

# The Universal Preservation Format

## Part 1: User Requirements

Thom Shepard  
UPF Project Archivist, WGBH  
thom\_shepard@wgbh.org

Dave MacCarn  
Chief Technologist, WGBH  
dave\_maccarn@wgbh.org

URL: <http://info.wgbh.org/upf/>

WGBH Educational Foundation

Draft Document  
revised March 14, 1999

# Universal Preservation Format

## Part 1: User Requirements

Thom Shepard  
thom\_shepard@wgbh.org  
Dave MacCarn  
dave\_maccarn@wgbh.org  
WGBH Educational Foundation  
revised March 14, 1999

1. Purpose of this Document
2. Definition
3. Scope
4. Background
5. Glossaries
6. Assessing User Requirements
  - 6.1 Survey of the Literature
  - 6.2 User Survey
  - 6.3 SMPTE meetings
  - 6.4 Conferences
7. Summary
8. Conclusions

### 1. Purpose of this Document

This report attempts to provide a context for the **Universal Preservation Format** by documenting how the UPF initiative builds upon other standards and technologies, how it serves the specific needs of the WGBH Educational Foundation, and how it fulfills broader archival requirements expressed by other institutions and organizations. In a sense, this document functions as metadata to the essence of the companion document, **Technical Requirements of the UPF**.

### 2. Definition

The **Universal Preservation Format** is a data file mechanism that utilizes a container or wrapper structure. Its framework incorporates metadata that identifies its contents within a registry of standard data types and serves as the source code for mapping or translating binary composition into accessible or useable forms. The UPF is designed to be independent of the computer applications used to create content, independent of the operating system from which these applications originated and independent of the physical media upon which that content is stored. The UPF is characterized as “self-described” because it includes, within its metadata, all the technical specifications required to build and rebuild appropriate media browsers to access contained materials throughout time. Objects within the UPF are branded with a unique identifier that travels with that object throughout time. Any modification made to the content of the object must be reflected in its identifier.

### 3. Scope

While archivists working with analog materials have traditionally separated issues of preservation and access, the digital age has blurred these lines to such an extent that many prominent writers in the field (Hedstrom, Waters, et al.) use the terms “digital libraries” and “digital archives” almost interchangeably. The **Digital Library Federation (DLF)**, an organization founded in 1995 for the purpose of “creating, maintaining, expanding and preserving a distributed collection of digital materials accessible to scholars, students, and a wider public,” has drafted a definition of digital libraries that seems to encompass the functions of both preservation and access:

Digital libraries are organizations that provide the resources, including the specialized

staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

(Donald J. Waters, "What Are Digital Libraries," CLIR issues, No. 4, July/August 1998.  
<http://www.clir.org/pubs/issues/issues04.html>)

In his analysis of this working definition, Donald J. Waters, director of the **DLF**, acknowledges that preserving the integrity of digital information and ensuring the persistence of digital works are ideals that member organizations achieve on a variety of levels. Nevertheless, he includes these functions as "central to the concept of a digital library."

**The Society of Motion Picture and Television Engineers/European Broadcasting Union (SMPTE/EBU) Task Force for Harmonized Standards for Exchange of Program Material**, which recently fulfilled its stated mission to "implement new technologies [...] to [establish] standards that will support the vision of future systems," acknowledged in its final report that the metadata required for a digital archive is inherently different from the requirements of acquisition file formats.

Ideally, an archive system needs a superset of the acquisition Template plus production history. Sufficient Metadata needs to exist to allow fast and efficient identification of the Content. Additional Metadata may be included to describe detailed characteristics of the Content to allow precise searching of the Content. Extensibility is once again a high requirement in order to allow inclusion of customer-specific data such as links to computer databases or billing systems. Applications may use hierarchical storage to contain such Metadata.

(SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams, Copyright (c) 1998 European Broadcasting Union and the Society of Motion Picture and Television Engineers, Inc. .  
[http://www.smpte.org/engr/tfhs\\_out.pdf](http://www.smpte.org/engr/tfhs_out.pdf))

The **UPF** assumes that a fundamental difference exists between acquisition file formats and a format to serve archival requirements. This difference is not simply between access and preservation, but between malleability and permanence, between clay and stone. In its report on preserving digital information, a task force sponsored by the **Commission on Preservation and the Research Library Group** defined digital archives as:

repositories of digital information that are collectively responsible for the long-term accessibility of the nation's social, economic, cultural and intellectual heritage instantiated in digital form.

(Task Force on Archiving of Digital Information, "Preserving Digital Information: Executive Summary," May 1, 1996, p.8. URL: <http://www.rlg.org/ArchTF/>)

While other initiatives work toward making digital materials publicly accessible today, we are investigating technologies that will save them for many tomorrows. Our success depends, in part, upon instilling a new awareness within the computer industry: that future digital storage products must be designed to fulfill the documented requirements of users. Despite some promising storage technologies on the horizon, the computer industry has yet to market affordable digital storage vehicles which will last as long as their analog counterparts, nor have they been convinced that is it in their commercial interest to do so. By concentrating on elemental concepts of how data and information about that data might be stored *through time*, the UPF initiative helps to establish working relationships between those who make and market technical specifications and those who must learn to use the tools of technology to preserve the potentially decaying fruits of our cultural heritage.

#### 4. Background

In 1977, two Voyager spacecrafts left Earth on a mission to explore and send back information about our solar system and beyond. Attached to each vehicle was a gold-plated phonograph record which contained 115 images, a collection of Earth's natural sounds, greetings in 55 languages, and musical samples from the collected works of Bach, Beethoven and Chuck Berry. Each record included a stylus, and inscribed on its protective aluminum jacket were visual instructions on how to play it. The disks were intended by astronomer Carl Sagan and his team of managers and engineers as a kind of packaged time capsule for any aliens or distant cousins who, several centuries from now, might be rocketing along.

Voyager's Interstellar Record is an example of a "self-described" storage mechanism. As an analog product, it has a significant advantage over current digital storage in that an "analogy" exists between how its information was stored and fundamental principles of physics, a relationship that does not exist in products consisting of zeroes and ones. For this binary information to be readable, an intermediary interpreter is required. The UPF's notion of a "self-describing" digital file format provides interpreters with all the technical information required to retrieve its contained materials. In effect, these stored algorithms constitute a standardized blueprint for reconstructing both the data types and the physical mechanism itself, upon which the data are recorded.

Awareness of the need for a universal preservation standard grew out of meetings between leaders of two departments within the **WGBH Educational Foundation**: Chief Technologist Dave MacCarn and the Director of the Media Archives and Preservation Center, Mary Ide. Concerns were expressed not only for moving image material, but also for the many other data types that the broadcast facility generates: audio, text, database files, captioning, descriptive video information and the whole gamut of original digital content produced for the World Wide Web. Though it was clear that a storage standard could have an enormous impact on the broadcasting industry, it remained to be seen whether UPF's ability to store all data types would appeal to other archival and library communities.

As a major public broadcasting station and content producer, WGBH has developed one of the more important media archives in the industry. Our unique collection not only has obvious production and historical value, but serves as a continuing source of revenue. Unfortunately, these very fragile possessions come in a variety of formats, many requiring antiquated machines to access their contents. Despite new media's promise of easy access and portability, digital technologies are compounding the problem of long-term storage by flooding the marketplace with new file formats and proprietary storage devices.

Two years ago, the WGBH Educational Foundation was awarded a grant (97-029) from the **National Historical Publications and Records Commission of the National Archives** to produce a Recommended Practice. At that time, Thom Shepard was hired as project archivist. Since then, we have presented UPF to a variety of engineering, computer, and archival groups. We have established an impressive Review Board, a formidable Web site (<http://info.wgbh.org/upf>) and a rapidly growing listserv. We have also collected a large volume of material from those working in the trenches of digital preservation. Among these documents are our user survey and minutes from UPF Study Group sessions within the **Society of Motion Picture and Television Engineers**. These quarterly meetings, described elsewhere in this document, bring together engineers and archivists to exchange ideas, voice concerns and help untangle the semantics that have hindered effective dialogue in the past.

#### 5. Glossaries

The need for a **cross-domain glossary** was recognized in **UPF Study Group** meetings between engineers and archivists. While many archivists could not define technical terms such as "data stream" or "file format," engineers did not understand basic tenets and

terminology of cataloguing and collection description as practiced by professional archivists and librarians. It became evident that the underlying concepts of many terms could be mapped for better communication. Several terms were seen to describe the same concept, and other terms had different meanings for different groups. What was needed, the group agreed, was a glossary that mapped the languages of the library to the languages of the computer. Though several specialized glossaries and dictionaries exist both in print and online, little effort has gone into correlating concepts across domains. We have taken the first rudimentary steps in addressing this need through our cross-domain online dictionary, available through our web site (<http://info.wgbh.org/upf/glossary.html>). Through hundreds of hypertext links, we have attempted to identify relationships among key concepts.

In keeping with the spirit of the cross-domain glossary, we will examine how the words that make up our name can have many different meanings for different groups of people.

The **Universal Preservation Format** is composed of three highly charged concepts. In a way, our name “self-describes” one of UPF’s key characteristics: it serves as a container for individual concepts stored within it.

Our use of the term **universal** carries two meanings. First, it relates to a specific philosophical approach. Problems of long-term digital storage cannot be solved or even addressed intelligently without initiating a universal collaborative effort. The responsibility for preserving our cultural heritage supercedes that of any single special interest group.

The second level of “**universal**” applies to the technological component of the UPF. We propose the creation of a system for universal storage that will serve as a safe haven for electronic media created in the past, present and future: for current digital materials, for migrated analog materials and for hybrid materials that may be developed in the future. And though we initially focused on technologies and standards recognized by the broadcasting industry, we speak of the UPF as possessing a universal application, meaning the same UPF system will work with equal success on text, image, and sound. The UPF also addresses the concern that basic requirements for any digital systems vary among disciplines and types of repositories.

When we first presented our ideas for a universal preservation format to various archival and library communities, we quickly learned that the meaning of the very term “**preservation**” differed depending upon which group we were addressing.

Paul Conway, head of the Preservation Department at Yale University, writes:

At one time, advocates for the protection of cultural artifacts, including books, primary source documents, and museum objects, used the terms “conservation” and “preservation” interchangeably. Today, preservation is an umbrella term for the many policies and options for action, including conservation treatments. Preservation is the acquisition, organization, and distribution of resources to prevent further deterioration or renew the usability of selected groups of materials.

(Paul Conway, “Preservation in the Digital World,” January, 1998. URL: [www.clir.org/cpa/reports/conway2/](http://www.clir.org/cpa/reports/conway2/))

Arguably, the so-called digital revolution has corrupted the terminology of the archivist by coining new meanings from old lexicons. The computer industry in general is notorious for stealing terms from the “analog” world to label, describe and explain its products. The very term “archive” carries an entirely different meaning in the world of personal computers than in the world of the traditional archivist. When we find our hard drives filling up, we often “archive” them into compressed “zip” or “sit” files.

For some media archivists, “digital” and “preservation” do not belong in the same sentence. Digital information is too easily damaged, too easily manipulated. The highly controversial colorizing of classic black-and-white movies comes immediately to mind. Entire movies can be digitally constructed from the “bits” and pieces of classic films. One need only watch Forest Gump stumbling through historical footage or the recent television commercial that displays Fred Astaire dancing with a vacuum cleaner to witness the power and threat of digital technology. Simon Pockley writes:

In spite of Marshall McLuhan's book-bound insights ('the medium is the message') **the process of digitization now allows us to separate content from the medium which carries it.** This loosening of the bonds imposed by medium, allows data in almost any form (text, sound, image) to be reused and recombined with a facility which we have yet to learn how to exploit.

(Simon Pockley, “Killing the Duck to Keep the Quack,” updated January 19, 1998  
<http://www.cinemedia.net/FOD/FOD0055.html>)

What may be lost on some analog archivists is that “digital” has its own innate integrity. Technically, a digital object is a primary source the moment it is first generated, whether it is authored or composed through a computer or whether it functions as a **digital surrogate**; that is, it has been scanned or migrated through some other method from analog. It is a given that a data file consisting of a facsimile of *A Tale of Two Cities* can never fully equate to a first edition of the original book. We must understand, however, that the data file itself can be considered a digital primary source and that any versions made from it should be referenced in metadata.

### The Universal Preservation Format as a file format.

What do we mean by the term “file format?” In the early days of home and office computers, file formats were associated with a single data type: a text file, a picture file or a sound file. These data types can be broken down further into software specific formats; for example, a text file can be ASCII, Rich Text Format or a proprietary word processing format, such as Word, WordPerfect or AmiPro. A complete definition of file format must include the concept of a **specification**, itself defined as a file's organizational requirements. A file format, then, may be defined as the

...specification for holding computer data that dictates what information is present in the file and how it is organized within it.

(Northern Micrographics: Imaging Dictionary: <http://www.normicro.com/glossary.htm>)

As stated above, file formats stemming from the early years of computing tended toward a single data type. The **UPF** is based on the model of a **compound document** which is a file format that contains more than one data type.

Related to compound document are the concepts, **essence** and **metadata**. **Essence** is data that represents pictures, sound or text. In the broadcast industry, **essence** is also described metaphorically as a “stream.” In a digital environment, this **essence** “flows” as binary code to some processing pool. **Metadata** is defined simply as “data about data.” A thorough definition of **metadata** would incorporate its functionality as described in detail by the “Pittsburgh Project.” For our purposes, **metadata** is defined as having four functional characteristics: **format**, **description**, **association** and **composition**. Metadata informs us about the **format** or binary code of data types. It also serves as **description** or cataloguing information. In addition, it may relate information about how one component **associates** or links to another. Finally, metadata can consist of **compositional** information, needed to combine data components into a structured sequence, such as a strip of video or staff of music.

**Essence and metadata**, as described above, may be bonded together within the same file format. The generic term is the **container**, but definitions used to describe this relationship vary from one initiative to another. For example, the **Committee on the Preservation of Electronic Documents**, a Task Force affiliated with the **Public Record Office & British Standards Institute**, calls its **bundles** “self-supporting” collections of electronic objects and their associated software.

Bundling is an electronic transit envelope that contains not only the electronic document files of whatever type, it contains a viewer and other navigation software to be completely self-contained. This envelope or Bundle of documents takes with it the correct version of viewer or browser that can display the documents in it, and to navigate around those documents.

(Public Record Office & British Standards Institute (UK), “A Mechanism for the Perpetual Preservation of Electronic Records of Value,” IDT/1/4 (A Working Group transferring to a Committee Status) TECHNICAL REPORT (Version 0.6))

The **Reference Model for an Open Archival Information System (OAIS)**, which is maintained by the **Council of the Consultative Committee for Space Data Systems (CCSDS)**, describes a similar framework for the archiving of digital information. OAIS introduces the term, “**Archival Information Package (AIP)**,” defined as:

An information packaging concept that requires the presence of **Content Information** and all the associated **Preservation Description Information** that is needed to preserve the **Content Information** over the long term. It has associated **Packaging Information**.

(Consultative Committee for Space Data Systems, “Reference Model for an Open Archival Information System (OAIS),” White Book (CCSDS 650.0-W-4.0), September, 1998. Available at [http://ssdoo.gsfc.nasa.gov/nost/isoas/ref\\_model.html](http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html))

This “**package**” consists of data required to “bind and identify” the **Content Information** (the “primary target of preservation”) to the **Preservation Description Information** (information required to make long-term sense of the content). The OAIS model corresponds to the notion of **Wrappers** as prescribed by the **Society of Motion Picture and Television Engineers/European Broadcasting Union (SMPTE/EBU) Task Force for Harmonized Standards for Exchange of Program Material**. A Wrapper gathers together program material and related information. Though the term “program material” may suggest broadcast content, wrappers in actuality apply to a range of materials, including stationary data types, such as a virtual strip of still images, frames containing scanned documents, and even a series of electronic records. Information within a wrapper may include pointers to materials stored outside the wrapper, a notion that corresponds to the **Warwick Framework**, which is described below.

(SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams, Copyright (c) 1998 European Broadcasting Union and the Society of Motion Picture and Television Engineers, Inc., [http://www.smpte.org/engr/tfhs\\_out.pdf](http://www.smpte.org/engr/tfhs_out.pdf))

## Identifiers

The idea of a **unique identifier** for electronic media stems from the domain of relational databases, but there is a key difference between records in a relational database and objects stored in UPF. If you edit a record’s content in a relational database, you do not necessarily change the unique identification of that record. But when speaking of archival content, we should assume that the record is locked. Changes are made to a copy or clone of that record, which takes on a new identifier that associates itself or references the original. To state this another way, when content is duplicated, it keeps the same unique identifier; however, if the copy is changed in any way, the changed file is considered a subset of the original, and a new unique identifier must be assigned. This concept is called **versioning**.

The **Bento Specification**, using the term **globally unique names** to mean unique

identifier, provides a naming mechanism “that can be used by large numbers of developers without registration,” and still provide reliably unique names. Bento objects also utilize a unique identifier called a **persistent ID**, which is “unique within the scope of its container.” Though it is not a Bento requirement, objects may have “additional IDs and/or names that are unique in larger scopes.”

The **(SMPTE/EBU) Task Force for Harmonized Standards for Exchange of Program Material** calls for the establishment of the **Unique Material Identifier (UMID)** as a component of its Wrapper. These identifiers would track content as it passed through a production system. Materials originating from the same source would be uniquely identified, whether they were source materials or “intermediate content elements.” **UMIDs** would also contain information “to trace copyright information and ownership of both finished and programs.”

(“SMPTE/EBU Task Force for Harmonized Standards for the Exchange of Program Material as Bit Streams,” Copyright (c) 1998 European Broadcasting Union and the Society of Motion Picture and Television Engineers, Inc., [http://www.smpte.org/engr/tfhs\\_out.pdf](http://www.smpte.org/engr/tfhs_out.pdf), p.75.)

Another concept for uniquely identifying material is a **digital signature**, which was proposed by the Department of Defense in the document, “Automated Interchange of Technical Information.” This digital signature is “a digital data file that authenticates the identity of the approving authority or sending agent by the use of a computationally unique string of numbers, and that enables the detection of unauthorized modifications to the contents of a signed data file.” (MIL-STD-1840C, “Automated Interchange of Technical Information,” <http://www-cals.itsi.disa.mil/core/formal/1840.html>) An example of a digital signature for image files is the **digital watermark**, a method of encoding images to create a hidden and traceable pattern which indicates proof of ownership but is invisible to the naked eye.

Of the many standard initiatives concerned with identifying digital documents, perhaps the best known is the **Digital Object Identifier**. A DOI is a permanent Web document identifier. If the internet address of a document changes, users are automatically redirected to its new location. The DOI system was conceived by the **Association of American Publishers**, in partnership with the **Corporation for National Research Initiatives**, and is now administered by the **International DOI Foundation**. Though the numbering system proposed by the UPF is local, similar principles are at work.

Among other initiatives for unique identification are ISO's **International Audiovisual Number (ISAN)**, Sony's **Unique Material Identifier**, Microsoft's **Advanced Streaming Format** and US Government's proposed **Digital Signature Standard (DSS)**.

Although these numbering systems may be inadequate for identifying the stored contents of a universal preservation format, much can be learned by studying their respective structure.

Lessons could also be gleaned from systems that have worked in the past. The most persevering classification scheme is the **Dewey Decimal System**. Dewey is a hierarchical classification scheme. Numbers represent concepts. The longer the number, the more specific the information. For example, to catalog information about “batting,” a pyramid from general to specific could be constructed:

700	The arts. Fine and decorative arts
790	Recreational and performing arts
796	Athletic and outdoor sports and games
796.3	Ball games
796.35	Ball driven by club, mallet, bat
796.357	Baseball
796.3572	Strategy and tactics
796.35726	Batting

Dewey has already been applied to digital objects on the World Wide Web. The “Scorpion Project,” sponsored by the **Online Computer Library Center (OCLC)**, is building a tool to automate the indexing and cataloguing of Web materials using the principles of the **Dewey Decimal System**.

(Keith Shafer, “A Brief Introduction to Scorpion,” <http://orc.rsch.oclc.org:6109/binintro.html>)

If we substitute subject content with other kinds of metadata, we might see how a Dewey-like system could be constructed to describe complex relationships, including versioning and levels or degrees of dependency. For example, segments might refer to specific hardware, operating systems, software data types and source programs.

**Traceability** is a concept related to unique identifiers. It means that versions of objects should be traceable to the original through some quality within the unique identifier. In other words, the unique identifier would be an **intelligent number** that would establish a relationship with other numbers, much in the same way that **Dewey Decimal** numbers provide specific information about the subject matter of published materials. A controversial implementation of identifiers is the recent revelation that **Microsoft** uses a “digital fingerprint” in its Windows 98 software. This code, which is generated when the Windows 98 operating system is installed, represents specific information about a user's hardware configuration. When the system software is registered over the Web, the number decodes and the information enters Microsoft's customer database. Identifiers are also embedded in documents generated by other Microsoft products, such as Word. Microsoft has claimed that it never intended to violate the privacy of its consumers and has vowed to remove it in the future. (Hiawatha Bray, “Privacy advocates decry digital ‘fingerprints’”, *Boston Globe*, March 9, 1999.) One might argue that a revised and less nefarious version of Microsoft's traceable **digital fingerprint** might have a positive practical application for a system serving long-term storage needs.

The term “**self-describing**” is often used in connection with storage and file formats. The term appears in the book “*Essential Distributed Objects Survival Guide*,” where authors Orfali, Harkey and Edwards conclude their chapter on Bento with the words:

So what do you think of these self-describing files within files? We [...] believe that these new compound document file systems -- OpenDoc's Bento and OLE's compound files -- offer tremendous opportunities for developers and system integrators. Just imagine all the good things you can do with files filled with self-describing data.

(Orfali, Roberd, Dan Harkey, Jeri Edwards. *The Essential Distributed Objects Survival Guide*. New York: John Wiley and Sons, 1996. p. 385)

Many agree that among these “good things” is the **Warwick Framework**. Developed through the **Dublin Core** workshops, the **Warwick Framework** is a “container architecture” that describes how digital objects might either be embedded in a source file or referenced to external files or other storage areas. This information might include domain specific descriptions, terms and conditions for document use, pointers to all manifestations of document or archival responsibility. The **Warwick Framework** recommends that bit streams within this architecture be self-describing.

(<http://cs-tr.cs.cornell.edu:80/Dienst/Repository/2.0/Body/ncstrl.cornell/TR96-1593/html>)

Tagging **ASCII** or plain text with bracketed codes is another method for self-describing documents. In this context, the **Standard Generalized Markup Language** or **SGML** might be considered one of the first tools for self-description. An ISO standard “metalanguage” since 1986, **SGML** is used to create data files that are platform- and application-independent.

The **University of Illinois Pablo Research Group** has designed a meta-language called the **Self-Defining Data Format (SDDF)**. This “performance data file format” specifies a “data record structure” or “packet” to describe the layout of records, as well as the “data record instances” or actual data to be processed. (<http://www-pablo.cs.uiuc.edu/Project/SDDF/SDDFOverview.htm>)

Self-description is central to one of the most exciting Web developments to emerge in the past few years: the **eXtensible Markup Language** or **XML**. Most World Wide Web developers are familiar with **HTML**, the system of tagging format and linking used on the World Wide Web. Like **HTML**, **XML** has roots in **SGML**. Also like **HTML**, documents tagged with **XML** are intended to be understood by cross-platform browsers. Unlike **HTML**, however, **XML** was devised to enable content creators to create their own sets of markup tags. These more robust tags could specify complex relationships among elements within a page, among many documents within a site and across sites. Though the **XML 1.0 Specification** has only recently been published as a **World Wide Web Consortium (W3) Recommendation** (<http://www.w3.org>), **XML** has been used to create several “vocabularies,” including the **Astronomical Markup Language (AML)**, the **Bioinformatic Sequence Markup Language (BSML)** and **Microsoft’s Channel Definition Format (CDF)**. One **XML**-based language with enormous potential for the broadcasting industry is the **Synchronized Multimedia Integration Language (SMIL)**, which is currently supported by **AT&T/Bell Labs**, **Microsoft**, **Netscape**, **Philips**, **RealNetworks**, and **Sun Microsystems**.

The impact of **XML** on Web development is likely to be enormous. Its impact on digital preservation remains to be explored. Michael Day, metadata research officer at the **UK Office for Library and Information Networking**, writes, “Text-only 'bootstrap standard' metadata [might be] attached to the data which would provide contextual information and an explanation of how to decode the record itself.” (Michael Day, “Extending Metadata for Digital Preservation,” *Ariadne: the Web Version*, May 19 1997 issue 9. <http://www.ariadne.ac.uk/issue9/metadata/>) The **Research Library Group’s Working Group on Preservation Issues of Metadata** has demonstrated how its recommended list of “preservation metadata elements” might be incorporated into an **XML** record. (“*RLG Working Group on Preservation Issues of Metadata*,” Appendix 3: **XML Implementation**, <http://www.rlg.org/preserv/metaapp3.html>)

Utilizing **XML** for the **Universal Preservation Format** could involve the creation of new tools to assist in the binding of **essence to metadata** and to access stored materials through next generation versions of popular browsers. **XML**, through its related standards, **Xlink** and **XPointer**, might also be used in the creation of the **UPF’s digital Rosetta stone**.

The **World Wide Web Consortium** recently announced an initiative to reformat **HTML** as an application of **XML**. This W3 project shares at least one source of inspiration with **UPF**; its code name is “**Voyager**.”

## **6. Assessing User Requirements**

As we investigated the user requirements for a digital archiving system, we gathered and processed information from many different sources, including academic papers, recommended practices, technical specifications, working and study group documents. These documents have been supplemented with fieldwork: soliciting input through user surveys, conferences and meetings.

In keeping with our cross-departmental origins, we sought the participation of organizations from a wide range of domains: the **Association of Moving Image Archivists**, the **Society of American Archivists**, the **Music Library Associations**, **Boston Art**

**Conservation and Conservation On-line.** We continue to exchange ideas with individuals who may not be members of these groups but who are actively involved in preservation issues. In addition to our own **UPF listserv**, which currently has close to 80 members, we routinely post updates about this project to the **AMIA listserv**, the **Archivist & Archives listserv**, and the **Electronic Records listserv**. We also send mailings to archivists who do not have email addresses.

### 6.1 Review of the Literature

There are many initiatives concerned with accessing digital materials, but few deal with long-term digital storage. In his essay, "Archiving and Authenticity," David Bearman, editor of **Archives and Museum Informatics**, points out that digital preservation initiatives have focused on one of four areas: the digital signal, intellectual content, software independence and social and legal standards. Bearman is well known for his pioneering work in network standards, including the "**Business Acceptable Communications (BAC)**," a reference model published in 1995 "to facilitate implementation of environments in which electronic evidence can be created." The BAC reference model was an early proponent of encapsulation as a method to promote software independence and interoperability.

We envisage transactions taking place as metadata encapsulated objects, although records might not be physically stored in this manner. When transmitted, the contents of the transaction would be preceded by information identifying the record, the terms for access, the way to open and read it, and the business meaning of the communication much as a train of baggage cars is preceded by an engine.

(David Bearman and Ken Sochats, "Metadata Requirements for Evidence,"  
URL:<http://www.lis.pitt.edu/~nhprc/BACartic.html>)

One important paper concerned with the signal level of digital preservation level is Jeff Rothenberg's highly influential "**Ensuring the Longevity of Digital Documents**," published in the January 1995 issue of *Scientific American*. Rothenberg argues that present technology is not sophisticated enough to mimic software behavior. "To replicate the behavior of a program," he writes, "there is currently little choice but to run it." (Rothenberg, p47) In his paper, "Metadata to Support Data Quality and Longevity," he asserts that current data files are both more complex and depend more upon the applications that created them. His conclusion is that one must access digital records through the original application or through emulators.

...there may be no inherent way to know whether to interpret a given sequence of bits as text, number, pointer, image, sound, video, program or new formats...

(Jeff Rothenberg, "Metadata to Support Data Quality and Longevity," para 1 under section "The Assault on the Longevity of Digital Data" URL: [http://www.computer.org/conferen/meta96/rothenberg\\_paper/ieee.data-quality.html](http://www.computer.org/conferen/meta96/rothenberg_paper/ieee.data-quality.html))

Both Rothenberg and the **UPF** initiative promote the concept of the container, which Rothenberg calls "virtual envelopes." These would contain the bit streams along with descriptions of contents and "transformation history."

[...] contents would be preserved verbatim, and contextual information associated with each envelope would describe those contents and their transformation history. This information must itself be stored digitally (to ensure its survival), but it must be encoded in a form that humans can read more simply than they can the bit stream itself, so that it can serve as a bootstrap. Therefore, we must adopt bootstrap standards for encoding contextual information; a simple, text-only standard would suffice. Whenever a bit stream is copied to new media, its associated context may be translated into an updated bootstrap standard. [...] These standards can also be used to encode the hardware specifications needed to construct emulators.

For data to be readable through time, Rothenberg says one must save the hardware specifications in a platform-independent environment. This idea corresponds to the UPF's implementation of a digital **Rosetta stone**, a term Rothenberg uses generically. The **UPF Rosetta stone** would get at the stored data types through platform-independent algorithms. These instructions, saved in a standard plain text format to the storage media, would contain the registry of the data types stored on the media and include their mathematical definitions. The **UPF Technical Requirements** suggests storing this "blueprint" information in analog format at a marked segment within a hybrid storage medium.

The phrase "digital Rosetta stone" appears in a number of other preservation initiatives. Massachusetts Institute of Technology's **Time Capsule File System (TCFS)**, discussed more fully later in this document, uses the phrase to describe its "ideal archival format," which would be "portable across media and platforms and would be easy to reverse-engineer." (<http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/index.html>) The phrase is central to Alan R. Heminger and Steven B. Robertson's paper on their process to assure long-term access to digital documents. Their **Digital Rosetta Stone (DRS)** model would contain:

multiple levels of knowledge about specifications and processes by which information is stored on various types of storage media. It would also contain archives of knowledge about how to meaningfully interpret that information so that the original meaning can be recovered.

(Heminger & Robertson, "Digital Rosetta Stone: Conceptual Model for Maintaining Long-term Access to Digital Documents," published in *European Research Consortium for Informatics and Mathematics*, Sixth DELOS Workshop, Preservation of Digital Information, June 1998, 35-43.)

Heminger and Robertson call the storing of format information, "knowledge preservation." This **metaknowledge**, which resides in a database, defines how data is stored on specific media and how specific software applications format their digital documents.

The authors describe their **Digital Rosetta Stone (DRS)** as a three step process: **knowledge preservation**, the process of gathering and storing the metadata about storage media techniques and file formats; **data recovery**, the process of migrating data from an obsolete medium; and **file reconstruction**, the process of "interpreting digital documents" through knowledge about how specific software formats its information.

The UPF's self-described file format and Heminger and Robertson's data reconstruction model contain many similarities. Both models contrast with Rothenberg's emulation proposal. Though today's file formats are more complex than yesterday's, as Rothenberg asserts, hardware emulation may prove to be less feasible than the preservation mechanisms proposed by the UPF and DRS models. In the past, the more successful emulators have depended upon the installation of special hardware helpers: cards that contain proprietary chipsets or ROMs that contain the original operating system.

MIT's **Time Capsule File System (TCFS)** was developed in response to this perceived drawback. TCFS Project Head Brian K. Zuzga writes, "Emulators increase in complexity as the original hardware increases in complexity, so this problem only would get more difficult with time." ([http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2\\_5\\_2\\_3.html](http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2_5_2_3.html)) Complicating matters are the special configurations that software applications increasingly place on hardware. Even a machine running the correct version of an operating system must be upgraded to run some of today's more demanding COTS products. In short, computer hardware consists of so many components -- RAM, video and sound cards, even analog converters -- that it is not enough to emulate a generic processor.

Adoption of Rothenberg's ideas will unquestionably salvage volumes of material that might otherwise be lost. And however one might criticize emulation as a long-term solution, successful emulators already exist for a wide range of obsolete systems, many available as freeware. It remains to be seen, however, whether emulation addresses the problem of its own perpetual integrity. One can easily envision running one emulation within another emulation to view data in a way that mimics the look and feel of the original application. Why is the emulation approach superior to migrating to new file formats which might also alter in some small way the original? On the surface, it seems a trade-off between system performance and file format fidelity, but if one factors in the costs and inconvenience of continually upgrading emulation software, not to mention the uncertainties of maintaining the interest of future programmers to do this kind of work, versus the potential loss of how data is formatted from one generation to another, then one might lean to the latter solution.

Emulation may be appropriate when it is necessary to preserve not just data but original programming. David Bearman has pointed out that

the task of continuing to devise methods to read old signals in old media is becoming more complex as the media proliferate, recording and layout methods become more proprietary, and firmware plays a greater role in decoding.

(David Bearman, "Archiving and Authenticity," The Getty Art History Information Program: Research Agenda for Networked Cultural Heritage, 1995. URL:<http://www.ahip.getty.edu/gi/new/ranch/archiving/archiving1.html>)

Fundamental questions must be raised about the inherent value of software. Are applications merely the means of generating and accessing data, or might a software program be considered an aesthetic experience, an original creative work, as worthy of preservation as a novel or a painting? If software as self-contained experience is machine- or system-dependent, what are the limits to the archivist's responsibility in assuring its accessibility through time?

**The Commission on Preservation and the Research Library Group Task Force on Archiving of Digital Information** has grappled with this aspect of preservation through its guidelines for migration:

Migration is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to retain the ability to display, retrieve, manipulate and use digital information in the face of constantly changing technology. Migration includes refreshing as a means of digital preservation but differs from it in the sense that it is not always possible to make an exact digital copy or replica of a database or other information object as hardware and software change and still maintain the compatibility of the object with the new generation of technology.

(Commission on Preservation and The Research Library Group, Task Force on Archiving of Digital Information, "Preserving Digital Information," May 1, 1996. <http://www.rlg.org/ArchTF/>, p4)

Refreshing relies on someone in the future taking on responsibility for the process. Maggie Exon, a medieval historian, states in a paper delivered at the 1996 Second National Preservation Office Conference in Canberra, Australia:

A task like information transfer or refreshment which needs to be repeated over and over again, will at some point fail to happen. This failure may take place a few years from now or a few hundred years but it will surely take place.

(Maggie Exon, "Strategies for Long-Term Preservation," Canberra: National Library of Australia 1996 URL: <http://www.nla.gov.au/3/np0/conf/np095me.html>, para 5)

Klaus-Dieter Lehman, **National Librarian of Germany**, in an essay published in

**Daedalus** points out that the current popularity of Java applets, used to develop both distributed documents and distributed programs, complicates strategies to preserve digital applications.

It is no longer necessary to install complete programs on a hard drive; instead, small program modules are written for each specific purpose. These so-called objects -- Java refers to them as "**applets**" -- are available on various computers within the network and can be downloaded as needed and combined to perform a particular task. An interpreter simulating a second computer on a PC insures that the different browser programs understand the applets. The interpreter runs the applets as soon as it receives them. (Klaus-Dieter Lehman, "Making the Transitory Permanent: the Intellectual Heritage in a Digitized World of Knowledge," *Daedalus*, Fall 1996 v125 n4 p307)

In theory, at least, Java itself could have a positive impact on long-term digital storage. As a standard machine-independent programming language, Java could be used to build applications or JavaBeans specifically for accessing or reconstructing data types within an archives.

As a blanketed archival solution, emulation arguably goes against the grain of current archival discourse because, while it does address the integrity of data, it does not solve the issue of informational integrity. Emulation is designed as a tool to interact with data, not simply to view it. The point of an archives is to maintain the content of the data object.

Another approach to promoting data longevity is the widespread adoption of **file format interchange**. One hears the term "native format" quite frequently these days, but what constitutes a "native format" is open to debate, especially considering that most applications offer several options for saving files. Exon posits the question: What information from the period 1995 to 2000 will still be around in 3000?

The information [likely to survive] will be software independent, and standardized formats will be used which are as simple as is possible consistent with the information still being usable. This, of course, poses a problem with much commercial and non-commercial multimedia, in which complex relationships are set up between image, sound and text. However, the move towards **standardized software and hardware formats** has been very strong in recent times and I can see this continuing, leading to the **possibilities of finding ways of translating these formats into relatively simple migratable formats**. (Maggie Exon, "Strategies for Long-Term Preservation," URL: <http://www.nla.gov.au/3/np0/conf/np095me.html>, para 10)

Though the majority of commercial software programs generate their own proprietary file formats, the actual content is rarely glued to these applications. For example, users of the painting program **PhotoShop** can substitute its own proprietary format with a variety of standard formats that can be read by other painting programs, such as **PaintShop Pro**. Word processing documents saved in either ASCII or RTF standards can be loaded into other word processors on a variety of platforms. Other applications, such as FileMaker, allow its proprietary data files to travel seamlessly across platforms.

In addition, open standard formats exist for many individual data types, and their specifications can be easily obtained through the Internet. Standard interchange formats for compound documents are also being developed. The **Advanced Authoring Format**, announced by Microsoft at the 1998 **National Association of Broadcasters**, is described as "an industry-driven, cross-platform, multimedia file format that will allow interchange of media and compositional information between AAF-compliant application." ("AAF Specification") The AAF standard is being developed through a collaboration among several software companies -- Adobe, Avid, Matrox, Microsoft, Pinnacle, Softimage, Sonic Foundry and TrueVision -- as a way to address the shortcomings of currently available digital media

file formats.

The Advanced Authoring Format resembles other container formats, which include Sun's **JavaBeans**, Apple's **QuickTime 3.0**, Avid's **Open Media Management** and Microsoft's own **Advanced Streaming Format**. This should come as no surprise since AAF is based on Avid's **Open Media Framework**, which is built upon **Bento**. Like its predecessors, AAF uses encapsulation to allow applications to identify specific, even proprietary data types. Before these container technologies existed, applications generated separate files for each data type. To move these materials into another application, one often had to convert each of these data types into a series of new formats. Relationships among the converted materials had to be re-established, which was often a laborious procedure. Container technologies, on the other hand, allow for these complex relationships to travel with the data types.

The **AAF** specification attempts to resolve the difference between an authoring system and a delivery system through "file flattening," a process that strips AAF files of its metadata. Two years ago the **UPF** incorporated this file flattening in the form of a media compiler, an apparatus which would "remove the acquisition format from the archives requirement."  
(MacCarn, "Toward A Universal Data Format for the Preservation of Media")

Metadata is a crucial component of an archival system, and the **UPF** is designed to incorporate as much metadata as needed. The integrity of digital information is dependent upon the quality of metadata. Peter S. Graham, associate university librarian for technical and networked information services at Rutgers University, in his paper, "Preserving the Digital Library," has described levels of preservation modes: **medium preservation** (tapes, disks, CD-ROMs, etc.); **technology preservation** (storage formats, operating systems); and **intellectual preservation**, "which addresses the integrity and authenticity of the information as originally recorded."

(Peter Graham, "Preserving the Digital Library," 1995, paper presented at "Long Term Preservation of Electronic Materials," JISC/British Library Workshop, Nov. 1995), <http://www.ukoln.ac.uk/services/elib/papers/other/preservation>)

Guidelines for all levels or modes of information have been described in great detail by the **Commission on Preservation and The Research Library Group Task Force on Archiving of Digital Information** under the section, "Information Objects in the Digital Landscape." The **Open Archival Information System (OAIS)**, discussed above, has adapted these guidelines as components of its **Preservation Description Information**.

**Content** is defined by **Task Force on Archiving of Digital Information** on several levels, from its very bit stream to its intellectual essence.

the preservation challenge for digital archives is to migrate [...] **intellectual content** using standard interchange algorithms and other appropriate migration strategies so that the **ideas** available in the end are identical to those contained in the original object. The measure of integrity in the preservation process thus turns, at least in part, on informed and skillful judgments about the appropriate definition of the content of a digital information object -- about the extent to which content depends on its configuration of bits, on the structure and format of its representation, and on the ideas it contains -- and for what purpose.

(Task Force on Archiving of Digital Information, "Preserving Digital Information: Executive Summary," May 1, 1996, p.13. URL: [http:// www.rlg.org/ArchTF/](http://www.rlg.org/ArchTF/))

Margaret Hedstrom, associate professor in the School of Information and Library Studies at the University of Michigan, has examined the complexities of digital content in her seminal paper, "Digital Preservation: a Time Bomb for Digital Libraries." Among her observations is that many librarians and archivists are resorting to transferring digital information to analog

formats as a preservation strategy. Hedstrom also points out that some types of digital materials do not lend themselves well to these "print equivalents."

(Margaret Hedstrom, "Digital preservation: a time bomb for Digital Libraries," <http://www.uky.edu/~kiernan/DL/hedstrom.html>)

**Fixity** refers to digital information as a discrete object or individual thing, unique from other things. The **Task Force on Archiving of Digital Information** offers the publication of books and the broadcasting of radio and television programs as analog examples. Defining fixity for digital objects is more complicated. As Peter S. Graham of Rutgers points out, fixity for printed material is a given:

The printed journal article I examine because of your footnote is beyond question the same text that you read. Therefore we have confidence that our discussion is based upon a common foundation. With electronic texts we no longer have that confidence.

(Peter Graham, "Intellectual Preservation: Electronic Preservation of the Third Kind" 1994, <http://sul-server-2.stanford.edu/byauth/graham/intpres/>)

Graham reports that one way digital fixity might be achieved is through **digital time stamping**, defined as "a means of authenticating not only a particular document, but its existence at a specific time." (Graham, "Intellectual Preservation: Electronic Preservation of the Third Kind")

The **digital watermark**, described above, is another method to help ensure fixity. Still another method might be to incorporate into the digital object a built-in or embedded database that records a kind of chronicle of changes. Donald Waters, Director of the Digital Library Federation, has expanded upon this concept:

some digital information objects are better modeled as continuously updated databases for which the preservation choice is whether to compile a complete record of changes or to capture snapshots of the database as the means of preserving information integrity.

(Donald Waters, "The Impact of Electronic Publishing on the Academic Community: Session 5: Digital Libraries and Archiving of Electronic Information: Choices in Digital Archiving: the American Experience," 1997, <http://www.portlandpress.co.uk/books/online/tiepac/session5/ch6.htm>)

OAIS defines its **Fixity Information** as a "protective shield" to ensure the authentication of its stored contents. A functional example of **fixity** in this sense might be the checksum at the end of an ISBN number, while the ISBN number itself would fall under the category of **Reference Information**.

**Reference** as an aspect of integrity means that information objects must consistently be findable and identifiable among other objects. The **Task Force on Archiving of Digital Information** cites the new MARC 856 field "to incorporate reference to digital objects." (Task Force on Archiving of Digital Information, p15.) The Study also singles out the work being done by the **Internet Engineering Task Force (IETF)** for the unique identification of digital information objects: the **Uniform Resource Name (URN)**, the **Uniform Resource Location (URL)** and a proposed framework called the **Uniform Resource Characteristics (URCs)**, which would include intellectual rights, property rights, context and provenance

**Provenance** is a term familiar to most archivists. James O'Toole identifies it as the principle by which archives are generally organized, and:

is based on the deceptively obvious insight that the person or organization producing the records determines their content. A nuclear engineer will produce files and documents that are essentially different from those produced by an impressionist painter, even if both of them talk about a whole range of subjects in their letters

(O'Toole, James, *Understanding Archives and Manuscripts*, Chicago: Society of American Archivists, 1990, p.55)

Elizabeth Yakel cites provenance as the key difference between how archival records and library materials are organized, a distinction that has clear application to digital realms:

Books are created to provide information on a specific topic and are physically organized by subject. By their nature, archival records are an organic byproduct of an institution, activity, or person. Therefore, maintaining the context in which archival materials were created is absolutely essential to future historical understanding of an organization, individual, or activity. Archivists have found that the best method of maintaining the context is to organize records according to their "provenance" or creator.  
(Yakel, Elizabeth, *Starting an Archives*, Chicago: Society of American Archivists, 1994, p.39)

The **Task Force on Archiving of Digital Information** describes provenance as the "assumption [...] that the integrity of an information object is partly embodied in tracing from where it came." (Task Force on Archiving of Digital Information, p15) For digital objects, provenance would include tracing migrational history on several levels: through individuals as sources and owners of objects, corporate affiliations and underlying policies, scientific data and instrumentation, and migration activity within one's own organization, called "the chain of custody." (Task Force on Archiving of Digital Information, p17)

In a word, provenance is history. **OAIS** defines its **Provenance Information** as information that documents the origin or source of its stored contents, as well as any changes made to it.

**Context** is defined as how digital information objects "interact with elements in the wider digital environment." (Task Force on Archiving of Digital Information, p18) On a technical level, digital information depends on a specific configuration, such as a specific chip or operating system. Another level is the "linkages" of these objects, best illustrated by Web pages. For archives, to maintain integrity, objects and linkages would have to reside within the same storage system. On the communications level, attributes such as bandwidth and security would have to be considered as part of an information object's integrity. The final factor under context is the social environment of digital information. The study cites email as both a personal communication tool and a viable "vehicle for formal communication among academic or business colleagues." (Task Force on Archiving of Digital Information, p19) **OAIS** adds to its **Context Information** the requirement to address the question of why the content was created in the first place.

**Information integrity** has been discussed in several other papers using other terminology. Graham, quoted above, refers to this quality as **Intellectual Preservation**. More recently, the **Society of American Archivists** issued its "Statement on the Preservation of Digitized Reproductions." The document points out that the integrity of digital information "begins with limiting the loss of information that occurs when a file is created originally and then compressed mathematically for storage or transmission across a network." Echoing the recommendations of the **Task Force on Archiving of Digital Information**, SAA urges archivists to "migrate valuable digitized data, indexes and software from one generation of computer technology to a subsequent generation."

(<http://www.archivists.org/governance/resolutions/digitize.html>)

Perhaps the most practical attempt to ensure the integrity of information is the NHPRC-funded project, "Functional Requirements for Evidence in Recordkeeping" (Grant No. 93-030). Better known as "**The Pittsburgh Project**," this initiative, directed by Richard Cox and James Williams, developed metadata "products" for specific professional domains, including legal, medical, accounting, information and records management. Each of these reference models includes six levels of metadata: **Handle, Terms and Conditions, Structural, Contextual, Content and Use History**. (<http://www.sis.pitt.edu/~nhprc/>)

## UPF as Tangible Product

Re-formatting as a means of converting obsolete videotape holdings poses two major dilemmas for the archival world: the lack of an ideal video format and the growing volume of material to be copied. Beyond the need to go to a digital format to avoid generational loss, absent from the archival field is anything remotely approaching what might be called an ideal format or a "preservation copy." Until an ideal or universal preservation format is introduced, video preservation should be viewed not as a tangible product but a continuing process aimed at protecting information that can migrate from one technology to another as the need arises. The current merger of video technology and computers suggests that the ideal format in the future may not be videotape but bitstreams of compressed data recorded on disks.

("Television and Video Preservation 1997: A Report on the Current State of American Television and Video Preservation" Report of the Librarian of Congress, October 1997 v1 p41.)

The above passage begs the question, Can an "ideal or universal preservation format" be introduced now? Is the necessary technology available to handle digital media of all types? We believe it to be so, and in a paper published in the **SMPTE Journal**, WGBH's Dave MacCarn cites Apple's **Bento Specification** and Avid Technology's **Open Media Framework Interchange Specifications** as "both media technologies that approach the UPF concept."

Bento defines a standard format for storing multiple different types of objects and an application program interface to access these objects... 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.

(Dave MacCarn, "Toward a Universal Data Format for the Preservation of Media," *SMPTE Journal*, July 1997 v106 n7 p477-479.)

The **Open Media Framework Interchange (OMFI)** builds upon Bento to establish a standard format for the interchange of digital media data among different platforms. It encapsulates in a single file whatever information it needs to transport digital media across platforms.

In addition to **Bento** and **OMFI**, other file formats and mechanisms have been developed for compound documents, some of which resemble the proposed functions of the UPF.

The **Hierarchical Data Format (HDF)**, for example, is designed for the scientific-visualization market. This standard is maintained by Fortner Software and Research Systems, Inc. and supported by the **National Center for Supercomputing Applications**. HDF can import data sets from virtually any format. (<http://hdf.ncsa.uiuc.edu/>)

HDF data files are "self-describing" in the sense that they include information describing the type, storage structure, and location of the data in the file. They also provide convenient structures for applications to store application-specific metadata. HDF files are also "architecture-transparent" in the sense that a file's contents is represented in a form that can be accessed by computers with different ways of storing integers, characters, and floating-point numbers. HDF files also provide structures that facilitate efficient direct access, so that a small subset of a large dataset may be accessed efficiently, without first reading through all the preceding data.

(Mike Folk, "HDF as an Archive Format," (A part of the ISO Archiving Workshop Series) <http://ssdoo.gsfc.nasa.gov/nost/isoas/dads/DADS16.html>)

The **Time Capsule File System**, developed by the Massachusetts Institute of Technology's

Artificial Intelligence Lab and Laboratory for Computer Science, is a "universal framework for preserving any archival file in a platform- and medium-independent fashion." (<http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/thesis.html>) MIT's decision to develop its own file framework is instructive. As Brian K. Zuzga reports, he and Dr. Alan Bawden considered several options when faced with the challenge of preserving "rooms full of 7-track and 9-track magnetic tapes in various states of decay." ([http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/section2\\_5\\_1.html](http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/section2_5_1.html)) They rejected preserving only the raw byte stream because they determined that "the knowledge to interpret data is being lost much more quickly than the data on the tapes themselves." ([http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2\\_5\\_2\\_1.html](http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2_5_2_1.html)) They considered maintaining the original hardware, but decided that the expertise required to make repairs is also vanishing. Emulators were also considered and rejected for similar reasons; they necessitated an indeterminate amount of time and money to port the emulator to operating systems of the future. And though they were already maintaining full copies of the original operating systems and applications, they determined that this method inadequately solved the problem "of reducing the number of formats that we need to decode in order to read all of our data." ([http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2\\_5\\_2\\_4.html](http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2_5_2_4.html)) Finally, they considered migrating their files to a small group of standard file formats. Their problem with this approach was the uncertainty as to whether these file formats would still remain standards or would require them to "translate our documents each time a new format comes along." ([http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2\\_5\\_2\\_5.html](http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/subsection2_5_2_5.html))

**Mikhail Popov**, affiliated with the Centre for Physics and Information Science (Tsentr fizicheskikh issledovaniy i informatiki pri OIYaI), has created the **SDF (Self-describing Data Format)** software to assist in "the transfer of binary data between different machines and different architectures," as well as in the "storage and retrieval on a single architecture." (<http://unix4.jinr.dubna.sa/sdrintro/sdintro.html>)

In 1992 **Adobe** (<http://www.adobe.com>) introduced its **Portable Document Format (PDF)** as a "universal electronic publishing tool." Though the software driver required to create PDF files is a commercial product, the **PDF Reader** is freely available for several platforms. The PDF format is generally regarded as the "de facto" standard for publishing professional and technical documents on the World Wide Web. As a measure of PDF's predominance on the Web, Netscape plans to incorporate "direct PDF format capability" into its next generation of Web browsers. (<http://www.imaging-resource.com/acrobat.html>)

## 6.2 User Survey

Available on the UPF Web site since August 1997, the **UPF User Survey** is intended as a continuous thread of input throughout the UPF project. Though questions are presented in a multiple choice format, the real value of the survey may be found in opportunities for comments and personal anecdotes that follow each question. Many archivists responded with "mini-essays" that illustrate some of the key points expressed in the academic literature. Others raised practical concerns that have not been significantly addressed by theorists.

Underlining most of the comments was the issue of funding. Many archivists reported that their departments couldn't justify the expense of investing in new digital technologies, especially in equipment that may become obsolete after a few years. Some expressed frustration over their inability to control technologies already purchased. The maintenance and replacement of obsolete hardware was cited as a major expense. Several archivists felt that standardizing digital equipment or parts would alleviate this problem dramatically.

Budgets also affected hiring practices. A few managers reported choosing staff with computer skills over those with experience or training in cataloging or records management. As a

result, decisions formerly made by long-term strategists are often made by junior-level staffers who may have technical know-how but who also may be implementing short-term fixes.

No one who responded to the survey seriously contemplated replacing analog collections with digital surrogates. In general, archives are experimenting with digital formats to provide selective online public access to their collections. The scope of this phase may broaden, but commentators wrote that their institutions are waiting for open systems before they will consider migration in any serious, systematic fashion. One archivist suggested that until open standards are in place, money would be better spent on developing metadata systems that protect the integrity of media and persevere through time.

The theme of integrity and identification of original works also ran throughout the commentaries. It was imperative, wrote several contributors, that a standard system of digital document identification be developed and administered by a central registry. Any changes to a registered document would result in that document's being given its own unique ID which would reference the original or parent version.

Finally, several archivists who responded to our survey suggested that continuing education should play a stronger role in preparing veteran archivists for new technologies. Archival societies should consider on-line tutorials for members. Archival conferences should include sessions that introduce digital formats and teach specific digital preservation strategies. A universal glossary or thesaurus would also be a greatly appreciated tool, wrote one archivist. A desire for a cross-domain lexicon also was echoed throughout several **SMPTE** sessions.

### **6.3 SMPTE meetings**

**The Society of Motion Picture and Television Engineers (SMPTE)** is an international technical society, founded in 1916, "devoted to advancing the theory and application of motion-imaging technology including film, television, video, computer imaging, and telecommunications." It has 8,500 members in 72 countries. On September 22, 1997, **SMPTE** assigned the UPF an official **Study Group (ST13.14)**. Titled "Requirements for a Universal Preservation Format" and chaired by WGBH's Dave MacCarn, the group first met to establish an agenda and a statement of objectives, which includes gathering input from the archival community through surveys, meetings and conferences. **SMPTE** meetings are held quarterly. Because the engineers who attend our meetings are also deeply involved in other **SMPTE** standard-setting groups, archivists whom we have invited have a unique opportunity to effect the direction of emerging technical standards.

The first meeting of the **SMPTE UPF Study Group** to include archivists was held on December 9, 1997 at the Sony Headquarters in San Jose, California. It was attended by twenty engineers, including the chairs of other related **SMPTE** committees: Stephen Long from the **National Imagery and Mapping Agency**, chairing the **Work Group on Metadata**, and Juergen Heitmann from AV Media Technology, chairing a **Study Group on Automated Storage and Retrieval**. We were joined by Robin Dale, who shared with us some concerns of her organization, the **Research Library Group**, specifically its "Working Group on Preservation Issues of Metadata."

The members of the **RLG** are mostly involved with the migration of static still images. Because there is no standard practice for saving these images in a single file format, digital images are stored on a wide range of formats. Dale expressed the hope that a universal file format will be flexible enough to transport any kind of still image file.

Though the idea of a preservation format appealed to her organization's membership, Dale

pointed out that most research librarians would resist the imposition of any given standard. Librarians, Dale said, “prefer their own methods of doing things.” For example, several standards exist for cataloguing. Recommended practices are frequently being appended or replaced to accommodate new media. An initiative that is likely to win acceptance in her community will be flexible in incorporating practices already established by research libraries and archives. She added that a “best practice” framework would have greater appeal to her group than a single standard. This framework could define the types of data that might be included in a preservation file format.

The next meeting of the SMPTE Study Group was held on March 12, 1998 in Atlanta, GA at the Turner Entertainment Building. Despite the early morning meeting time, seven representatives of local archival institutions attended.

Should a digital preservation specification consist of a series of standard file formats or would a “preservation mechanism” that ensures lossless migration to new technologies better serve the archival community? While this question was discussed but not resolved, all agreed that obtaining the perspective of the archivist was crucial to developing a standard for digital storage. Some of the engineers admitted that the companies they represented needed to catch up with what the archival communities have known for decades: the realization that materials must be identified.

“There are things that this community takes for granted,” said one engineer, “things that other communities have to understand.” Solutions that serve the day-to-day needs of those working in video post-production may not be useful “in terms of maintaining this for years, let alone decades.”

Specifically, the video data stream should contain the kind of information traditionally found in the title pages of books. This information would remain with the video from the moment it is created or recorded as a primary source, or the point at which authorship begins.

It was pointed out that because **SMPTE** lacks the expertise to define all the classes of metadata, content experts in other domains must actively define their own classes. **SMPTE** can provide standards to enable the transport of these classes of metadata in a way that will make it easier “to migrate content with metadata from one physical carrier to the next one in 15-20 years.”

One conservator assured the engineers that domain-specific systems of metadata have already been developed for electronic records. “What we have not done is address the technical issues.” The study group was cautioned not to accept all input collected from archival and library communities as “equally valid.” Levels of experience and expertise vary, and not all spokespersons recognize the complexity of the problem. The study group pointed out that UPF activities are examined and guided by a review board, whose members are well regarded in the preservation field.

The fourth meeting of the UPF Study Group was held on June 10, 1998 at the Microsoft Headquarters in Redmond, WA, attended by a dozen archivists and 20 engineers.

For the first time we discussed how specific characteristics of the **Universal Preservation Format** might serve the mission of the archivist. Engineers asked if archivists thought digital materials should be stored in the format on which they were created or in formats that increased their chances of survival. One archivist expressed the view that it was important to preserve the “look and feel” of the original media so that future users could recreate “the context in which a particular item was intended to be experienced.”

The impact of home computing on emerging storage technologies was another topic raised at this meeting. The issue is relevant to a universal preservation format because consumer needs may eventually define the boundaries between what is considered an acquisition or program format and one designed for platform-independent long-term storage. When video cameras replaced 8mm film, many families rushed to have their old 16- and 8mm home movies “preserved” on video. As equipment to digitize becomes more affordable, families are scanning their family albums and converting their home videos into digital formats in the belief that these materials are being preserved for future generations. The sad truth is that digital devices marketed for consumers may prove to be less adequate for long-term storage than the medium they are replacing. The hope was expressed that public awareness and consumer demand for better computer storage products will help archivists in their quest for more affordable and durable storage tools and devices. At the same time, one engineer emphasized, “archivists must demand that manufacturers provide tools for migration management, that make it easy to migrate from one file format to the next file format technology that may come with the next operating systems.”

The most recent meeting was held on September, 1998 at the Avid headquarters in Tewksbury, MA, during which the issues outlined above were further discussed and developed.

#### 6.4 Conferences

From the moment Dave MacCarn introduced the UPF at the 1996 annual meeting of the **Association of Moving Image Archivists**, archival and library conferences have been important vehicles for exchanging ideas and raising public awareness about UPF. To date, our team has presented the UPF at the following:

Aug. 1997	<b>Society of American Archivists (SAA):</b> Electronic Records Section	Mary Ide
Oct. 1997	<b>Music Library Association</b> , New England Chapter	Thom Shepard
Nov. 1997	<b>Association of Moving Image Archivists</b>	Shepard, MacCarn
Feb. 1998	<b>Music Library Association</b> , National Conference	Thom Shepard
June 1998	<b>European Research Consortium</b> <b>Sixth DELOS Workshop</b>	Dave MacCarn
June 1998	<b>American Institute for Conservation</b> <b>26th Annual Meeting</b>	Thom Shepard
June 1998	<b>Electronic Media Special Interest Group</b> <b>American Library Association</b> <b>ALCTS PARS Reformatting Discussion</b> <b>Group</b>	Thom Shepard
Sept. 1988	<b>Museum Computer Network</b>	Thom Shepard
Dec. 1988	<b>Association of Moving Image Archivists</b>	Paul Messier Eddy Zwaneveld Dave MacCarn

In general, our call for collaboration between archivists and engineers in adapting existing technologies for long-term digital storage has been met with both enthusiasm and guarded optimism. Specific qualities of the UPF that audiences seem to understand and support include the **container** structure (metadata and essence stored together) and the **hybrid model** (digital materials stored with analog instructions). Other archivists have expressed their conviction that no storage system that depends upon software to retrieve its contents

can ever be considered truly archival.

Coming full circle, the UPF presented its final session at the AMIA Conference in Miami, FL on December 9, 1998. This year, we invited speakers from outside the project to discuss how a UPF storage standard would impact upon digital preservation. Dave MacCarn presented an overview and introduced our guest speakers.

Ed Zwaneveld, director of Technical Research and Development for the **National Film Board of Canada** and chair of the **AMIA Preservation Committee**, pointed out that video recording systems were originally conceived to be “interoperable in all production and post-production phases.” The trend now is toward “application-specific equipment,” each of which has its own proprietary “standard.” He spoke about the need for true interoperable standards that will last 50 years or more, and discussed the work of SMPTE and EBU working groups to develop such a standard. He pointed out the UPF’s connection to these technical initiatives, particularly the **SMPTE/EBU Task Force for Harmonized Standards for Exchange of Program Material**, and its call for the development of wrapper technology specifically designed for archival use.

Paul Messier, founding member of the **Boston Arts Conservation** and founder of the newly formed **Electronic Media Specialty Group of the American Institute for Conservation**, focused on the role of the UPF in helping to ensure “the integrity of digital cultural materials.” He reminded the audience that “a digital work is not only information, but potentially a piece of our shared material culture.” He pointed out that becoming custodians of digital information is a new challenge for archivists. “We have difficulty assessing where the value of a particular digital work lies.”

Both speakers contrasted UPF with migration. Zwaneveld demonstrated that playback quality can diminish dramatically as moving image material migrates from one digital format to another, while Messier pointed out that if digital materials are determined by conservators to have intrinsic value, then preservation strategies other than migration must be developed. “A migration strategy based on creating a new digital surrogate is not a preservation technique,” he said, “since the new digital object was often significantly altered as compared to the original.”

## 7. Summary

At the March 12, 1998 session of the SMPTE UPF Study Group, Juergen Heitmann (AV Media Technology, Seeheim-J, Germany) told us, “I look forward to a very short list of requirements coming out of this group.” According to responses from SMPTE UPF SG meetings, our User Survey and other sources described in this document, these requirements may be itemized as follows:

- A digital preservation system should be designed to last at least 100 years.
- Archivists must be able to store their materials in a platform-independent way.
- Instructions must be provided on how to retrieve material if the software application used to create it is no longer available or accessible.
- Digital solutions purchased today must not be obsolete in the future.
- Documents stored in an archival system should be uniquely identified. Changes or versions of these documents must be identified as both separate from and related to the source document.
- Many archivists call for a central registry of identifiers.
- The only “complete” system for long-term digital preservation must incorporate both analog and digital elements.
- A table of contents for a digital storage system should be viewable without the need of

- special digital devices.
- A storage system should allow for input from a limitless number of sources. For example, one should be able to export from a graphics application, a word processor, or a database program seamlessly onto a long-term digital storage medium.
- Ideally, the system should allow for the various data types to “know about each other”; that is, the user could build associations among the stored objects. In a sense, the storage system itself could become a database with its own search and retrieval mechanism.
- A storage system must include an easy, perpetual, and foolproof method to determine the stability of the storage medium and the integrity of the stored contents. This test may be the equivalent of a standard test pattern.

## 8. Conclusion

Given the current mindset of the computer industry, which seems to regard digital equipment, storage media and file formats as disposable as paperclips, archival institutions and organizations are struggling courageously with technical obsolescence either by developing in-house tools or by adapting commercial products that are woefully inadequate.

In the early stages of this project, we sought to bring together engineers and archivists because we believed that digital technologies and products have been developed and marketed without direct input from the professionals who use them. We quickly learned that problems of communication were not limited to these two camps. Archival communities are far from reaching a consensus on digital preservation. Though much work is being done to make digital materials accessible and searchable over networks, less effort has gone into developing open standards or shared systems designed specifically for long-term digital storage.

For many years, analog archivists and librarians have grappled with the kinds of issues outlined in this paper. As Margaret Hedstrom has pointed out:

Preservationists within the library and archival community have been instrumental in developing an array of tools and methodologies to reduce the decay of traditional materials and to restore books and documents that have deteriorated to such an extent that their longevity and usability are threatened.

(Margaret Hedstrom, “Digital Preservation: a Time Bomb for Digital Libraries,”  
<http://www.uky.edu/~kiernan/DL/hedstrom.html>)

It is time to study the past accomplishments of librarians and archivists, map their findings and continually solicit their input in the common quest to preserve digital materials. For example, along with “retro-active” solutions mentioned throughout this paper, one would be wise to study “pro-active” strategies used in the past, such as the successful effort to standardize acid-free paper.

Finally, as we advocate a specific set of standards for the long-term storage of electronically generated media, we seek to raise public awareness about the need for universal computer storage standards. The integrity of digital information is a moral issue. It affects both institutions and individuals. The day is here when many of history’s primary text and visual sources -- photos, films, letters, and other documents of the human record -- are generated electronically. The authenticity of this “evidence” must be assured. As we have demonstrated, the technological foundation for a universal preservation format has existed for several years, but for any digital preservation model to succeed, we must have storage vehicles that will last at least as long as their analog counterparts. Our hope is that the

Universal Preservation Format initiative will be instrumental in creating a self-described mechanism as durable as the Voyager's "long playing" record.