The Universal Preservation Format
Background and Fundamentals

Thom Shepard, Dave MacCarn
thom_shepard@wgbh.org, dave_maccarn@wgbh.org
WGBH Educational Foundation
Boston, MA, USA

Sixth DELOS Workshop
Preservation of Digital Information

I. Background

> *Film, video and sound recordings are vital components of our collective memory... This vast source of information, inspiration and creativity—the most known contemporary archive of our society—is threatened. ...[We] are losing large parts of our recorded past."*[1]

According to a recent Library of Congress report, video materials in the public and private sector are estimated to exceed several hundred thousand recorded hours. The same report judges the amount of feet of news film and other film used to record television programming to total in the several millions.[2] Much of this historical material is in danger of being lost.

> *[The] preservation of television and video materials faces enormous obstacles, in particular, the vulnerability of videotape to adverse storage conditions, abusive handling, and technological obsolescence.*
>
> William T. Murphy, Coordinator
> Report on the State of American
> Television and Video Preservation
> In a letter dated January 16, 1996

With more than one hundred sixty thousand hours (160,000) of video programming, WGBH has first-hand knowledge of these vulnerabilities, which is rapidly being translated into monetary considerations. For example, while income recovered from the first airing of a production used to be 100%, that figure is now 55%, with an additional 44% of the total income generated being made from additional sales of materials from the production. In short, recovering production costs relies heavily on materials from our archives.

Technological obsolescence, in particular, has hindered the preservation of film and video materials by contributing to the enormous expense of accessing stored materials. The standard format for recording television programs has taken several forms over the past fifty years, including kinescope, 2» videotape, 1» videotape and digital tape. As the standard for recorded programming continues to evolve, the equipment used to access materials produced in earlier formats has become increasingly difficult to find and, accordingly, more expensive to use.

Archivists and industry members have addressed the problem by transferring older formats to digital tape, thus attempting to maintain the quality of the original material. However, this costly process simply puts off—rather than solves—the preservation problem. The enormous and rapid changes taking place in digital technology have resulted in a veritable explosion of formats. Fourteen different digital tape formats are available at present (D-1, D-2, D-3, D-5, D-6, Digital Betacam, Betacam SX, Ampex DCT, Consumer DV, DVCAM, DVCPRO (D-7), DVCPRO-50, Digital S and D-VHS) with several more for High Definition Television. With these format wars

heating up, many of these formats may soon become obsolete, making them unsuitable for preserving media information. In addition, digital non-linear editing systems have internal proprietary media formats.

From an archivist perspective this is a nightmare. On one side of the room you store the tapes and on the other side the tape machines and spare parts.

As with the videotape materials produced during the last fifty years, technical obsolescence may make digital formats that are common today inaccessible tomorrow. There is a significant need for a Universal Preservation Format (UPF), designed specifically for digital technologies, that can store compound content (not only the media itself but also information about it) so that it can be accessed easily both today and into the indefinite future.

This thinking is reflected in a national plan for redefining film preservation put forth by the Librarian of Congress in August of 1994. As one of his eight recommendations, he suggested that:

> *[The new electronic technologies] are already transforming film access but archives should insist that certain stringent criteria be met before new technologies are adopted as preservation media.*[3]

Paul Messier, Conservator of Photographs and Works on Paper for the Boston Art Conservation, also called for the establishment of criteria for assessing digital video as a preservation medium in a paper that he presented at Playback '96: A Round Table on Video Preservation. His suggested criteria were adapted from those suggested for still images by Basil Manns, Research Scientist at the Library of Congress in his article «The Electronic Document Image Preservation Format.» These «criteria» anticipate the technical specifications necessary for the selection and description of data to be preserved through a UPF. However, no further action has been taken within the field of archives.

Let there be no ambiguity on this key point: we are emphatically not advocating yet another acquisition format, universal or otherwise. Additionally we need to differentiate between digital archives, which are concerned with the timeless storage of digital materials, and digital libraries, which is primarily concerned with timely issues of access. Our proposed solution is to establish a format expressly for the permanent archival storage of digital materials.

Background

The concept of a universal preservation format was introduced at the Society of Motion Picture and Television Engineers Conference, October 1996. Sponsored by the WGBH Educational Foundation and funded in part by a grant (97-029) from the National Historical Publications and Records Commission of the National Archives, the Universal Preservation Format initiative may be summed up in the following passage:

> *a platform-independent Universal Preservation Format, designed specifically for digital technologies, that will ensure the accessibility of a wide area of data types -- especially video formats -- into the indefinite future. WGBH will work with both technology manufacturers and archivists to determine a UPF that meets the needs of both non-commercial and commercial interests. At the end of the two-year grant period, WGBH will submit a Recommended Practice to the Society of Motion Picture and Television Engineers (SMPTE), a standards-creating organization, and the Association of Moving Image Archivists (AMIA).*
>
> (Project Summary, "Statement of Purpose," p. 1)

Mission of UPF

Our mission, as outlined in the original UPF grant proposal, includes the following key tasks:

- to analyze the problem of preserving the video digital data contained in electronic records
- to raise awareness of electronic records preservation
- to build support for an effective solution
- to develop a Recommended Practice for a UPF and encourage its adoption as a cross-industry standard
- to present the concept of a universal preservation format to both archivists and technology manufacturers at professional conferences and working groups, and through articles in professional journals.

Working with representatives from standards organizations, hardware and software companies, museums, academic institutions, archives and libraries, this project will submit the final draft of the Recommended Practice to the Society of Motion Picture and Engineers (SMPTE). It will suggest guidelines for engineers when designing computer applications that involve or interact with digital storage. We expect to make the process of preserving and accessing electronic records (both original and migrated) more efficient, cost-effective, and simpler.

Reaching out

From the very beginning of this project, we have actively solicited the participation of several organizations. Our ever-growing database of contacts includes members from the Association of Moving Image Archivists, the Society of American Archivists, the Music Library Associations, Boston Art Conservation, and Conservation Online, as well as individuals who may not be members of these groups but who nonetheless are actively involved in preservation issues. In addition to the UPF listserv[4], which currently has close to sixty members, postings go to the AMIA listserv, the Archivist & Archives listserv, and the Electronic Records listserv. Mailings are also send to archivists who do not have email addresses. A task item planned is to set up teleconferences between engineers and archivists.

On September 22, 1997, the Society of Motion Picture and Television Engineers assigned an official Study Group (ST13.14). At this first meeting, the following objectives and tasks were established:

Statement of Objectives

- The study group will document requirements of data formats for the preservation of electronically generated media and related information.
- Extensive input from the archival community will be gathered through the use of surveys and meetings.

Specific Tasks and/or Documents

- Some areas of study: Containers, Objects, Labels, Metadata, Composition, Rosetta stone for future coding and translation, and the coordination of other SMPTE activities that may be useful.
- Explore the possibility of a universal format and guide to the storage of collections.
- Based on the requirements gathered, the study group will investigate available technology and explore configurations in order to provide a basis for a working groups' recommended practice or standards.

(SMPTE Engineering Committee Work Assignment/Work Statement)

User Survey

What are these needs and concerns? UPF established a user survey.
Though many archivists said that they realized they would have to "migrate" at some point, most could not justify the costs of either migrating to digital or investing in new digital equipment that will only become obsolete in a few years. Running throughout these commentaries was the frustration that archivists had no control over new technologies. And while digital has qualities that are enormously appealing to archivists -- searchability, mobility, longevity -- computer technologies seem disposable, like snakes shedding their skins.

Some archivists reported that they are feeling pressure from administrators to go digital for all the wrong reasons: consolidating their collections, for example.

For those already involved in some form of digital conversion, the strategy has generally been to convert from analog to digital in an ad hoc manner. No one has developed strategies for replacing their analog collections with digital formats.

The published results of the survey are on the web site[5]. In addition there are posted follow-up questions that invite comment.

II. Technical Specifics of the UPF

The first step is to separate the data format from the storage format. We can look at this separation through the use of essence, metadata, wrappers and identifiers.

Essence

Media can be thought of as a data object. These objects can be data types, such as video or music. The term often used to label these data types is «essence.» Many software applications are capable of interacting with a range of these essence files through the use of interchange formats. For example, in the transfer of word processing documents across applications and even operating systems, the Rich Text Format (RTF) is often used. For video, a standard for interchange is SMPTE 259M.

Metadata

A data object can also take the form of information about the data types. In terms of function, this information, for which the term "metadata" has been coined, may be divided into four basic categories: format, description, association, and composition.

Wrapper

The wrapper -- or container -- is a file format for storing "essence" along with the information that describes it. The wrapper is a file format that has a framework structure. Anyone familiar with the Dublin Core metadata initiative, specifically the Warwick Framework Architecture, may have some understanding of frameworks as a method for managing data. Warwick describes a metadata structure in which material describing certain objects may either be embedded in the source or be referenced to files or storage areas external to the source. This information might include domain specific descriptions, terms and conditions for document use, pointers to all manifestations of document, archival responsibility, and even structural data.

Unique Identifier

Identifying digital objects as unique entities is essential to establishing archival integrity, especially when it is so easy to misplace, corrupt or delete digital information. The UPF is looking at initiatives dealing with unique identifiers and expects to include such a system or systems in our Recommended Practice. Basically, each object carries an ID that is unique within its container. As this object undergoes changes, often called "versioning," each new generation is assigned its own identifier, which always references its parent.

Self Describing Format

The foundations of essence, metadata, wrappers and identifiers can create a «self describing» format. The UPF uses a "digital Rosetta stone" to get at the range of data types held in a digital storage bank. The digital Rosetta stone serves as a key, defining the data types and encapsulating algorithms for deciphering the file. The use of platform-independent algorithms is used to decode file types. The Rosetta stone might also serve as local registry for unique identifiers.

Use of existing technology

These self-describing technologies are already available. Along with the surge of digital formats are technologies that are designed to handle digital media of all types. Apple Computer's «Bento Specification»[6] , Avid Technology's «Open Media Framework Interchange Specification»[7] and the Society of Motion Picture and Television Engineers/European Broadcast Union's «Harmonized Standards for the Exchange of Television Program Material as Bit Streams.[8]» are media technologies that approach the UPF concept .

Bento

Apple Computer's «Bento Specification» is the underlying technology of Apple Computer's OpenDoc Standard Interchange Format.[9] Bento is a specification for storage and interchange of compound content. Bento defines a standard format for storing multiple different types of objects and an API to access these objects. An object

container is just some form of data storage (such as a file.) This storage is used to hold one or more objects (values) and information about the objects (metadata.) Bento containers are defined by a set of rules for storing multiple objects, so that software that understands the rules can find the objects, figure out what kind of objects they are, and use them correctly.

Bento objects can be simple or complex, small (a few bytes) or large (up to 2^{64} bytes, approximately 2^{27} hours of D-1 video.) Bento is designed to be platform and content neutral. so that it provides a convenient container for transporting any type of compound content between multiple platforms. The Bento code currently runs on Macintosh, MS DOS, Microsoft Windows, OS/2 and several varieties of Unix.[10]

OMF

Avid Technology's «Open Media Framework» (OMF) Interchange, a standard format for the interchange of digital media data among different platforms, adopted the use of Bento containers. Additionally, the OMF format encapsulates all the information required to transport a variety of digital media such as audio, video, graphics, and still images, as well as the rules for combining and presenting the media. The format includes rules for identifying the original sources of the digital media data, and it can encapsulate both compressed and uncompressed digital media data.

OMF Interchange provides for a variety of existing digital media types and the ability to easily support new types in the future. A single OMF Interchange file can encapsulate all the information required to create, edit, and play digital media presentations.

While OMF Interchange is designed primarily for data interchange, it is structured to facilitate playback directly from an interchanged file when being used on platforms with characteristics and hardware similar to those of the source platform, without the need for expensive translation or duplication of the sample data. OMF Interchange provides for the development and integration of new media and composition types.[11]

SMPTE/EBU

Based on Bento and OMF with the addition of unique identifiers.

Bento, OMF and SMPTE/EBU as a Preservation format

Preservation requires the handling of many different recording formats—such as 2» videotape, 1» videotape, D-1, D-2, D-3, and others—which can be thought of as having data types (the way the video is encoded, e.g. 4:2:2, $4f_{SC}$.) Although Bento allows for any data type, the OMF Interchange only defines a <u>minimum</u> number of data types (e.g. TIFF, RGBA and AIFF). By adding <u>additional</u> standard data types to the these technologies would result in a storage container format that will be able to encompass all present recording forms and allow for all future forms. In moving from the raw recording format (e.g. videotape) to a data tape (or other media) format that incorporates the UPF, the number of formats that archivists need to preserve will be substantially reduced. The UPF breaks the bond between the recording format and the machine through which the format is accessed.

Hybrid Technology

Although these technologies bring us a self-describing format there is still the hurtle of reading them with out the native machine that recorded then.

«Reading and understanding information in digital form requires equipment and software, which is changing constantly and may not be available within a decade of its introduction. ... We cannot save the machines if there are no spare parts available, and we cannot save the software if no one is left who knows how to use it.»

> «Preserving Digital Information» - 1996 Report of the Task Force on Archiving of Digital Information, Commission on Preservation and Access and The Research Libraries Group.

The answer is in our old friend analog. We still use microfiche. We can apply this to digital storage in the form of a hybrid solution. An out of the world example in the 1977 Voyager Interstellar Outreach Program[12]. The Voyager spacecraft included a gold plated copper disk with recording from the planet earth. The playing

instructions are in a symbolic language. By taking this idea to the UPF one can create the ultimate self-describing format. The analog portion would contain information about the contents along of the disk with the instructions for the creation of a machine to read the disk (blue prints). There are currently technologies to accomplish this from Norsam Technologies[13]. Norsam focuses on the need to greatly increase the storage densities of micro fiche as well as digital recording systems. Using these two technologies the UPF becomes a preservation format that is whole, unto itself.

Media Compiler

A media compiler would perform the actual moving of data. It would remove the baggage of the acquisition format as it imported into the archive. It would optionally export whatever metadata you needed from the archive. Specifically, you could pre-select which set of relationships or media formats you wish to transport for a given need, such as Internet access. And because the relationships among your data objects would be built-in, you could very easily "package" information. For example, you could extract certain media objects, along their associative text files, based on a scholar's search patterns. These materials could then be burned into a CD-ROM or transferred onto some other portable storage.

Summary

• Working with standards organizations, hardware and software companies, museums, academic institutions, archives and libraries.
• Self-Describing Format, Immune to Technological Obsolescence.

A worthy standard for long-term digital storage will carry forth the traditional practices of analog collections. Specifically, a recommended practice must enable provenance and original order. Its framework must be robust, allowing for certain types of metadata to be embedded with the media, while others to be referenced externally. By concentrating on elemental concepts of how data and information about that data might be stored through time, the Universal Preservation Format initiative is attempting to construct a bridge between engineers and information scientists, between those who make and market technical specifications and those who must learn to use the tools of technology to preserve the rapidly decaying fruits of our cultural heritage.

References

[1] From Fading Away: Strategic Options to Ensure the Protection of and Access to Our Audio-visual Memory. Task Force on the Preservation and Enhanced Use of Canada's Audio-Visual Heritage, National Archives of Canada, June 1995, pi.

[2] Library of Congress, «Redefining Film Preservation: A National Plan» August 1994.

[3] Redefining film preservation: a national plan; recommendations of the Librarian of Congress in consultation with the National Film Preservation Board [coordinated by Annette Melville and Scott Simmon]. August 1994. Recommendation 3.7.

[4] UPF Listserv: UPF@info.wgbh.org: «Subscribe UPF your name» to listserv@info.wgbh.org

[5] UPF [http://info.wgbh.org/upf]

[6] Open Doc underlying technologies, now maintained by the Sunrise project at the Advanced Computing Laboratory at Los Alamos National Laboratory. [http://www.acl.lanl.gov/sunrise/DistComp/OpenDoc/overview.html]

[7] Open Media Framework Interchange Specification, Copyright © 1995 Avid Technology, Inc. [http://www.avid.com/omf/]

[8] SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams, Copyright (c) 1997 European Broadcasting Union and the Society of Motion Picture and Television Engineers, Inc. [http://www.smpte.org/engr/tfhs_out.pdf]

[9] Open Doc underlying technologies [see 6]

[10] Bento Specification. [see 6]

[11] Open Media Framework Interchange Specification.[see 7]

[12] Voyager Interstellar Outreach Program [http://vraptor.jpl.nasa.gov/voyager/record.html]

[13] Norsam Technologies, Inc. [http://www.norsam.com/]