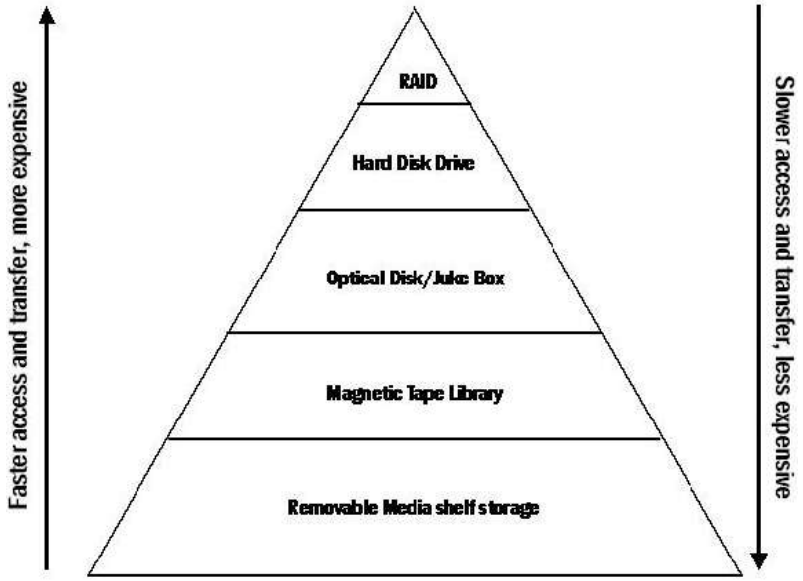


Figure 2: Hierarchical Storage Management (InterPARES, 2002, p. 45)

Current Mass Storage Technology

If the multiple terabytes of data created by the Texas Legacy project are to be made available in any active form, whether as first generation copies or as products optimized for the end user, a mass storage solution will be required. Several current technologies are described below.

Network Attached Storage (NAS)

NAS is a scalable storage infrastructure that relies on a TCP/IP network for access to data storage devices. A typical NAS server will contain a large disk array, usually in a RAID configuration (see below). NAS servers are accessed via a network through a special operating system known as a network file system (NFS), a minimal system that handles input and output operations for the disk array and network, but provides no other application services (Spalding, 2003, pp. 36-37). Such a configuration provides easy access to large bodies of data by multiple users or application servers and centralizes data backup resources. NAS is ideal for applications that require file-based or read only access to files, or for data that must be accessed by multiple application servers. Unfortunately, applications that require specialized, direct access to disks, such as databases that write data directly, may require special configuration in order to work with an NAS architecture (Spalding, 2003, p. 44).

Storage Area Networks (SAN)

SANs use optical fiber channel communications to effectively simulate internal devices for multiple servers at once. SANs are typically used for data warehouse and multimedia storage applications. Current SAN configurations use a switched fabric network or an arbitrated loop to create a high bandwidth (typically 100-200 Mb per second) internal network for any number of application servers (Troppens, 2004, pp. 78-85). Fiber channel networks may be connected to SCSI devices using special hardware, thus expanding the types of storage devices that may be

used. Both disc based and tape based storage devices may be connected to a SAN network, providing a heterogeneous storage environment that can address different storage needs within the same storage-application infrastructure. Unlike NAS, however, the input/output infrastructure of a SAN is more difficult to work with unless specialized hardware or software is used to simulate different logical configurations. Additionally, fiber channel device costs are high owing to their relatively recent introduction (Spalding, 2003, pp. 56-62).

Redundant Array of Independent Disks (RAID)

Multiple hard drives may be connected sequentially to a computer, but such a configuration does not protect data against failure of one or more of the devices. To guard against device failure, a disk array may be configured using hardware or software into a RAID configuration. RAID assembles many drives into one virtual drive by partitioning data across the disk array. Because of this redundancy, defective drives may be replaced without losing access to data in the interim (Spalding, 2003, pp. 87-90). RAID service levels are characterized by the configuration of partitions and devices. Each RAID level writes data to the physical devices in different ways and, as such, has different levels of fault tolerance, read and write performance, and space requirements (See Table 1). The performance of a particular RAID configuration is ultimately dependent upon the hardware or software implementation as vendors may choose to improve upon or not implement the basic specifications (Troppens, 2004, pp. 20-21). RAID arrays can be configured as a disk subsystem for direct server storage, or as a separate device for NAS or SAN (Troppens, 2004, p. 14). Many RAID devices are available commercially in varying storage capacities into the tens of terabytes. Arrays may also be built from commercially available parts at relatively low cost. With current hard drive prices ranging from $0.50 to $1.00 per GB, multi-terabyte RAID servers typically cost from $1.25 per GB or more.

RAID Level	*Fault Tolerance*	*Read Performance*	*Write Performance*	*Space Requirement*
RAID 0	none	good	very good	minimal
RAID 1	high	poor	poor	high
RAID 10	very high	very good	good	high
RAID 4	high	good	very, very poor	low
RAID 5	high	good	very poor	low

Table 1: Qualitative comparison of different RAID configurations (Troppens, 2004, p. 34)

Magnetic Tape

Magnetic tape was the first electronic data storage medium. As described above, magnetic tape typically has a higher per unit capacity than disk or optical media. Data is stored sequentially which accounts for slower access times relative to disk drives. Numerous large capacity tape storage formats are available, including Digital Linear Tape (DLT), Advanced Intelligent Tape (AIT), and Linear Tape Open (LTO), which can store compressed or uncompressed data in a variety of capacities of up to 800 GB each. Multiple tape drives may be connected using

controller hardware and software to function as a single mass storage device. Such devices, called tape libraries or auto-loaders, may be configured to provide error correction, buffering of data, and redundancy similar to RAID arrays (Spalding, 2003, p. 90). Unlike disc drives, the cost of tape storage is divided between the tape device and the media. Tape media costs will depend on the type of tape used. Current commercial retail prices for high capacity tapes, including DLT, SDLT, LTO, and AIT, cost from two to ten cents per GB. Tape device costs vary greatly depending on the type of tape device and the configuration. Single tape drives cost in the hundreds of dollars each while all-in-one multi-terabyte tape libraries or auto-loaders cost into the tens of thousands of dollars.

Optical Media

Other than the differences in disk surface and access, optical media devices may be configured in much the same way as disk or magnetic tape devices. Optical media libraries and auto-loaders are available that function identically to their magnetic tape counterparts, using Compact Disc (CD) and Digital Video Disc (DVD) media with less per-item capacity than tape. Unlike tape, however, optical media have the benefit of random data access via a filesystem, thus behaving somewhat like write protected hard disks (Troppens, 2004, p. 279). Writable CD and DVD media are currently available that store compressed or uncompressed data at capacities from 800 MB and 4.7 to 9.0 GB respectively. The main challenge currently facing DVD storage formats is that of competing and incompatible formats (Baumann, 2002, p. 8). Additionally, newer but less ubiquitous disc media formats are available that can carry several times the amount of conventional discs (Glossary Tech, 2005, Disc), but their longevity and market tenacity is questionable. Despite the smaller capacity of optical media, mass storage applications for audio and video are becoming more common (Spalding, 2003, p. 94).

Emerging Mass Storage Media

Disk arrays are expected to comprise the dominant active storage media for the foreseeable future. An upper bound for magnetic disk media capacity has been identified, but the technology that will replace it is as yet unknown (Baldi, et al., 2005, pp.49-50; Jones, 1997). In the meantime, upgrading active disk arrays is a relatively simple process of copying data from one device to another over a network.

Magnetic tape is continually increasing in capacity and marginally in speed. However, tape media will continue to be slower than disk arrays or optical media due to its sequential access limitations. Optical media, on the other hand, is expected to match and exceed the capacities of magnetic tape in the near future, thus providing promising new migration paths for mass storage. Three recently developed technologies are discussed below.

Blu-Ray Disc (BD) & HD-DVD

Blu-Ray media uses smaller wavelength blue-violet laser light, instead of the red light used for CDs and DVDs, to achieve a higher density of data storage on standard size discs. BD holds 25 GB per layer in one or two layer configurations, resulting in 25 or 50 GB per disc. This translates to approximately 13 hours of standard definition video or 2 hours of HDTV video per 25 GB disc. Prototype devices exist, but devices and media are expected to become available

commercially by 2006 (Blu-ray.com, 2005).

Blu-Ray is one of four major formats that are competing for the high definition DVD market. The others include another blue laser format supported by NEC and Toshiba, and two red laser formats that use compression techniques to increase capacity (DVD Town, 2005). Whether the market competition in the high definition disc market will stabilize into a single format is as yet unclear. The fact that a dominant format has not yet emerged suggests caution before selecting such media for migration, especially since even larger optical disc capacities may soon overtake these formats (see below).

Multiplexed Optical Data Storage (MODS)

Researchers at the Imperial College of London introduced a disc technology that diverges from the binary representation of data on discs. Instead of each "pit" representing a single bit of data, the pits are angled to return multifaceted reflections that can represent multiple bits of data. This technique can yield up to 250 GB per layer on a standard CD sized disc. A double sided, double layered disc could hold up to one terabyte of data (Imperial College London, 2004). No commercial implementations are known to be in development at this time.

Holographic Versatile Disc (HVD)

Optware, Inc. announced an implementation of a laser interference technique, called Collinear Holographic Data Storage, for encoding data on standard sized discs. Initial reports indicate that up to a terabyte of data could be encoded using such technology with high transfer speeds of up to one GB per second (Optware, Inc., 2004). Products using this technology are in development with planned release in 2006 at capacities up to 200 GB per disc. One terabyte models are planned for the future (Kallender, 2004).

Repository Systems

Any discussion about digital preservation is incomplete without a discussion of the information systems necessary to keep track of and provide access to the stored digital objects. This section covers two general classes of repository systems: academic & customized video repository projects and discreet repository software systems.

Digital Video Archive Projects

Over the past ten years, several notable digital video projects have been initiated, either as an academic or research initiative or as a response to a specific need. Most published articles discovered in the course of this research treated about proof of concept projects in retrieval, storage, and indexing of digital video. Several projects were discovered, however, that are broad-based research initiatives that are similar in scope to that desired by the Texas Legacy project. Below is a brief discussion of each project that focuses on creation, storage, and indexing of digital video materials and, when possible, the underlying technology.

Informedia Digital Video Library

The Informedia Digital Library was created by Carnegie Mellon University in 1994 as a testbed for developing educational digital video retrieval systems (Christel, et al., 1995; Informedia, 2005). The initial phase of the project encoded over 1000 hours of VHS quality news and documentary video in MPEG-1 format, accounting for about one terabyte of storage. Primary research goals included development of automated techniques such as speech processing, video OCR, and image recognition for the development of search indexes and metadata (Wactlar, 1999). The testbed is still in operation and has provided technology for similar digital libraries, such as the European Union's ECHO project (Informedia, 2005, Collaborations), and source materials for the Open Video Digital Library (described below).

Visual History Archive

The Shoah Foundation's Visual History Archive (VHA) contains 180 terabytes of oral history interviews with World War Two Holocaust survivors (Shoah Foundation, 2005). A proprietary retrieval system was developed for the project to store and access MPEG-1 encoded videos via select Internet-2 nodes. The source videos were produced using Sony Beta SP, then encoded to MPEG-1 for online access. A Digital Betacam preservation copy of the source tape is made and stored off-site and two VHS access copies are made available for on-premises viewing. The VHA system implements a hierarchical storage regime similar to that described by the InterPARES report. Online storage is provided by a one terabyte active storage array that is served by a 400 terabyte robotic tape archive using 50 GB (100 GB compressed) AIT-2 cartridges. If a requested video is not available on the local or network disk caches, the robotic unit retrieves the cartridge and transfers the requested data to the local disk cache. Typical access times for tape transfer ranges from five to ten minutes (Gustman, et al., 2002). With respect to metadata, the interviews are manually indexed using a thesaurus of over 30,000 index terms. A single interview typically requires 7 hours of combined effort to completely index and describe (Shoah Foundation, 2005, The Archive).

Open Video Digital Library

The Open Video Digital Library (OVDL) is a technology testbed developed by the University of North Carolina (The Open Video Project, 2005). The project contains 1800 video segments in MPEG-1, MPEG-2, MPEG-4, and QuickTime formats totaling over half a terabyte of data. The video segments are collected from a variety of sources, including Informedia and various U.S. government agencies, that provide a data set for research into video retrieval and indexing techniques. The retrieval system was built using open source PHP middleware and MySQL database platforms. OVDL metadata is encoded using a Dublin Core metadata set and exposed via Open Archives Initiative (OAI) protocols for remote search (Marchionini, 2004).

Físchlár

The Físchlár Digital Video Systems are on-demand video retrieval applications that provide access to recorded Irish television news and entertainment programming (Físchlár, 2005). Early systems behaved much like a digital video recording system (e.g.: TiVo) by capturing live television broadcasts, but newer systems are available that provide educational materials for

distance learning. The Físchlár systems use a common XML-based architecture that determines the appropriate client viewing technology then streams appropriately encoded data to the user. Additionally, metadata is stored and exchanged using the MPEG-7 metadata encoding standard (Smeaton, 2004).

DigitalWell

DigitalWell is a collaborative project between the University of Washington and ResearchChannel (DigitalWell Project, 2005). The system provides access to selected news and radio programs and a variety of other documents in a variety of formats including Windows Media, MP3, QuickTime, MPEG-2, and MPEG-4. The system was developed using Internet-2 protocol frameworks in order to explore the delivery of high-definition video over the nascent networks. Metadata is provided using extended Dublin Core and MPEG-7 profiles and may be accessed via OAI protocols (DeRoest, 2003).

Repository Software

An alternative to the development of a customized repository as described above is the implementation of one of several open source digital library and repository software packages that are currently available. These products may be suitable for use by Texas Legacy with relatively little customization, depending on the final project scope.

DSpace Repository

DSpace is a collaborative project between MIT and Hewlett-Packard, Inc. (DSpace Federation, 2005). The software is built in Java and is based upon the OAIS framework. DSpace is configured with basic Dublin Core metadata support, an OAI harvesting mechanism, a simple collections model, and a configurable institutional work flow model. Objects may be individually or batch imported, optionally using METS encoding for export. An applications programming interface (API) allows for additional customization. Given its simplicity, DSpace may require some customization to accommodate a large video collection, especially in the ingest process, metadata model, and data storage architecture.

The iSchool hosts a DSpace testbed and recently upgraded the software and hardware. It is hoped that a robust proof of concept will encourage the adoption of DSpace or a similar repository system by the U.T. Libraries and archives.

Fedora Repository

Fedora is a collaborative project of the University of Virginia Library and Cornell University begun in 1999 (Fedora Project, 2005). Fedora is built in Java and stores objects in an SQL database in XML wrappers. The software uses a Web services model for interaction with clients, users, and eventually, other repositories. Fedora supports OAI discovery and harvesting which is capable of exposing Dublin Core, Metadata Encoding and Transmission Standard (METS), and MPEG-21 metadata. Objects may be individually or batch imported, optionally using METS and MPEG-21 encoding for both ingest and export (Fedora Project, 2005, Documentation).

Fedora is currently at release version 2.0 and in phase two of development. Upcoming features include a federated (peer-to-peer) repository model, improved submission processes (inspired by

DSpace), and Resource Description Framework (RDF) indexing (Fedora Project, 2005, Publications).

Greenstone Digital Library

The Greenstone Digital Library was created by the New Zealand Digital Library Project at the University of Waikato in the late-1990s (Greenstone, 2005). The current version of the software is a PERL-based CGI application with some Java-based administration tools. The greatest strength of Greenstone is an automatic indexing and metadata capture program called MG (Managing Gigabytes). It is possible that MG could be extended to extract embedded metadata from video files, but no such support currently exists. Greenstone uses a software-like build/compile mechanism for building collections. This is fairly well-suited to static document collections, but makes it difficult to update collections. Additionally, customization of the user interface and retrieval functions are somewhat labor-intensive.

Preservation Plan

The needs analysis and research findings establish the context for a detailed discussion of potential preservation activities for existing and new Texas Legacy project materials. Meetings and email correspondence occurred in March and April that included representatives of CHAT and the iSchool, the purpose of which was to formulate pragmatic options for the future of the Texas Legacy materials. This section presents the findings of these communications. The methodology of the overall plan is discussed, followed by a description of the initial migration process. Next, criteria for migration strategy after completion of the initial pilot program are presented. Finally, the section ends with a discussion of organizational arrangements to be made concerning permanent digital repository and succession plans.

Preservation Plan Methodology

Semantic Organization

In order to establish a pragmatic preservation strategy, the basic semantic units of information to be preserved must be defined. All Texas Legacy materials are oriented on two main products: video interviews and topic segments produced from these interviews and "b-roll" footage. These semantic units and their supporting documents are recorded on various physical media, such as mini-DV tapes, and virtual media, such as electronic text transcripts and interview logs. The original video materials should be kept intact as they are converted into digital form.

From a current and future user's standpoint, however, the physical and virtual containers are not of as much concern as the semantic groupings. For example, a researcher will not seek reel #2011, but will be interested in the content of the reel: an interview with Mickey and Bob Burleson. From this perspective it is the interview that forms the basic unit of preservation, regardless of the physical storage medium or formats selected. With this understanding, however, it is important to note that any new semantic organization will maintain links to the original organization, also known as provenance, via metadata.

Constraints

In addition to the constraints identified in the needs analysis, there are a number of practical constraints that were identified during the implementation discussions. The following areas refine the parameters of the near-term preservation strategy:

- Storage Capacity and Cost: Digital storage for most documents is effectively free given their relatively small size. The same is not true of digital video storage since video data requires several orders of magnitude more storage capacity. The cost of storage is still declining as new and larger media are introduced, but, unfortunately, we have not yet reached the point where a video collection such as that of the Texas Legacy Project can be stored cheaply. It is the opinion of those involved in generating this plan that this threshold will be reached at some time in the near future; perhaps no more than several years from now. In the meantime, it is desirable to avoid excessive new hardware costs.

- Establishment and Refinement of the Migration Process: In addition to the cost-benefit analysis of storage options, there is a significant learning curve involved with equipping CHAT with the knowledge base and hardware necessary to perform the required conversions. Before equipment can be procured, however, there are issues involved in determining the optimal encoding platform and environment that relates to the selection of storage media which adds complexity to the decision process.
- Source Tape Longevity: Finally, the source tapes that contain raw interviews are steadily aging. Some of the earliest tapes are over eight years of age – very near the 10 year lower service limit of tape media. Furthermore, there is concern over the time needed to convert the source materials onto newer media. The above constraints suggest that more time is needed to establish a preservation regime, however, the advancing age of the earlier materials, combined with the one-to-one conversion time for video tapes indicates that efforts must begin soon to migrate the data to newer media.

Implementation Strategy & End State

In order to accommodate both the time and cost constraints described above, the preservation plan will proceed as a multi-staged migration process. The first stage consists of a pilot migration program with the following goals:

- Encode the highest-risk source tapes; particularly those on VHS and Hi-8.
- Establish and refine an encoding environment and develop the required personnel skills.
- Allow more time to observe digital storage costs and trends in the expectation that mass digital storage will become significantly cheaper in the near future.
- Allow time to negotiate long-term digital repository arrangements and determine a viable succession plan for Texas Legacy materials.

It is estimated that the first 90 hours of Texas Legacy video source footage is within the "at risk" category, that is, those at or near 8 years of age at this time. The process for the pilot program is described in detail below.

At the conclusion of the pilot migration program, this plan will be revisited to integrate changes to the procedures, technology, and techniques used in the pilot program, as well as organizational changes, into a revised plan to complete the conversion of Texas Legacy materials. It is expected that the majority of changes to the initial pilot plan will be related to the technologies involved rather than to the essential preservations practices that are used. Specific criteria to be evaluated at the end of the pilot program are described later.

The desired end state of the overall plan is to convert all of the Texas Legacy source materials to specified, standard formats residing on computer manipulable storage formats in multiple, geographically dispersed copies. Additionally, all digital materials will be thoroughly documented and the processes used in the conversion recorded to assist in future migration efforts. Finally, an institutional affiliation will be established to help guide and potentially provide resources for future preservation efforts.

Pilot Migration Plan

Migration Process

This section describes the information conversion and migration workflow shown in Appendix D. This workflow diagram is the starting point for conversion of existing video materials, and continues after task #3 of the current information workflow process for newly created video materials (Appendix C). It is important to note that this process diagram deliberately avoids prescribing specific storage media, formats or other technologies in order that it may accurately depict the required processes. The pilot migration will initially adhere to this strategy, then provide suggestions for refinement as they are identified.

The process begins after completion of one of three processes. New interviews are filmed and documented as described previously, or an existing interview is selected for conversion. Alternatively, short interview segments or topic segments are created using source interviews, copied in its original format, and shipped to the CHAT office in Austin, TX (#1). Source video tapes will be digitized using an encoding workstation at the CHAT offices in Austin. These will be identified using the original reel numbers and saved to digital storage. The encoded videos will then be combined with preexisting lead and trail footage (#2) which will result in a complete working file of the semantic unit (#3). The assembled semantic unit will then be assigned a unique number that will be used in subsequent file naming and identification.

The working file will then undergo several transformations into various derivative products. First, the working file will be exported to a Web-ready video format for online distribution as a use copy in a digital repository or Web site (#4). Second, the working file will be encoded and copied onto consumer-grade video media such as DVD to serve as physical reference copies (#5). These reference copies will be distributed between the CHAT offices in TX and CA, to individual narrators (for newly produced materials), and to physical repositories such as the Center for American History. Finally, the working file will be exported to an archival file format with embedded metadata (#6). A conversion log will be maintained that captures the identifying information for the tape that was converted, the date of conversion, and notes about any departures from routine procedures or other such issues that arise (#7). This log will be used to record subsequent maintenance operations in addition to inventory control. Additionally identifying and contextual information about the semantic unit will be logged and recorded as separate metadata.

Upon completion of the encoding and conversion, the archival file will be written to mass storage media and verified for integrity (#8). As soon as practicable thereafter, all digital supporting materials such as transcripts and still images will be converted to standard file formats, named in accordance with the unique number of the related semantic unit, then written to mass storage (#9). The data contained in mass storage will be periodically tested for integrity and migrated according to an appropriate schedule, to be determined by the longevity of the chosen storage media. To complete the process, a duplicate copy will be made of all materials written to mass storage (#10). The media format of this duplicate copy will ultimately be determined by the technology selected for mass storage. The duplicates will also be periodically tested for integrity and replaced as appropriate.

Asset Identification

In order to uniquely identify digital source materials, as well as link these source materials to their supporting materials, a file naming system must be devised. A new unique ID will be generated for each new semantic unit while the digitized source reels will retain the original reel number. The file naming algorithm will use this index number in a concatenation of the following:

- unique index # (4 digits)
- dash
- sequence number (for sequences of files, 1 digit)
- dash (if sequenced)
- subject keyword
- dash (if material type included)
- material type, if applicable
- dot
- windows file type extension (3 chars)

3004-burleson-transcript.rtf

represents an RTF text transcript that supports a video about "Burleson," which would have the filename:

3004-burleson.avi

All digital filenames shall not exceed 32 characters if possible to avoid cross-platform compatibility problems. Additionally, this scheme will be used for files made available on the Web and, as such, URL unsafe characters shall be avoided (see Berners-Lee, 1994, section 2.2). Additionally, characters that are reserved for Windows file names (such as * and ") shall also be avoided.

Selected Technologies

The technologies selected for preservation must of necessity be selected from those technologies that are currently available and expected to be reliable into the future. This section describes formats selected as a starting point for the pilot migration program. During the course of the pilot migration alternatives to these formats may be introduced or discovered. If newer formats prove to be well supported, open standards that are superior to existing preferred formats, the workflow process may be modified to accommodate the change, and ideally, applied retroactively.

Storage Media

After assessing the current costs of storage and storage media, digital tape remains the best value considering the volume of data that is to be stored. Unfortunately, the cost of multiple copies and the long term risks inherent in tape media make an exclusively tape-based media archive

unattractive. Alternatively, there are hard drive systems available that can hold multiple terabytes in a single, redundant array, but these are prohibitively expensive and will soon become obsolete, especially when considering the constantly declining cost of storage media.

After much deliberation, an inexpensive bulk media strategy is favored for the pilot stage. Inexpensive magnetic hard drives that will be procured on an as-needed basis during the pilot migration. Prices for large capacity, internal hard drives are constantly declining as a result of the periodic introduction of newer drives. Consequently, there tends to be a "sweet spot" in price for drives at an optimal point in the capacity versus cost curve. Currently, the optimal price resides at roughly $100 for a 200 GB hard drive ($0.50 per GB); drives of larger and smaller capacities are more expensive per GB due to supply and demand dynamics. Therefore, it is suggested the most economical hard drive capacity be determined when more storage capacity is required.

The strategy to be used for the pilot migration is as follows:

- When new encoding capacity is required, purchase a pair of internal hard drives at the optimal cost described above.
- Place these drives into a set of external Firewire drive enclosures and connect them to the encoding workstation. (software or hardware RAID may be implemented)
- Resume the encoding process until the drives are effectively full, with duplicate content on each drive.
- Carefully remove the drives from the enclosures, label the drives, return them to the original packaging, label the containers, and begin the process again.

Using this process it will be possible to inexpensively encode well over a terabyte of data, with redundancy, on a random access storage medium. All data encoded using this process can be readily accessed if needed and quickly copied to newer storage media when required.

Video File Format and Codecs

Of the video file formats studied, none simultaneously meets the requirements of long-term, archival viability and non-proprietary, open architecture. The two that come closest, MPEG-4 and AAF/MXF, either have not matured as standards, or have potential proprietary restrictions. Thus, it is preferred to retain AVI as the preferred digital video container format. AVI is an openly documented standard that is not encumbered by specific proprietary restrictions. Additionally, AVI is widely supported in audio-visual software and hardware and can adequately support the quality level of the Texas Legacy video materials (NDIIPP, 2005). An uncompressed DV codec will be used for video encoding while audio will be compressed using an optimal codec to be determined during the pilot migration.

Text File Formats

Transcripts and other documents related to the video segments are currently stored in Microsoft Office formats (Word, Excel, etc.). Although these formats are widely used, the long-term efficacy of the formats cannot be guaranteed since they are proprietary and subject to forced obsolescence due to closed specifications. Furthermore, the nature of the information stored in

these files does not require the majority of the functionality afforded by the Microsoft formats. Unless Microsoft releases its formats or a viable, open-standards based competitor to Microsoft Office becomes available, it is desirable to convert these files into an openly documented, non-proprietary format that can be interpreted by a maximal number of applications. Therefore, Microsoft Word documents will be converted into ASCII encoded Rich Text Files (RTF; Biblioscape, 2005) and Microsoft Excel spreadsheets will be exported to tab-delimited ASCII text before transferring to digital storage. Additionally, existing databased transcripts and database contents should be converted to RTF and delimited formats respectively. This process can be easily automated using the existing PHP/MySQL architecture. Using these ASCII text formats will ensure that the content of the files will remain in a common, non-binary format that is supported by any text processing program. The original Microsoft files shall also be copied to tape in the interest of maintaining the original bitstream from which the converted files originated.

Image File Formats

In addition to textual supporting items, there are a number of still images that relate to various video segments. Supporting images shall also be converted to a standard format prior to transfer to digital tape. Both the Tagged Image File Format (TIFF) and Portable Network Graphics (PNG) graphics formats meet the quality requirements for long-term storage. TIFF is an industry standard format for professional graphics that stores an uncompressed bitmap and optional metadata. The format is well documented and is stable, but is a proprietary standard owned by Adobe, Inc. As an alternative, PNG is an open, ISO standard that was meant to replace the proprietary Graphics Interchange Format (GIF) in Web browsers. Although the format was designed for Web graphics, it adequately supports high quality images with lossless compression and descriptive metadata. Since PNG is an open standard that is also supported natively by Web browsers, it is preferred that images be converted to PNG before being archived. Again, the original bitstream will also be archived to maintain provenance.

Metadata

Most of the file formats described above support at least some embedded metadata. Embedded metadata should be completed as thoroughly as possible prior to writing the data to tape. The procedures for editing embedded metadata vary depending on the file type and the software used to create the files. Values for the various metadata fields will be derived from the interview and conversion logs generated during the production and conversion processes.

Additionally, archival metadata will be generated on a per-semantic unit basis. All digital items relating to an identified semantic unit will be described with a single metadata file. These metadata files will assist the development of digital archives for CHAT materials in the future.

U.T. archives and other traditional archives typically use Encoded Archival Description (EAD), however, EAD is better suited to the functions of an archive, not so much for the maintenance and description of digital records. The Metadata Encoding and Transmission Standard (METS) serves this purpose much better (Library of Congress, 2005). METS is an XML standard that is designed for use in describing individual digital files or groups of files in a rigorous manner. METS encoding can document the description of files as well as their technical attributes and

relationships to each other. Current practices by U.T. archives will be taken into consideration during the development of a schema and tools to create METS files. These tools will be developed prior to or during the pilot program.

Implementation Schedule

Encoding Priorities

Existing interviews and segments will be converted based on media type, analog media first, then in the order of creation date, oldest first. Source media are listed in order of this priority in the first column of Table 2.

Work Effort and Duration

Given the tasks described in the conversion plan, it is estimated that one interview or produced segment video will require as much as three hours to convert and create derivative products. This estimate includes the time necessary to:

- Prepare for digitization
- Play the video into the editing environment
- Assemble, add metadata, and save the raw video file
- Create a Web encoded version
- Create DVD reference copies
- Gather and convert supporting materials
- Write all materials to digital storage, backup storage, and update conversion log

This process will take somewhat longer at the outset of the project. Automation through scripting will reduce the required time and allow for other tasks to be performed during the process. Additionally, opportunities may arise for encoding assistance from academic programs or interns studying digitization.

Given the time to encode each source tape, it is assumed that no more than one semantic unit can be encoded per week (no more than 52 per year) by CHAT staff without arranging for outside assistance. A projected migration schedule for the pilot program based on these assumptions is shown in Table 2.

*Reel Numbers (qty)**	*Source Media*	*Source Date(s)*	*Migration Window*
1020 (1)	VHS	Jun 1999	Jul 2005
1001 – 1015 (15)	Hi8	Feb – Aug 1997	Jul – Oct 2005
1016 – 1019 (4)	Hi8	Jan – Jul 1998	Nov 2005
1008 (1)	Hi8	Oct 2000	Nov 2005
2002 – 2065 (63)	Mini-DV	Jun – Oct 1999	Jan – Sep 2006
2070 – 2131 (61)	Mini-DV	Feb – Oct 2000	TBD (post-pilot)
2133 – 2170 (38)	Mini-DV	Mar – Aug 2001	TBD
2171 – 2249 (78)	Mini-DV	Apr – Oct 2002	TBD
2250 – 2295 (46)	Mini-DV	Oct 2003	TBD
2296 + (N/A)	Mini-DV	TBD	TBD

Table 2: Encoding priorities and source groupings.

**Not all sequence ranges are contiguous, hence, quantities do not match the numerical differences.*

Estimated Budget

The estimated budget for the duration of the project is shown in Table 3 below. The projection includes costs necessary to establish a digitization environment for CHAT and the costs associated with migrating 90 hours of digital video to digital storage (approximately 54 interviews accounting for 1.2 TB of data). Low and high cost estimates are provided for reference, as determined via multiple Web searches of vendor and e-commerce sites. Final costs for each item will depend on market availability and additional features.

Media costs are a high estimate based on current costs. There are several factors to consider in media pricing. First of all, not all of the required media must be purchased at one time. Second, media costs constantly decrease over time as new, denser formats become available and reduce demand for previous formats. Therefore, actual media costs are likely to be lower than those shown over the life of the pilot project.

Item	*Low Estimate*	*High Estimate*
Digitization workstation: Macintosh OS X platform with internal DVD burner	$2,500	$3,500
200+ GB Internal Hard drives, 24 total (maximum) 2400 GB main storage and 2400 GB backup storage	$2,400 (@ $0.50/GB)	$2,880 (@ $0.60/GB)
External Firewire Hard Drive Enclosures, 2 total	$90	$120
Mini-DV Cassette Player with digital video output	$1,500	$1,600
Mini-DV cleaning tapes, 2 total	$10	$30
DVD-R Media, 162 discs total	$59 (@ $0.36 ea.)	$150 (@ $0.92 ea.)
Authoring software: iMovie (free), iDVD (free), and Cleaner 6 (commercial)	$500.00	$550.00
Total:	***$7,059***	***$8,921***

Table 3: Estimated pilot migration budget.

Post-Pilot Assessment

Assessment Criteria

Upon completion of the pilot migration, this plan must be revisited and revised to incorporate possible improvements. Two major areas should be evaluated before continuing migration of assets after the pilot plan is complete: storage media and formats.

Storage media should be evaluated primarily in terms of cost versus capacity as described earlier. Additionally, any newly developed storage technologies that become available should be considered in terms of the factors described in the Literature Review, especially longevity, open standards, and market adoption. Any critical lessons learned during the course of the pilot program should also be taken into account. Finally, changes in the organizational relationships between CHAT and other institutions may present new opportunities. For example, a sufficiently willing institution may have the ability to store some or all of the Texas Legacy digital materials.

File formats and codecs should be evaluated primarily with respect to the seven factors described by the NDIIPP (NDIIPP, 2005). Any lessons learned during the pilot program should also be applied. In any case, changes in format should constitute a net improvement over formats used during the pilot program. In all cases, the original, physical video tapes and the first-generation digital files created during the conversion process will be maintained in their original form, regardless of the relative merits of newer technologies. Under no circumstances will these original records be “thrown out.”

Maintenance

Storage

All physical storage media shall be housed and maintained according to the media manufacturers suggestions, avoiding exposure to heat, humidity, pollutants, magnetism, or vibration. Source tapes shall be stored and handled as outlined in Van Bogart (1995, Section 5.2). Media shall be handled as little as possible to reduce the chance of accidental damage or other exposure that could threaten the data contained therein. Additionally, special physical storage containers that prevent media exposure should be considered. In the event that damage to media is suspected, it should be tested according to the procedure below.

Testing

One drive, chosen at random, shall be tested every three months for data integrity. Testing consists of the steps below:

- Retrieve the drive from storage and place in an enclosure.
- Perform integrity tests on media using drive testing software.
- Attempt to copy a file both to and from the drive, taking care not to alter the original data.
- View the contents of the drive and evaluate integrity of files.
- Remove temporarily transferred files and return to storage.

In the event of unrecoverable file corruption or transfer problems, the companion drive will be obtained from storage and tested in the same manner. If the companion drive exhibits similar problems, all files will be copied from one of the drives to the encoding workstation. If any files are unable to be copied successfully from the first drive, the copy will be retried using the companion drive. The original video tapes will be re-digitized in the case of any unrecoverable video files. Otherwise, if the companion drive is free of defects, the contents will be re-copied to the original drive using the contents of the companion. All events encountered and actions taken will be described in detail and recorded in the conversion log.

Repository & Succession Planning

Two remaining areas must be addressed in addition to the pilot migration. These areas are somewhat independent, but may benefit from being negotiated and resolved together. First, the current method for retrieving digital use copies from the Texas Legacy Web site should be updated. Such an update can take many forms (in order of difficulty):

- Redesign the existing Web site to make better use of the technologies available for presenting use copies.
- Implement a digital library or repository system as described in the Literature Review.
- Negotiate to use the resources of an existing digital repository.

Second, a succession plan must be devised to ensure that all Texas Legacy materials are maintained by some other agency in the event that CHAT is unable to maintain them for some reason. A suitable succession plan includes responsibility for all of the physical and digital assets, as well as the digital use copies.

Ideally, these plans will be unified such that the agency responsible for succession may immediately take some or all the responsibility for the deposit of physical use copies (as the CAH does now) and digital use copies as well as copies of source tapes. These plans are expected to take time to develop fully and may involve interim steps to reach fruition.

Succeeding institutions should have greater resources than CHAT and have existing archival resources as well as sharing similar organizational goals. Some possible institutions include:

- Academic institutional archives.
- Texas environmental, history, or humanities organizations.
- National environmental, history, or humanities organizations.

Appendix A: References

Baldi, L., Bez, R., Bechevet, B., Chappert, C., van Haaren, J., Samson,Y, et al. (2005). *Memories for future electronic systems: A road map for European research and development*. Innovative Mass Storage Technologies "White Book". Retrieved from http://www.ex.ac.uk/dsnet/white_book.htm.

Baumann, C, Larabie, C. & Atkinson, S. (2002). *Managing archive storage systems and the transition from tape to disk*. SMPTE 144th Technical Conference and Exhibition, October 23-26, 2002, Pasadena, CA. Retrieved from http://www.leitch.com/resources/applicationNotes/servers/archiveStorage.pdf.

Berners-Lee, T. (1994). *Uniform resource locators (URL)*. Internet Engineering Task Force, RFC 1738. Retrieved from http://www.ietf.org/rfc/rfc1738.txt.

Biblioscape (2005). *Rich text format (RTF) version 1.5 specification*. Retrieved from http://www.biblioscape.com/rtf15_spec.htm.

Blu-ray.com (2005). *Home page*. Retrieved from http://www.blu-ray.com/.

CCSDS (2002). *Reference model for an open archival information system (OAIS)*. CCSDS 650.0-B-1 "Blue Book". Retrieved from http://www.ccsds.org/documents/650x0b1.pdf.

Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S. & Wactlar, H. (1995). Informedia digital video library. *Communications of the ACM, 38*(4), 57-58.

Cornell University Library (2003). *Digital preservation management: Obsolescence.* Retreived from http://www.library.cornell.edu/iris/tutorial/dpm/oldmedia/index.html.

DeRoest, J. (2003). DigitalWell: Screaming media asset management. *D-Lib Magazine, 9*(4). Retrieved from http://www.dlib.org/dlib/may03/05inbrief.html#DEROEST.

Digital Preservation Testbed Project (2001). *Migration: Context and current status*. Digital Preservation Testbed White Paper. Retrieved from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf.

Digital Preservation Testbed Project (2003). *Emulation: Context and current status*. Digital Preservation Testbed White Paper. Retrieved from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf.

DigitalWell Project (2005). *UW media: Home page*. Retrieved from http://www.digitalwell.org/.

DSpace Federation (2005). *DSpace*. Retrieved from http://www.dspace.org/.

DVD Town (2005). *HDDVD.org: The different formats*. Retreived from http://www.hddvd.org/hddvd/difformatsblueray.php.

Fedora Project (2005). *Fedora*. Retrieved from http://www.fedora.info/.

Físchlár (2005). *Físchlár Digital Video Systems*. Centre for Digital Video Processing, Dublin City University. Retrieved from http://www.fischlar.dcu.ie/.

Gilliland-Swetland, A. (2000). Setting the stage. In Murtha Baca (Ed.), *Introduction to*

Metadata. Retrieved from http://www.getty.edu/research/conducting_research/standards/intrometadata/pdf/swetland.pdf.

Glossary Tech (2005). *Glossary-Tech.com*. Retrieved from http://www.glossary-tech.com/.

Greenstone (2005). *Greenstone Digital Library*. Retrieved from http://www.greenstone.org.

Gustman, S., Soergel, D., Oard, D., Byrne, W., Picheny, M., Ramabhadran, B. & Greenberg, D. (2002). Building and using cultural digital libraries: Supporting access to large digital oral history archives. *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, 18-27.

Imperial College London (2004). *Peter Török's research group - Research topics: Optical data storage*. Retrieved from http://www.imperial.ac.uk/research/photonics/pt_group/peter_torok_research_topics_ODS.htm.

Informedia (2005). *Informedia digital library*. School of Computer Science, Carnegie Mellon University. Retrieved from http://www.informedia.cs.cmu.edu/.

Jones, M. (1997). *The limits that await us*. Presented at the THIC Meeting, Falls Church, VA, April 23, 1997. Retrieved from http://www.thic.org/pdf/Apr97/mitre.mjones.pdf.

Kallender, P. (2004). Japan's Optware advances holographic disc storage. *Computerworld* (August 2004). Retrieved from http://www.computerworld.com/hardwaretopics/storage/story/0,10801,95446,00.html.

Lawrence, G., Kehoe, W., Rieger, O., Walters, W. & Kenney, A. (2000). *Risk management of digital information: A file format investigation*. Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub93/contents.html.

Library of Congress (2005). Metadata encoding and transmission standard (METS). Retreived from http://www.loc.gov/standards/mets/.

Lindner, J. (1996, February). Magnetic tape deterioration: Tidal wave at our shores. *Video Magazine*. Retrieved from http://palimpsest.stanford.edu/byauth/lindner/tidal.html.

Marchionini, G. (2004). A briefing on the evolution and status of the Open Video digital library. *International Journal on Digital Libraries, 4*(1), 36–38.

NDIIP (2005). *Digital formats for Library of Congress collections*. Retrieved from http://www.digitalpreservation.gov/formats/.

Open Video Project (2005). *The Open Video Project*. School of Information and Library Science, University of North Carolina at Chapel Hill. Retrieved from http://www.open-video.org/.

Optware Inc. (2004). *The world's first movie recording on a preformatted holographic disc*. Retrieved from http://www.optware.co.jp/english/what_040823.htm.

Shoah Foundation (2005). *Survivors of the Shoah Visual History Foundation*. Retrieved from http://www.vhf.org/.

Smeaton, A., Lee, H. & McDonald, K. (2004). Experiences of creating four video library

collections with the Físchlár system. *International Journal on Digital Libraries, 4*(1), 42-44.

Spalding, R. (2003). *Storage networks: The complete reference*. Berkeley, CA, London: McGraw-Hill Professional.

Theora.org (2005). *Theora: Free video compression from xiph.org*. Retrieved from http://www.theora.org.

Troppens, U., Erkens, R. & Müller, W. (2004). *Storage networks explained: Basics and application of fibre channel SAN, NAS, iSCSI and Infiniband*. West Sussex, UK: John Wiley & Sons, Ltd.

University of Leeds (2004). *Survey and assessment of sources of information on file formats and software documentation*. Final report of the Representation and Rendering Project. Retrieved from http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf.

US-InterPARES Project (2002). *Findings on the preservation of authentic electronic records: Final report to the National Historical Publications and Records Commission*. Retrieved from http://www.gseis.ucla.edu/us-interpares/pdf/InterPARES1FinalReport.pdf.

Van Bogart, J. (1995). *Magnetic tape storage and handling: A guide for libraries and archives*. Commission on Preservation and Access and the National Media Laboratory. Retrieved from http://www.imation.com/government/nml/pdfs/AP_NMLdoc_magtape_S_H.pdf.

Video Development Initiative (1999). *Digital video for the next millennium*. Retrieved from http://www.vide.net/resources/whitepapers/video/1.shtml.

Wactlar, H., Christel, M., Yihong Gong & Hauptmann, A. (1999). Lessons learned from building a terabyte digital video library. *Computer* (February 1999), 66-73.

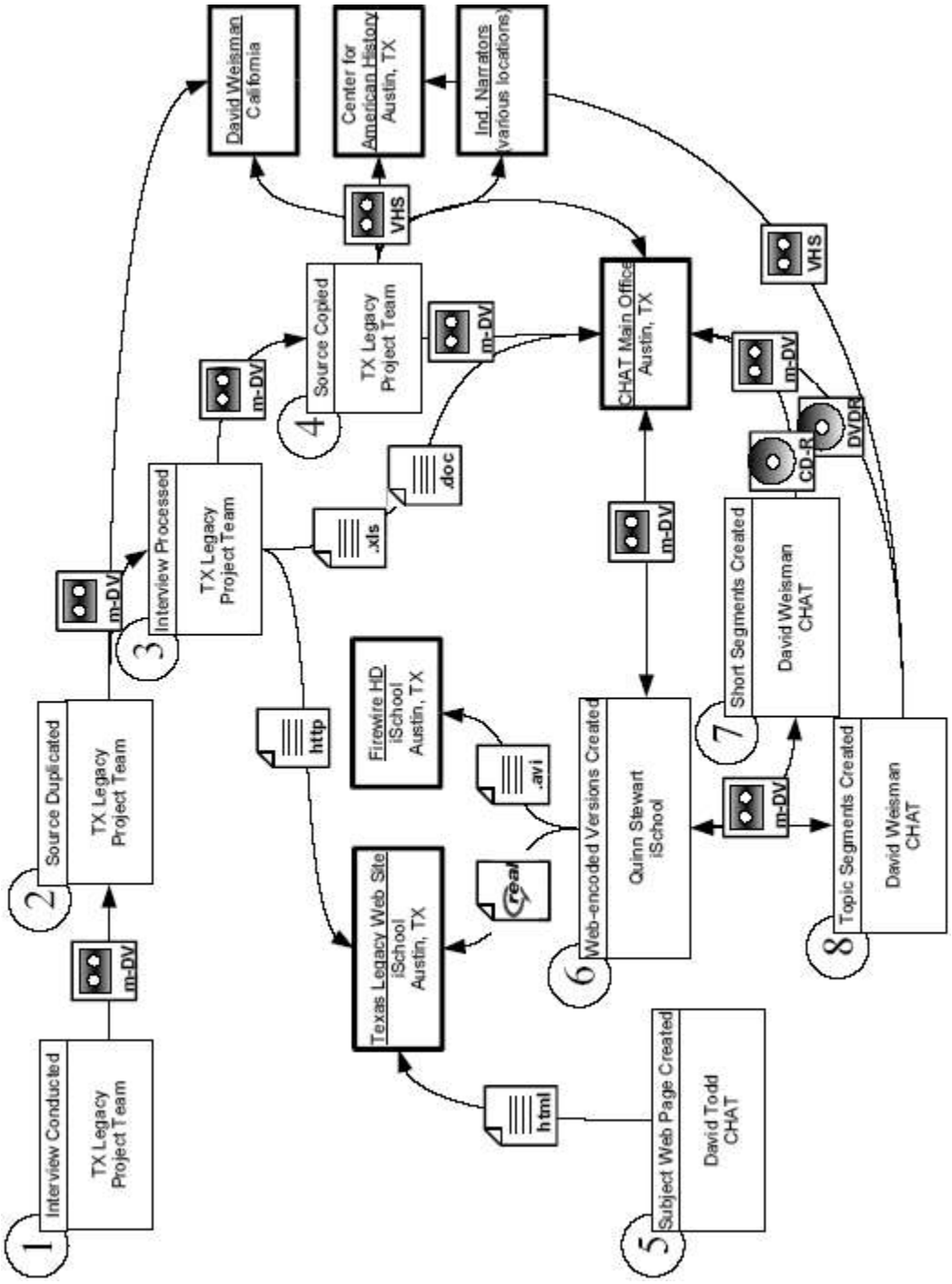
Wactlar, H. & Christel, M. (2002). Digital video archives: Managing through metadata. In *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*, National Digital Information Infrastructure and Preservation Program, Library of Congress. Retrieved from http://www.informedia.cs.cmu.edu/ documents/Wactlar-CLIR-final.pdf.

Appendix B: Inventory of CHAT Information Assets

(As of 10 January 2005)

- 40 hours of hi-8 tape recordings of interviews #1001-1020 (20 tapes)
- 40 hours of DAT audio recordings of interviews #1001-1020 (20 tapes, as backup)
- 40 hours of backup VHS recordings of interviews #1001-1020 (20 tapes, as backup)
- 54 analog audio backup tapes for interviews #1001-1020
- 301 hours of mini-DV recordings, #2001-2301, of interviews and landscape
- 301 hours of backup VHS recordings, #2001-2301, of interview and landscape
- 160 photographic color 35mm negatives, prints, and scans of narrators
- 10422-line, 15-column Excel spreadsheet of hi-8 and mini-DV interview and landscape log entries
- 624-line, 10-column Excel spreadsheet of black-and-white film landscape log entries
- 505-line, 4-variable Excel spreadsheet of environmental history events
- 155 transcripts, averaging 12,000 words or 90k each
- Web server containing:
  - “Static” HTML files and supporting graphics
  - 2 45-minute videos on citizenship (on Jim Hightower and Diane Wilson), with RealMedia, Quicktime, and full-resolution mini-DV versions
  - 2 28-minute videos on water (groundwater and surface water), with RealMedia, Quicktime, and full-resolution mini-DV versions
  - 1 28-minute video on ecotourism, with RealMedia, Quicktime, and full-resolution mini-DV versions
  - 1 15-minute video on northeast Texas issues and narrators, with RealMedia, Quicktime, and full-resolution mini-DV versions
  - 1 5-minute video on energy (featuring Smitty Smith), with RealMedia, Quicktime, and full-resolution mini-DV versions
  - 1 4-minute video on the cattle industry (featuring Terry O'Rourke), with RealMedia, Quicktime, and full-resolution mini-DV versions
  - 41 2-3 minute narrator profile videos as modem-size and broadband-size RealMedia streaming files, with RealMedia and Quicktime versions on PC, and full-resolution mini-DV versions
- 40 2-3 minute narrator profile videos, with RealMedia and Quicktime versions on CD, and full-resolution mini-DV versions being digitized for streaming
- 21 2-3 minute narrator profile videos (in edit)
- 7 ecoregion ("Texas Vista") 2-minute videos, in RealMedia and Quicktime, on CD and workstation PC
- 9 3-minute "Arts&Conservation" videos, in RealMedia and Quicktime, on CD and workstation PC
- 1 351-record Claris Filemaker database holding contact information for narrators, technical support, archival contacts, etc.
- Various paper correspondence, background research, board meeting minutes, etc.

Appendix C: Current Information Workflow Process



Appendix D: Proposed Information Workflow Process