AUDIO ARCHIVING FOR 100 YEARS AND LONGER: Once We Decide What to Save, How Should We Do It?[1]

A broad interchange of multimedia digital documents that include audio portions is now available. For the widest possible utility, interchange standards are essential. The standards extensions believed necessary are under consideration by international standards bodies and professional associations.

We can view an archived document as being shared by its originator with recipients many years in the future. Each recipient should be able to exploit the shared document in every way that its originator intended, but cannot dialogue with the originator to ask, "How should I read and interpret this portion of your document?" This letter sketches a way of ensuring long-term document preservation with reliable interpretability.

The number of audio bits generated is accelerating rapidly along with the growth of the Internet.[2] This audio information, like most new records—business, governmental, historical, social, cultural, and academic; written records, multimedia recordings, and data collections—are "born digital." Thanks to widespread prosperity and the astounding pace of innovation in digital storage and communications, more intellectual content will be created and shared in the current decade than in all prior human history [3]. For instance, we read [4]:

> Thanks to the Internet and the rapid global expansion of computing, humans and their machines will create more information in the next three years than in the 300,000 years of history dating to the earliest cave paintings and beyond. . . .
> [Industry participants were] quick to pitch the study to Wall Street, adding it to analysts' projections that spending on data storage products is drawing even with spending on computers themselves and that it will account for 70 percent of information technology budgets by 2005. . . . an individual . . . could easily have a terabyte (the equivalent of 250 million pages of text) of stored personal records, photos, and other data by 2005.

If all these records are worth generating (which is not obvious), some of them are worth saving. Society faces two challenges: deciding which works are worth saving and making sure that saving happens. The records at issue are the intangible product of the mind's work and, for those of us who live by and for work of the mind, are our legacy. They include information critical to democratic institutions and to our well-being. They also include records of cultural importance.

We believe that artists and engineers would like their best output to survive well into the future. We cannot today confidently assure anyone that it will be so. We see a snail's pace for a well-known problem: census data from 1960 that had to be rescued from obsolescence in the 1970s [5]; satellite telemetric data that reside on crumbling tapes in government warehouses, and so on. Neither libraries nor other durable institutions have created needed infrastructure; they face practical issues: missing funding, authorization, staff training, technology, and legal uncertainties [6].

What will surely become accepted "best practice" is determined by the following:

- Physical carriers inevitably degrade as they age—some faster, some more slowly, and all influenced by how they are managed
- Digital representations can be copied without any errors whatsoever
- Rerecording (copying) analog representations invariably distorts and degrades them
- Digital technology has become pervasive and inexpensive, and will continue to improve rapidly for the foreseeable future.

This letter focuses on creating archive-worthy digital representations of audio and essential metadata.[3] For many applications, considering audio in isolation would be imprudent. We describe some of the challenges facing digital library systems designers in incorporating audio as a component of multimedia documents.

1 Background

Our current thinking starts with a CPA-led[4] task force inquiry into preserving digital assets [5]. What we know of today's circumstances suggests that they have remained qualitatively unchanged since 1995. Subsequent inquiries and research projects such as [7]–[12] have confirmed and broadened Garrett's [5] conclusions. Recent work has increased since 1995, as has the awareness of the challenges within library and archive communities [13].[5]

We consider the technical requirements of all kinds archiving institutions, from the giant research libraries and national archives to what individuals might want to do on home computing systems.[6] These are more

[1] A refined version of a presentation at the 109th Convention of the Audio Engineering Society, Los Angeles, CA, 2000 September 22–25, Workshop 5: *Digital Libraries, Preservation, and Metadata*; revised 2001 June 29.

[2] The Council on Library and Information Resources reports that 100 years of recorded sound has left us with a legacy of the equivalent of more than 5 petabytes of professionally recorded audio. Libraries are already overwhelmed with preserving everything from cylinders to vinyl. They are drowning in a preservation crisis as they continue to accumulate media in extinct formats and as audio material proliferates at a pace they are unable to match. . . . J. A. Moorer estimates we are distributing terabytes (TB) of new garage band music every day [1], [2].

[3] Ancillary information is conventionally called *metadata*, whereas the actual content is often referred to as *essence*.

[4] The Commission on Preservation and Access (CPA) was founded about 15 years ago to combat the acid paper problem that threatens written works printed between about 1880 and about 1930. It has evolved to today's Council on Library and Information Resources (CLIR), a Washington-based not-for-profit organization working "to ensure the well-being of the scholarly communication system upon which knowledge creation depends."

extensive than the requirements of digital audio archives, such as those that came into service in radio stations in the 1990s. What "pure" audio archives need is approximately a subset of what we believe will eventually be wanted. For instance, we will touch on extensive work under way to facilitate multimedia document interchange, complete with all necessary descriptive data. Any such document might contain not only audio, video, image, and text data, but also scientific tables and computer programs.

The shortfalls in the systematic accumulation and management of digital holdings have been reconsidered in two National Research Council reports [14], [15]. Neither the Library of Congress (LC) nor the U.S. National Archives and Records Administration (NARA) fills the gap.[7]

The biggest digital library activity at LC, the American Memory project,[8] was focused on converting to digital form heirloom papers and photographs to make these treasures widely accessible, not to capture "born digital" records for posterity. However, the American Memory project funding, which was charitable, is exhausted. LC also has had a small project that included audio records [16]. Recently the U.S. Congress has provided one-time funding for digital content [17], but details of how it will be used have not yet been decided.

NARA does not archive broadly, and probably never will. Its mission is to preserve records essential to the proper functioning of the U.S. government. NARA's digital program is arguably too small for the challenges it faces, as illustrated by its difficulty in conforming to a judicial decree ordering government retention of e-mail [18]. It has chosen to focus its limited research resources on analyzing the practical organization of archival digital storage, doing this in a pilot project [19] that overlaps topics emphasized in this letter.

No work since Garrett et al. [5] expresses the challenges as broadly and effectively. To analyze progress since then would take more than appropriate space; we therefore summarize what Garrett et al. [5] teach, acknowledging some subsequent progress, but pointing out that the resources applied have not yet eliminated any challenge, and have not yet begun to address some topics. Our focus is on topics of interest to research libraries (members of the Research Libraries Group). The challenges include the following.

Administrative: In 1995 research librarians were mostly not authorized to engage in digital archiving, and not funded to acquire and manage the required resources. The issue of authorization seems to have relaxed since that time, but sufficient new funding has not been provided, and there is resistance to diverting funding from traditional activities. Funding discussions seem overwhelmed by talk of price inflation for scientific, technical, and medical periodicals.

Synergetic: Research libraries and parent institutions face numerous issues in deciding on how to share content. In traditional libraries, not everyone has equal access to collections.[9] Universities have long competed partly by offering their scholars unique research materials and seem more eager for access to others' content than to provide access to their own.

Ability: Libraries do not have the digital infrastructure and staff skills required for digital serving and archiving. This is exacerbated by an inferior salary structure for librarians relative to that for computing service personnel.

Cultural: To do what is necessary, libraries need to change. Change is difficult and happens slowly for any large institution, especially for departments emphasizing stability and client service along traditional lines. Some universities have combined their library and campus computing organizations. This may have helped toward the availability of digital periodicals, but does not seem to have helped significantly toward the collection of heritage digital content.

Legal: The liability risk of libraries as potential contributors to copyright infringement by their patrons are incompletely understood [14]. The available software is insufficiently helpful toward controlling how and where materials are reused, and there are unresolved tensions surrounding the "fair use" exemptions of the U.S. Copyright Act of 1976. Publishers are selling limited-period licenses for digital content, rather than selling physical carriers (such as books), and the pertinent contract law is unclear and changing.

Selection: The library cost for each accessioned holding is higher than the market price of the work alone, including for one-of-a-kind content whose management requires special skills and more human labor than do

[5] *RLG DigiNews* [13] is a bimonthly web-based newsletter intended to 1) focus on issues of particular interest and value to managers of digital initiatives with a preservation component or rationale; 2) provide filtered guidance and pointers to relevant projects to improve our awareness of evolving practices in image conversion and digital archiving; and 3) announce publications (in any form) that will help staff attain a deeper understanding of digital issues.

[6] Jim Gray of Microsoft points out that a terabyte (10^{12} bytes) of storage is sufficient to store all spoken communication an individual hears in her lifetime (personal communication). Today, this much disk storage can be added to a personal computer for about $3000. The cost of storage has been dropping at about 26% a year, and this is projected to continue for at least 5 more years and probably for 10 more years (personal communication by colleagues in the IBM Almaden Research Center). In a decade, to store "everything" will cost about $150.

[7] U.S. facts are cited for illustration; the situation is similar in many other nations.

[8] See http://lcweb2.loc.gov/.

[9] Recall the Index Librorum Prohibitorium, as described in the Encyclopedia Britannica V, 327, 1976 edition. A current example is the British Library, which restricts access to its reading rooms to researchers who can demonstrate that they have exhausted other sources. See http://www.bl.uk/information/reader-admissions.html. Recently Rayna Green reminded us of such practices in [2, pp. 47–49]. How limitations of access are often forced on archives is described in [20], which continues with an articulate recommendation of procedural steps for archivists, donors, and patrons. Such considerations are a source of requirements for provenance, copyright, and conditions metadata that archive document structure should accommodate.

widely published works. Furthermore, the number of new works is much larger than it was before the information age, partly because digital technology enables almost anyone to publish and partly because prosperity and wider literacy have vastly expanded the population with the knowledge and leisure to create interesting material.

Technical: Assuming progress against these challenges, there remain technical challenges. These can be partitioned into standards for metadata, saving bit streams,[10] and ensuring that our successors will be able to read and play the bit streams a century from now.

Digital technology and infrastructure are in their infancy compared with their counterparts for handling paper, which have been refined for about 3000 years. Anyone who might otherwise think that digital information will soon displace paper-based information should reflect on the fact that our largest civilian employer is devoted to moving paper; it is the U.S. Post Office, with over 700 000 employees.[11] For those of us concerned with archiving, there is a bright side—massive loss of digital information has probably not yet occurred; we would like to keep it that way.

2 What Is Worth Saving?

While the technical challenges will be easily addressed, among the other challenges, at least one seems difficult: choosing what to save. George [21] suggests how difficult this is for humanities documents. Selection even by disciplinary experts is as much governed by fashion as by stable, objective criteria, exacerbating an already formidable challenge. The difficulties are surely similar for musicians, and are compounded by questions of recording authenticity that can only be answered by including provenance information whose schema are yet to be agreed upon.

Contrary to what the section title might suggest, we intend to avoid issues of scholarly choice and artistic taste in favor of reminding the readers and ourselves of the motivations and circumstances of those who will make such choices or will administer content designated for preservation.

Common patterns for certain classes of paper records have parallels in sound recordings. For instance, governmental and international negotiation records are not massively reproduced and not items of commerce (other than, perhaps, in espionage). The situation is similar for original engineering, medical, and scientific data. The portions of those deemed to have durable value get special storage treatment, albeit perhaps not as good or as frequent as might be merited. Any of these have, or could have, audio components that deserve preservation.

Another pattern is apparent for memorabilia of statesmen, scholars, artists, and performers. Such artifacts languish for decades before collectors determine which of many candidates they care to save. For instance, it was not until late in the twentieth century that libraries took an interest in the youthful correspondence of Leonard Bernstein. Until the materials are recognized to be very valuable, they stay tucked away in relatives' and friends' closets, attics, and basements, often in less than optimal environments. Such materials might include young musicians' and speakers' audio that retrospectively are considered valuable—but not before the physical media that carry them start to become unreadable.

Our first instinct might be to focus on technical aspects of audio, dealing with these more or less independently of other modalities. However, practical examples illustrate that sound recordings are usually accompanied by critical related information—text and pictures—describing the recording circumstances, the music and its history, and promoting the recording [22]. Such ancillary information is essential for reasons well beyond what librarians might otherwise require for collection catalogs.

Records are documents accumulated in the course of practical activities. As instruments and by-products of those activities, records constitute a primary and privileged source of evidence about the activities and the actors involved in them. While records are often conceived in terms of textual documents, such as letters and reports, they can take any form. What differentiates records from documentary materials in general is not their form, but their connection to the activities in which they are made and received. If this link is broken, corrupted, or even obscured, the information in the record may be preserved, but the record itself is lost.[12]

Such considerations are the foundation of well-accepted proposals for integrated digital objects (see, for instance, Kahn and Wilensky [24]. Librarians might point out the good reasons why library catalogs are distinct from the collection elements they tabulate, and that these have parallels in digital collections. Indeed, objects held by IBM Digital Library [25] or its replacement, IBM Content Manager, can safely be stored far from the catalog. If standard metadata is held inside digital objects, a library can easily create catalog entries as part of accessioning objects.

3 System Structure

To relate the last decade's work on archiving for broadcasting, we need a model of software layering. For this I draw on personal experience with IBM Content Manager,[13] a product with several thousand enterprise

[10] At least within this letter *bit stream*, *file*, and *digital document* are synonyms. We mostly use *bit stream* because transmission through a single channel is the usual mechanism by which digital documents are shared among people using computer processes as their agents.

[11] See http://www.usps.gov/history/pfact98.htm.

[12] "For example, a map of [Belgrade] is a document, but a map of [Belgrade] known to have been used in making a targeting decision that led to the bombing of the Chinese embassy is an essential record of that action. The key difference between the document and the record is the specification of the context of action in which the record was involved. To preserve authentic records entails preserving the documents themselves and also their connections to the activities in which they were used" [23].

[13] Registered trademark of IBM Corporation.

customers and a software base for more than 50 IBM business partners. A recent project illustrates how Content Manager can be used by broadcast companies. A press release at the National Association of Broadcasters meeting in 2000 April in Las Vegas started, "IBM and Sony Electronics™ today announced the . . . delivery of a newly designed . . . digital asset management system, created to help CNN™ digitize its vast videotape library . . . making it more accessible to CNN journalists worldwide." The integrated Media Production Suite[14] package combines IBM middleware products (including Content Manager, MQ Series, MQSF, and Websphere) with Sony "to-air" digital video servers and tape libraries and draws on Sony broadcast video integration skills. It is intended to help with many aspects of television product life cycles.

The software that companies like IBM and its business partners product is layered similarly to TCP/IP and other communication protocol stacks. Fig. 1 depicts a possible layering and indicates which layers vendors sometimes split between network servers and client machines that might be end users' workstations. The higher the box in the diagram, the more it tends to be specific to an application class or even to a particular customer.

Between about 1988 and 1993, my work consisted mostly in the design of instances of 4 and 5, and secondarily in consulting with IBM and customer programmers who were implementing instances of 9. IBM Research colleagues worked on a relational database system (IBM DB2, an instance of 3) and a distributed file system for multimedia data (Tiger Shark [26], an instance of 2). Naturally, we discussed the interlayer boundaries in detail. The upper boundaries of each box are exposed as IBM-supported APIs (application programming interfaces) today. For 4 and 5, these are published APIs in IBM Content Manager, and are among the interfaces provided for programs written by IBM business partners, software contractors, and customers.

Layering software as suggested by Fig. 1 is undoubtedly more complicated than producing a monolithic product providing similar functionality, but it has compelling engineering and installation advantages that can be managed to achieve large cost savings, and that include the following:

1) As new hardware technology (storage disks and tapes, printers, analog-to-digital converters) becomes available, all that is needed is to extend instances of boxes 3 and 7. No other software needs to be changed and the new technology can be used in the same system as the old, for example, to allow gradual migration of data to more cost-effective storage media.

2) Most of the software is independent of changes in other system components. For instance, much of our 1993 document storage subsystem code is still in use in the most recent versions of the IBM Content Manager.

3) Server computer platform dependencies can be confined to box 2. One example of how this helps is that the IBM Content Manager executes on many different systems, ranging from $2000 personal computers to $5 000 000 multiprocessor mainframes. One way customers have exploited this is to develop their applications on PCs, deferring the acquisition of more expensive facilities until they are almost ready to go into production.

4) Operating system dependencies can be hidden from most of the software. For instance, IBM Content Manager operates identically under various flavors of UNIX, Microsoft Windows, and IBM OS/MVS.

Such software layering has spawned a large industry that did not exist 10 years ago—software integration contracting. How such a structure is exploited is determined by business factors rather than technical issues. Each software provider typically chooses which layers it will produce and which interfaces it will publish. Typical

[14] Registered trademark of IBM Corporation.

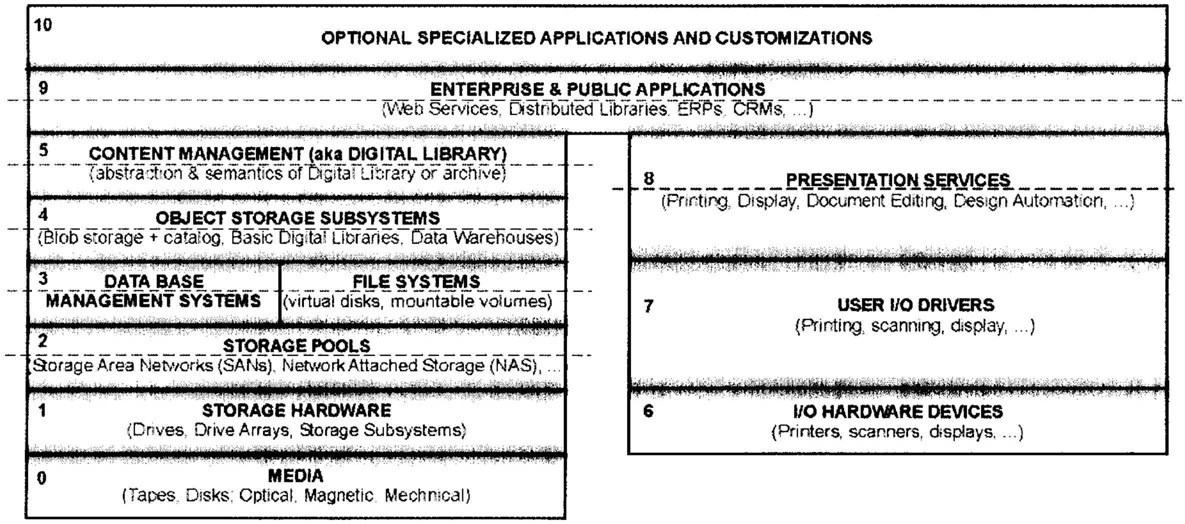
10
OPTIONAL SPECIALIZED APPLICATIONS AND CUSTOMIZATIONS
9
ENTERPRISE & PUBLIC APPLICATIONS
(Web Services, Distributed Libraries, ERPs, CRMs, ...)
5
CONTENT MANAGEMENT (aka DIGITAL LIBRARY)
(abstraction & semantics of Digital Library or archive)
4
OBJECT STORAGE SUBSYSTEMS
(Blob storage + catalog, Basic Digital Libraries, Data Warehouses)
3
DATA BASE MANAGEMENT SYSTEMS
FILE SYSTEMS
(virtual disks, mountable volumes)
2
STORAGE POOLS
Storage Area Networks (SANs), Network Attached Storage (NAS), ...
1
STORAGE HARDWARE
(Drives, Drive Arrays, Storage Subsystems)
0
MEDIA
(Tapes, Disks; Optical, Magnetic, Mechnical)
8
PRESENTATION SERVICES
(Printing, Display, Document Editing, Design Automation, ...)
7
USER I/O DRIVERS
(Printing, scanning, display, ...)
6
I/O HARDWARE DEVICES
(Printers, scanners, displays, ...)

Fig. 1. Layering of multimedia computing systems. Layer boundaries are fuzzy to depict incomplete standardization across vendors. Dashed lines depict optional partitioning of layers into client and server portions that can execute in different networked computers. Each box might have many implementations, even within a software suite from a single vendor, to accommodate different technologies and different applications. Boxes are numbered for easy textual references.

considerations for exposing an interface or keeping it proprietary are the costs of customer education and the implied costs of long-term commitment to the specification.

Some software vendors offer integrated packages, combining the top layers without exposing interfaces. An advantage for a customer who can reliably predict for what purpose and how a service will be used for two decades is that large economies and performance gains are possible in the lower layers. Such vendors typically work closely with a few customers to ensure pleasing products. However, specialization that pleases a few customers tends to make a product unacceptable for most other potential customers and difficult to adapt to changing markets.

The current letter focuses on the document structure that box 4, 5, and 9 instances use to combine audio data with other modalities. One advantage that this confers are large economies in layers 0 to 3, which are made mostly insensitive to the differences between audio data and image, video, and text data.

4 Technical Elements

The "integrated digital object" reasoning alluded to leads to models in which audio is a part of multimedia digital objects. These objects contain metadata describing what is being though necessary and affordable for audio playback. Since the playback purpose and manner cannot be confidently predicted, choosing the metadata schema and contents requires care and professional insight.

> En premier lieu, on considère l'état descriptif de l'enregistrement. Les documents édités qui comportent une documentation d'accompagnement (pochette, livret, etc.) ou un générique (film vidéo) sont généralement bien identifiés. Par contre, les documents inédits (enregistrements de terrain, fonds privés, etc.) restent décrits de manière souvent sommaire et aléatoire. L'insuffisance de description de l'enregistrement peut compromettre foundamentalement la gestion de celui-ci, et l'impossibilité à décrire peut constituer un premier critère de sélection, voire de maintien ou non dans une collection (Fontaine [8]).[15]

Our users will often start with audio and ancillary information that was generated years before they decided to prepare the document for long-term preservation. They need to consider several steps:

- Preparing the original artifact for optimal conversion from analog to digital format, for reasons, suggested by Hess [27]
- Providing missing metadata
- Converting the content to forms highly likely to be comprehensible in 100 years or more
- Ensuring that the transformed data and metadata survive.

Ensuring that the bit streams survive can itself be partitioned into two steps: maximizing the durability of storage volumes (tapes or disks, each carrying several bit streams) and "active migration" of bit streams. An extensive literature addresses "best practices" for media durability [28]. Although the details are important to custodians of large collections, for small holders they boil down to keeping the media at uniform temperature and humidity comfortable for human beings (a little cooler and a little drier might be a little better), avoiding imposing physical stresses while the media are idle, reading the data from the media only on properly maintained equipment, reading the masters only to produce secondary copies for playback, and checking data correctness by sampling volumes periodically (such as annually).

Cost-effective active migration is under investigation for large archives [19]. Small holders will have to adapt what the large archives do to their own circumstances, and will for the foreseeable future be driven to migrate their holdings mostly by rapid technology evolution. Copying files from one volume to another is already easy. How and when to do it will probably be addressed from time to time in the home computing trade literature. For example, migration from magnetic tape media to CD media is just now becoming inexpensive and timely. Table 1 summarizes the potential causes of loss and their remedies.

To some readers, the many details implied may seem discouraging. They should recall that significant positive aspects will emerge as by-products of the conversion of many different modalities, formats, and media to a standard, comprehensive, digital preservation format. Future individual users will not be faced, for each modality, with a multiplicity of storage formats to convert to whatever current playback format they require. Archives and libraries will not have to understand and provide for different storage and preservation resources for many different kinds of media. Instead, each will need to understand and manage a single storage and access facility. This simplification will more than offset the costs of applying information technology.

5 Metadata and Standards

Reporting the findings of a governmental records inquiry, reference [29, §2.3.1] recommends XML-based representation of documents, metadata, and access control rules, and that standard data structures and archive systems be developed according to the following criteria, which we extend somewhat.

Completeness: Digital archive systems should accommodate all potential kinds of documents, including at least multimedia documents, databases, and engineering design documents.

[15] The first thing to consider is the description of the record. Published documents that are accompanied by further literature (on the cover or a brochure or libretto), or credits (in the case of video film) are generally clearly identified. In contrast, unpublished documents (land registration, private trusts, etc.) often are only described in a summary and superficial way. Insufficient description of a record can fundamentally compromise its management and, if it is impossible to describe, can constitute an initial criterion for its selection or not into a collection.

Integration: Each individually archived object should encapsulate all pertinent documents, context, and authentication, avoiding dispersion of the record across multiple systems, that is, the representations of various kinds of data should be integrated.

Hierarchy and Instance Extensibility: The data structure should support data and metadata layers (onion model), and the structure of instances should allow any number of new components to be added.

Schema Extensibility: Both the metadata and the content schema should be extensible.

Authenticity: Every electronic record should be digitally signed to prove who created it, when it was created, and that it has not been modified since creation. If components are independently meaningful or independently created, they should also be independently signed. Furthermore, if it is important that several otherwise independent records were created by the same contributor and intended by the author to stand in specific relationships to each other, linking records should be signed and stored so that the entire complex of related items is reliably intact permanently.

Minimal Standardization: For each kind of record a minimum obligatory metadata set should be defined, and these should reflect that many varied business needs and record types will occur.

Playback Reproducibility: Metadata should include any recording parameter values that affect playback quality or comprehensibility.

Manageability: An archive should:

- Enable recovery and recreation of indexes from the stored documents
- Contain all needed information about authorization to inspect and alter any item(s) in the archive, including the authorization records themselves
- Separate policy data (such as access control, disposal schedules) from immutable records
- Separate indexing records and archiving records to allow specialized indexing and whatever browsing and searching aids might be desired, doing so in ways that permit future support for unanticipated patterns of information discovery.

Led by Australian work [11] several independently started discussions are progressing toward de facto or de jure standards for the interchange of multimedia data. The Open Archival Information System [30] is a high-level reference model developed by space agencies in North America, Europe, and Japan. Contrary to naïve opinions, it does not itself specify semantics and syntax sufficiently for automated information interchange. Another inquiry into essential semantics is the W3C RDF thread, which "integrates a variety of web-based metadata activities including sitemaps, content ratings, stream channel definitions, search engine data collection (web crawling), digital library collections, and distributed authoring, using XML as an interchange syntax" [31]. RDF is part of the basis for the current ISO/IEC proposal for a digital video and audio interchange standard [32]. The Advanced Authoring Format (AAF) Association [33], an industry consortium, is rapidly developing a vendor-independent, open-source media file interchange format for multimedia records. Finally, somewhat earlier than most of this standards-track activity, the Library of Congress mounted a prototype and pilot project of significant interest [34].

I believe that within five years we will have a broadly useful multimedia document interchange standard that uses XML for its packaging and structuring syntax. Since the current standards proposals do not explicitly

Table 1. Summary of technical challenges and remedies.

Possible Causes of Bit Stream Loss	How to Mitigate Risk
Short media lifetime	Media lifetimes can be improved by prudent storage and handling, as described in many writings, for example [28].
Finite media lifetime	"Active migration," that is, copy each bit stream to a new substratum before the loss probability becomes significant.
Finite device lifetime (maintaining obsolete technology will not work)	Market dynamics lead industry to replace devices with better alternatives. The characteristic time for this is shorter than that for media deterioration. "Active migration" is the only practical solution.
Suppliers' guarantees of lifetime	Suppliers mostly do not guarantee long technology lifetimes. Given our target of 100-year archiving, such guarantees would probably not be helpful even if given.
Interpretability of bit streams	This is the only incompletely solved technical problem today. See two sections forward.
Natural disaster	Earthquakes, fires, and floods threaten entire collections. Digital collections should be replicated in multiple sites, or in relatively safe sites. See the extensive literature on storage disaster recovery.
Government misbehavior	Sometimes governments deliberately destroy material in response to doubtful ideological fashions, for example, Nazi book burnings in the 1930s. The only remedy is to store replicas in libraries in other countries.
Destruction in war*	Of course the best measure is to avoid war. Failing that we should replicate data at risk in "iron mountain" facilities and in other nations.

* Twice in the last century the collections in Liege, Belgium, have been burned. More recently we have seen the Sarajevo library burned as an act of "ethnic cleansing."

mention long-term archiving, we need to investigate whether they are already adequate for it or need to be extended. Among other things, this inquiry will need to consider the semantics of multimedia metadata, and also the representations of sound and images.

6 Interpreting the Bit Streams in 100+ Years

The point at which we have arrived has audio data about to become bit stream components of multimedia documents expressed with XML syntax and with metadata components whose semantics and syntax will be internationally agreed. Let us assume that each audio recording is a single time sequence, deferring any question of synchronizing independent channels from a single performance. We must still consider the format of individual bit streams.

Presumably members of the Audio Engineering Society will be keenly interested that good choices are being made for audio. The schema and formats of other kinds of bit streams will be decided by other groups. Arguably the AES should be prominent in influencing proposed audio format and metadata standards. If so, it needs to inform itself promptly to preemption by other deliberative bodies.

Before addressing audio formats, let us consider more generally what needs to be saved. We have seen that archival documents will be represented as integrated objects whose components are metadata and bit streams. The components will be set off from each other by punctuation defined by standard XML schema. The metadata will be represented by ASCII character representations of finite set members. Some metadata will identify the data type of each bit stream, and other metadata will convey attribute values whose domains and meanings will be different for different bit stream types.

For each bit stream type, what remains is to define the attribute domains and also the bit stream encoding, which might depend on the attribute values. For instance, audio might require an attribute to indicate the encoding used—.mp3, .wav, .au, .aiff, or some other format. Thus it is not necessary to settle on a single encoding rule for audio, and we will probably choose to have a number of parameterized rules to support different optimizations.

Before deciding on such matters, we should consider another general issue-choosing between standard representations and programmatic representations. To illustrate this issue, consider a very simple case—a single frame of black and white television—and much more complex cases—programs such as computer games, editors, and even entire operating systems. A raster image file like the television frame is sufficiently orderly to permit its schema description in terms that are demonstrably precise and complete and also almost surely comprehensible a century from now. Moreover, this schema can be written to cover many billions of other raster images, so that the effort for the single frame can be amortized over a large body of information. In contrast, the description of a computer program is approximately as complex as the program itself, and the notion of schema is not useful for programs because even two programs that are mostly the same can have radically different behaviors in execution. Other kinds of files are of a complexity intermediate between those of raster images and computer programs. Although we believe it practical to save audio bit streams as instances of completely specified schema rather than along the lines needed for saving computer programs, no one has considered this question carefully as far as we know.

For computer programs, Rothenberg [35] proposes saving executable versions (the outputs of compiling and link editing), emulating today's target machines (such as, the Intel 486 architecture) on each "interesting" future computer, and executing the saved program by means of the emulator.

We believe this approach is impractical. For instance, it means writing an emulator for each "interesting" current machine to run on each "interesting" target machine. Let us assume that the current machine's description is complete and accurate, a far from trivial assumption, and that someone can afford to write the emulator. How is the correctness of that emulator to be proven? How is the future user to be sure that the saved program behaves exactly in emulation on the target machine as it did running on today's machine? The emulator cannot be written until the future machine exists. By then it will be impossible to compare how the saved program emulation runs on the new machine with how it ran on today's machine.

Lorie [36] avoids such problems in a subtly, but essentially different proposal: defining a "universal virtual computer" (UVC) that is Turing-complete, that is, sufficient to execute any program that works on a stored program computer. The UVC needs never be realized in hardware, but any Turing-complete real machine could be programmed to emulate it. Furthermore, it is possible to specify a UVC that is sufficiently simple that its description using one of several mathematical languages can be demonstrated to be absolutely correct and complete. It is further possible to use the mathematical language so that the UVC definition will surely be comprehensible in the future, even though its user cannot ask any question of its architect.

With UVC defined, today's user would translate the program of interest to be a UVC program (using well-known techniques of translating programs, such as compilation) that is archived. The user would also create a UVC emulator on today's real machine—the same machine as is used for executing the conventionally compiled target program. UVC target program executions could then be compared with conventional target program executions to validate the correctness of the UVC compiler.

The future user would build a UVC emulator for his target machine, and run the saved program on that machine. Presuming that the two UVC emulators—one on today's real machine and one on the future real machine—are correct, then the UVC target program will behave identically today and in the future. All the necessary testing can be done when the needed real machines are available.

Although Lorie's reasoning seems correct and practical, until it is shown in realistic pilot installations, we cannot persuasively assert that the problem is fully solved. So we have three tasks ahead of us for the audio portion of multimedia documents: 1) demonstrating convincingly that the UVC idea is correct and practical by defining a good UVC and showing that it correctly handles varied exemplary documents; 2) deciding whether audio preservation is tractable by the standards approach or deserves the UVC approach; and 3) demonstrating satisfactory archiving of audio, including all applicable metadata.

7 Conclusions

Audio recordings usually occur together with other data types. Audio applications occur in varied institutional contexts for which flexible software systems and storage management help toward managing cost and adjusting to constantly changing business circumstances. Many institutions require easy document interchange with other organizations and individuals whose identities cannot be predicted. These and other circumstances favor digital libraries whose contents are mostly packaged following durable, widely accepted standards.

Furthermore, the dramatic and continuing decline of digital technology prices is encouraging institutions of widely varying size and funding levels to consider using digital library technology to manage multimedia collections, including audio data. The ensuing technical attention has been focused on rapid, inexpensive document interchange, paying little attention to preserving documents for decades to centuries. This letter has examined the latter need, and sketched a program for meeting it. Although its emphasis has been on audio data, it teaches that the incompletely solved challenges are in related topics.

Saving the bits will be accomplished by periodically copying every file onto a new substrate. Creating and organizing metadata is well understood by research librarians and archivists. For digital records, standard XML will be used unless a better alternative appears. A practical multimedia archiving environment is being prototyped [19], but not carried to commercial software as of this writing. We even know in principle how to assure that the bits have meaning in some distant future [36]. However, the greatest challenges to implementations are not technical, but rather economic and political.

Knowing how to solve the technical problems is not enough. One has to do it with the expectation that doing so will expose challenges of reduction to practice, definition of quality measures and assurance procedures, and training of staffs. Four great libraries, the British Library, the Royal Dutch Library, the National Archives and Records Administration, and the Library of Congress, have recently made institutional commitments to address digital archiving; government libraries in Australia seem to be ahead [11], [29]. Notwithstanding such reasons for optimism, we must still assert that, with only niche exceptions, nowhere in the world is a public digital preservation program growing at a breadth, pace, and scale commensurate with the Internet.

This much is accepted in many professional communities, but has not yet been given much attention by the Audio Engineering Society. The best results require collaboration across disciplines that traditionally have had little conversation, much less collaboration—librarians, archivists, computer scientists, IT professionals, and perhaps even audio engineers. The AES should determine its proper role and participate.

8 Acknowledgment

Implicitly this report owes much to what I learned from electronic mail exchanged between 2000 December and 2001 February among the members of the then newly constituted AES Technical Committee for Archiving, Restoration, and Digital Libraries (TC-ARDL).

Our beginning inquiry into prospective standards for multimedia metadata and XML packaging was carried out by the following computer science students of Professor Edward Fox at the Virginia Polytechnic Institute and State University: Solomon Gifford, Mark Hawley, Vinoth Jagannathn, and Shawn Kolodgie.

I gratefully acknowledge many detailed discussions with Richard L. Hess, a fellow member of the AES TC-ARDL, and with Raymond Lorie, an IBM Research Division colleague. Hess provided extensive and helpful suggestions to improve an early version of this letter from what would have suited an I/T audience to something closer to the expectations of AES members.

9 Bibliography

[1] E. Cohen, "Preservation of Audio," in *Folk Heritage Collections in Crisis* (2001 May); http://www.clir.org/pubs/reports/pub96/preservation.html.

[2] Council on Library and Information Resources, *Folk Heritage Collections in Crisis* (2001 May); http://www.clir.org/pubs/reports/pub96/contents.html.

[3] H. Varian, P. Lyman, J. Dunn, A. Strygin, and K. Swearingen, "How Much Information?"; http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html, 2000.

[4] *The New York Times* (2001 Jan. 15).

[5] J. Garrett, D. Waters, P. Q. C. Andre, H. Besser, N. Elkington, H. M. Gladney, M. Hedstrom, P. B. Hirtle, K. Hunter, R. Kelly, D. Kresh, M. Lesk, M. B. Levering, W. Lougee, C. Lynch, C. Mandel, S. B. Mooney, A. Okerson, J. G. Neal, S. Rosenblatt, and S. Weibel, "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information," called for by the Commission on Preservation and Access and The Research Libraries Group (1996 May); http://www.rlg.org/ArchTF/.

[6] H. M. Gladney, "Digital Intellectual Property: Controversial and International Aspects," *Columbia-VLA J. Law and the Arts*, vol. 24, no. 1, pp. 47–92 (2001); http://home.pacbell.net/hgladney/Columbia.htm.

[7] M. G. Christel, H. D. Wactlar, and A. G. Hauptmann, "Improving Access to Digital Video Archives

through Informedia Technology," presented at the 109th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 48, p. 1107 (2000 Nov.), preprint 5280.

[8] J.-M. Fontaine, "Conservation des documents sonores et audiovisuels," in J.-M. Arnoult et al., *Protection et mise en valeur du patrimoine des bibliothèques: Recommandations techniques* (Direction due livre et de la lecture, Paris, France, 1998), chap. 10; http://www.culture.fr/culture/conservation/fr/preventi/documents/c10.pdf; http://www.culture.fr/culture/conservation/fr/preventi/guide_dll.htm.

[9] M. Hedstrom and S. Montgomery, "Digital Preservation Needs and Requirements in RLG Member Institutions" (1998 Dec.); http://www.rlg.org/preserv/digpres.html.

[10] S. Herla, J. Houpert, and F. Lott, "From Single Carrier Sound Archive to BWF Online Archive—A New Optimized Workstation Concept," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 49, pp. 606–617 (this issue).

[11] National Library of Australia, The PANDORA Project (Preserving and Accessing Networked Documentary Resources of Australia) is a collaborative initiative that aims to develop policies and procedures for the selection, capture, and archiving of Australian electronic publications and the provision of long-term access to them. See http://pandora.nla.gov.au/documents.html, 2000. http://pandora.nla.gov.au/pandora/.

[12] D. Schüller, "Preserving the Facts for the Future, Principles and Practices for the Transfer of Analog Audio Documents into the Digital Domain," *J. Audio Eng. Soc. (Communications)*, vol. 49, pp. 618–621 (this issue).

[13] RLG DigiNews (http://www.rlg.org/preserv/diginews/).

[14] Computer Science and Telecommunications Board of the National Academies, *The Digital Dilemma: Intellectual Property in the Information Age* (National Academy Press, Washington, DC, 2000); http://books.nap.edu/html/digital_dilemma/.

[15] Committee on an Information Technology Strategy for the Library of Congress, Computer Science and Telecommunications Board, National Research Council, *LC21: A Digital Strategy for the Library of Congress* (2000 July); see especially chapter 4, "Preserving a Digital Heritage"; http://books.nap.edu/books/0309071445/html/73.html.

[16] P. Bulger, "Saving America's Aural Legacy," *J. Audio Eng. Soc. (Letters to the Editor)*, vol. 49, pp. 626–627 (this issue).

[17] U.S. Congress, "Making Omnibus Consolidate and Emergency Supplemental Appropriations for Fiscal 2001," Public Law 106-554 (2000 Nov.).

[18] J. W. Carlin, Archivist of the United States, on the Report of the Electronic Records Work Group of NARA (1998); http://www.nara.gov/nara/pressrelease/nr98-148.html.

[19] R. Moore, C. Baru, A. Rajasekar, B. Ludaescher, R. Marciano, M. Wan, W. Schroeder, and A. Gupta, "Collection-Based Persistent Digital Archives," *D-Lib Mag.* (2000 Mar., Apr.); http://www.dlib.org/dlib/march00/moore/03moore-pt1.html.

[20] A. Seeger, "Rights Management," in *Folk Heritage Collections in Crisis* (2001 May); http://www.clir.org/pubs/reports/pub96/access.html.

[21] G. W. George, "Difficult Choices: How Can Scholars Help Save Endangered Research Resources?," CLIR Rep. Pub58 (1995); http://www.clir.org/pubs/abstract/pub58.html.

[22] S. Lyman, "Why Archive Audio Metadata," in *J. Audio Eng. Soc. (Communications)*, vol. 49, pp. 622–625 (this issue).

[23] K. Thibodeau, "Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration," *D-Lib Mag.* (2001 Feb.); http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html.

[24] R. Kahn and R. Wilensky, "Accessing Digital Library Services and Objects: A Frame of Reference," CSTR Project: Architecture of the Digital Library, Draft 4.4 (1995 Feb. 2); http://www.isoc.org/HMP/PAPER/243/html/paper.html.

[25] H. M. Gladney, "A Storage Subsystem for Image and Records Management," *IBM Sys. J.*, vol. 32, no. 3, pp. 512–540 (1993). Today the product version of this software is called IBM Content Manager; see http://www-4.ibm.com/software/data/cm/cmgr/.

[26] R. Haskin and F. Schmuck, "The Tiger Shark File System," in *Proc. IEEE 1996 Spring COMPCON* (Santa Clara, CA, 1996); http://www.almaden.ibm.com/cs/shark/cmpcon96.zip.

[27] R. Hess, "The Jack Mullin/Bill Palmer Tape Restoration Project," *J. Audio Eng. Soc. (Features)*, vol. 49, pp. 671–674 (this issue).

[28] G. Gibson, "Standard and Preservation of A/V Media and Data," *Int. Preservation News*, Newsletter of the IFLA Core Programme for Preservation and Conservation (PAC) (1997 May 14); http://www.ifla.org/VI/4/news/14b-97.htm.

[29] Public Record Office, Victoria, Australia, "Victorian Electronic Records Strategy," Final Report (1998); http://www.prov.vic.gov.au/vers/final.htm.

[30] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," CCSDS 650-0-R-1 Red Book (1999 May); http://ftp.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf.

[31] W3C Committee for RDF, "Semantic Web Activity: Resource Description Framework (RFD)," *World*; http://www.w3.org/RDF/.

[32] Motion Picture Experts Group (ISO/IEC working group in charge of standards development for digital video), "The MPEG Home Page" (2001); http://www.cselt.it/mpeg/.

[33] AAF Association, "Advanced Authoring Format Technical Information" (2000) http://www.aafassociation.org/.

[34] C. Fleischhauer, "Digital Formats for Content Reproductions," Library of Congress (1988 July); http://

lcweb2.loc.gov/ammem/formats.html.

[35] J. Rothenberg, "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation," Report to the Council on Library and Information Resources (1999 Jan.). See also, "Ensuring the Longevity of Digital Documents," *Sci. Am.*, vol. 272, no. 1, pp. 42–47 (1995); http://www.clir.org/pubs/reports/rothenberg/contents.html.

[36] R. Lorie, "Long-Term Archiving of Digital Information," *Proc. First ACM/IEEE-CS Joint Conf. on Digital Libraries*, pp. 346–352 (2001 June 24–28).

H. M. Gladney
HMG Consulting
Saratoga, CA 95070, USA
Email: hgladney@pacbell.net
Website: http://home.pacbell.net/hgladney/

THE AUTHOR

Henry Martin Gladney received a B.A. degree from the University of Toronto in 1960 and M.A. and Ph.D. degrees from Princeton University in 1961 and 1963, respectively. He was a research staff member in the IBM Almaden Research Center from 1963 until 2000, when he resigned to offer independent consulting services in digital asset management. He is a former member of the American Chemical Society, a member of the Association for Computing Machinery, and a Fellow of the American Physical Society. Since 1987 Dr. Gladney has worked on technical and business issues associated with digital libraries and digital security management. He is the author of 60 refereed papers and 10 patents (4 pending).