Conservation-quality digital media systems and distributed archives

Peter Cornwell, Joyce Shintani

CONSERVATION-QUALITY DIGITAL MEDIA SYSTEMS AND DISTRIBUTED ARCHIVES

Peter Cornwell
ZKM | Center for Art and Media
Center for Digital Conservation
Lorenzstr. 19
76135 Karlsruhe, Germany

Joyce Shintani
Shintani Multimedia
Nauener Str. 25
70597 Stuttgart, Germany

ABSTRACT

The explosion of interest in digital asset retrieval systems, including Musical Information Retrieval (MIR), combined with the convergence of fast, low-cost network technology and large storage components led the publicly funded Center for Art and Media (ZKM) in Karlsruhe, Germany to begin an investigation into the implications of these technologies on archiving, conservation and access in 2002. Today, ZKM is creating an international network of scalable, open system-based archives, powerful enough to manage heritage collections. The core is a disk storage system, operable by non-specialist personnel in the museum or archive environment, suitable for music, moving imagery, film and other media. The project is called the "Digital Heritage Program" and is housed in the ZKM Center for Digital Conservation. In 2004, the ZKM Storage Area Network (SAN) reached approximately 100Tb (Terabyte) capacity and will eventually house the audio and moving imagery collection of the ZKM | Media Library, as well as mirroring several other collections. The presentation presents specific experiences from the project to date and also reports on two developments slated for roll-out in 2005: the pilot implementation of "institutional guarantee" as well as public-private partnership.

1. INTRODUCTION

The limited lifetime and degradation in quality with each play out that is inherent in magnetic tape and film media has led to a long-term crisis in management of media collections, both commercially and institutionally. It is true that digital technologies have precipitated great advances in compression equipment. However, the compression techniques, no matter how good, result in significant loss of information compared with the original and cannot be used for conservation. Conservation archives require more than ten times the storage space of domestic compressed data. Such systems enable the creation at any time of tape copies which, when played on original equipment, produces results indistinguishable from the original at the time that it was digitized.

The development of a strategy to bridge current limitations of technology to create sustainable storage architecture that can evolve efficiently and minimize the complete loss of important material in the short term is one of the most important cultural goals of the decade. Of equal concern are the cost of ownership – capital and consumable equipment, operating cost, maintenance contracts, personnel, facilities – and the security of irreplaceable titles.

The Center for Art and Media (ZKM) in Karlsruhe, Germany, a publicly-funded center for media art and technology including electronic music, began an investigation into the implications of fast, low-cost network technology and large storage components on archiving, conservation, and access in 2002. In the following presentation, we present specific experiences from the "Digital Heritage Program", which is housed in the ZKM Center for Digital Conservation. In addition, we report on two developments slated for roll-out in 2005: the pilot implementation of "institutional guarantee", as well as public-private partnership.

The underlying approach taken in this work is simple: We have concentrated on the needs of digital media assets conservation by developing capture, secure storage, and dissemination methods for the very highest quality data that can be obtained from analogue masters – a "loss-less" digitization approach, to be described below.

The fundamental property of digital representation – that data-sets can be cloned to produce identical multiples – now enables the intrinsic limitation of access time of traditional media to be overcome. Loss-less digitization and copying of valuable collections on multiple disk systems at different locations within an installation offer a level of fault-tolerance equivalent to the reliability of existing tape media. Automatically maintaining identical copies of one installation's titles at another geographic location – a technique known as mirroring – gives unprecedented security. The breakthrough in this approach is that the fast, geographic cloning enabled by new communication and storage technology overcomes the lifetime limitation of individual physical media.

Before entering into details of the Digital Heritage Program, we will first describe the institution at which the archive architecture was developed.

2. BACKGROUND

The ZKM | Center for Art and Media in Karlsruhe is a unique institutional model which, responding to rapid developments in information technology and today's changing social structures, provides a forum for e-

search, production, presentation, and discourse in the fields of art, science, and media technology. The ZKM comprises the Media Museum, the Media Library, the Exhibitions Department, the Museum for Contemporary Art, and five research institutes. The different collections are made public via temporary art exhibitions in diverse in-house exhibition spaces, as well as via traveling exhibitions, conferences, and a comprehensive internet presence.

The ZKM | Media Museum has a thematic spectrum extending from interactive film to cyberspace simulation technology and current internet software applications. The main task of those working at the Media Museum is the storage, restoration, and presentation of media art.

The ZKM | Media Library comprises audio and video collections and a library. Occupying floor space amounting to some 600 sq.m., the entire Media Library offers a comprehensive collection of international video art, contemporary music, and literature relating to 20th century art. A central database allows search requests covering the inventory of about 1,300 digitized art videos, 12,500 audio works with a focus on electroacoustic music, 30,000 books, 120 reviews and 500 CD-ROMs and DVDs. Of special importance is the International Digital Archive for Electro-Acoustic Music (IDEAMA) from the medium's beginnings until 1970. Besides significant modern works and the archive of the Deutsche Gesellschaft fuer elektroakustische Musik (DegeM), the collection also hosts one of the largest soundscape collections worldwide, as well as works awarded the Karl-Sczuka-Award for radio since 1955. The target group consists of visiting academics, users in the ZKM institutes and the State Academy for Art and Design Karlsruhe, as well as the interested public.

Development from these resources towards a storage area network serving the whole organization is a large-scale task which will be completed in 2005 with the distribution of generic, fault-tolerant disk storage throughout the building. This will enable both load-sharing and mirroring to service very high bandwidth applications, such as the non-linear delivery of filmic resolution imagery into the Media Theatre, as well as a high level of automation of backup for research and administration and security for the Media Library collections. Network security and access protection schemes are of the highest importance in these plans.

This is the scenario that research at ZKM has addressed. A ‘farm’ of inexpensive digitization workstations with film as well as new and historic video tape transports has been developed, which has multiple very fast connections to large-scale storage, in a facility that can operate 24 hours a day.

3. SYSTEM DESCRIPTION

In examining the last sixteen years experience of development and maintenance of the ZKM | Media Library, the evolution of technology as well as archive methodologies have been crucial to the formulation of a sustainable architecture for a new generation of digital archives. Essential aspects of this archive architecture are the following:

3.1. Disk-Based Archiving

The technical feasibility of building huge disk archives has been demonstrated by companies such as Google [2], which routinely operates tens of thousands of computers in its caching of internet search keys. Also, experience with the Terastore jukebox at ZKM has seen the 5 meter machine of six years ago overtaken by a few disks costing five hundred dollars and fitting into a shoe box – and delivering much higher performance and functionality at the same time. A further trial was the collaboration with the Institute of Contemporary Art (ICA) in London and U.K. company Street Vision (Peter Cornwell, CEO) in 2002. In it, loss-less digitization techniques were developed for video which have since resulted in a Linux-based, very high resolution workstation for video and audio. The project digitized a sample of the ICA's collection of historical lectures from the last 20 years and proved that secure disk-based archives are now robust enough for deployment in a museum service.

A disk has a life expectancy in continuous service that is lower than that of a magnetic tape, although an analogue tape loses quality with each access, and no tape can be accessed continuously. On the other hand, a disk is a non-contact device, and its contents do not degenerate with access. If disks are "spun down" when not in use, their life expectancy can be of the same order as that of tape archives, and their content can be cloned in a matter of days. Apart from the risk to content, the huge cost in human intervention of copying a tape collection is prohibitive. Cloning a media collection on disk also results in "regeneration". The same is true of digital tape, however copying analogue tape or film results in loss of information.

Moreover, there are side effects of disk-based archive methodologies that actively contribute to the potential accessibility of collections. Firstly, the geographic mirroring of archives enables collection sharing, and it ensures survivability in the face of disaster at one or even several sites. This is simply not possible with large collections based on tape-based media due to transcription time. Materials in a disk-based archive can also be protected by access permission or even by encryption on disk, individually, so that copyright protection can be administered on a title-by-title basis.

Another side effect of direct-access, disk-based archives is background computation. Computation to search and access titles for users and to manage the consistency of geographical mirrors does not exhaust the capacity of this architecture. Either in "its spare time" or by scheduling through load-balancing, the storage network can also accomplish other computationally demanding tasks, such as the transformation of titles into

new format standards or the production of a variety of compressed formats.

3.2. Open System Architecture

Open systems are customarily defined as those systems that can be supplied by hardware components from multiple vendors and whose software can be operated from different platforms. They are opposite to closed or proprietary systems.

Together with the Internet Engineering Task Force's Storage Area Network standards (www.ietf.org) – the organization that sets Internet standards [3] – the impact of open systems on the sustainability of practical disk-based media archives is far reaching, because it insulates installations from dependence upon individual computer vendors over the very long lifetimes significant for archives. Consequently, the ZKM project has concentrated on an open system strategy. Important also, has been the establishment of strategic collaborations with other institutions and corporations for development of mirrors of collections and of new standards.

Storage equipment must support open and widely adopted data communications and file store standards so that, even if standards evolve, it is possible to migrate collections from old representations to new without loss of title information. Interoperation of products from different vendors as part of a storage area network is essential, as the long-term security of a collection cannot assume the long term survival of any single computer company. Arising from this issue is the need to perpetually upgrade installations into the distant future, rather than to discard machines at a definite end-of-life. In this area, the project has conducted specific trials with open systems.

3.3. Open Source Software

When the source code of a computer program is made available free of charge to the general public, it is known as open source. The aim of open source software is to produce more useful and bug-free products for general use. The concept relies on peer review to find and eliminate bugs in the program code, a process which commercially developed and packaged programs do not utilize. Since the use of open source software has the potential to significantly reduce costs of purchase, operation, maintenance, and enhancement, it was the choice for the ZKM project. It has demonstrated reliability and maintainability of large systems over very long periods, through widespread technical review and the ability to fix code defects. In addition, software can be customized to specific project needs, which is unachievable using the products of a single company.

4. SUSTAINABLE COMPUTING

One of the key lessons from the experiences of the last sixteen years is that new digital media archives must be sustainable. Basically, sustainability refers to doing something with the long term in mind – it is the concept of meeting the needs of the present without compromising the ability of future generations (of computing) to do the same. Today's decisions are made with a consideration of sustaining activities into the long term future. Thus, “sustainable computing” should endure, adapt to changing conditions, and also be affordable over time.

Condensing ZKM's experience, it seems that just a few issues fundamentally affect sustainability. They also reveal potential solutions to controlling the cost of ownership over extended periods, first in terms of equipment, and second in terms of retaining generations of personnel able to handle installations and title masters reliably. The following issues contribute to the quality of "sustainability":

4.1. "Future-proof"

Research at the ZKM during the last three years has produced computer architecture for archiving media art that is "future-proof". Unlike conventional computer products, it will not get outdated and require replacement after three or five or even ten years; rather it can be progressively upgraded with widely available components and maintained by institutional personnel. This has been achieved using open systems technology, as described above, which makes it possible for the first time to manage storage of the scale of hundreds of Tbs (Terabytes) economically and safely. This is assured through the use of loss-less digitization. By "loss-less", we mean any digitization method, which, in the case of moving imagery for example, records all of the information available in the original and stores it using a digital representation that preserves that information frame by frame. The use of open systems means that the archive will continue to grow in capacity at the rate of the state-of-the-art. It also means that the archive's life is constantly extended, and its reliability even improves. Most significantly, digital collections of the highest conservation quality remain completely portable, enabling huge volumes of data rapidly to be transferred and converted to future standards as they emerge.

4.2. Cost of Ownership

Open systems potentially cost much less to operate over extended periods than proprietary systems. However, their contribution to the institutional guarantee of digital conservation archives lies in their openness – not being locked into vendors and proprietary formats and the problems of maintaining them successfully over long periods. An important aspect for archive managers is the ability to select freely from a large competitive market of network storage equipment and qualified personnel, over a very long operational horizon.

A multi-vendor market is essential because of the intrinsic tendency of vendors to lock clients in to their product range to the exclusion of lower-priced competi-

tors. Today, the realization is growing that even the largest computer and software companies will not remain stable providers of the products that archives require for the extended lifetimes now being considered.

Confidence that personnel can be attracted to work on such installations, especially in an institutional context with salary levels lower than in industry, is an important factor in managing operational costs. In addition, the risk of losing control of the installation through the loss of critical people must be managed. 24/365 running, maintenance, and a perpetual upgrade program necessitates the retention of personnel with strong open systems skills, and continual research is essential for institutions to be able to reason effectively about industry and institutional trends. This in turn requires that some level of academic and engineering skill-base be maintained, which is easier to assure using an open system like Linux.

5. FORMAT ACCESSIBILITY

The diverse composition of ZKM Media Library users requires flexible means of access: as a service institution for the scientific staff and the exhibition team, as a place for research and teaching for the State Academy for Art and Design, and as an educational center on media art for the interested public. Thus, the accessibility of collections using widely available desk and internet-based and mobile equipment is of great concern. In particular, the rapidly changing formats and performance of both browsing interfaces and communication channels must be addressed if media collections are to remain widely and efficiently accessible. Consequently, another research project running at ZKM concerns the archive infrastructure: the transformation of master datasets into compressed formats suitable for delivery to a multiplicity of viewing environments (archive client system). In a service trial using wireless LAN and handheld personal digital assistants (PDAs), ZKM | Media Library titles and versions of collections formatted for different access methods are produced automatically or even "on-demand". This is particularly important as internet performance improves and users expect increasing quality of compressed materials to be delivered from web sites.

Due to the limited transfer rates of wireless technology and the limited processing available on mobile devices, the compression rate for digitization has to be carefully chosen. Wireless data rates are faster than most domestic internet connections, and so yet another set of constraints must be addressed. All of this contrasts with general archive principles concerning preservation, in which it is important to digitize material without compression to the extent that storage space is available. Eventually, a multiplicity of different digitized versions will be necessary in order to meet the demands of combinations of different communication channels and client viewing systems. The equation is complicated still further by the fact that specific clients only run specific application software. Thus, the archive interface must not only recognize the compression level appropriate, but also determine the permissible client application interface.

6. INSTITUTIONAL GUARANTEE

While technical and financial solutions now have been elaborated for mastering most of the challenges facing digital hard-disk archiving, many institutions are still reticent to commit to an implementation. Hart and Liu [1] have found a lack of trust in digital preservation to be one of the main impediments. They apply a concept derived from their analysis of monetary currency called "institutional guarantee" to the development of trusted systems for the preservation of digital information. They find that the creation of this guarantee is instrumental in increasing people's confidence and trust in digital media, since there are no precursors for the preservation of such documents.

The necessity for several independent and authoritative organizations to participate in the adoption of a technology before it is perceived to have gained a critical mass is essential to precipitate its widespread uptake by the community. Historically, examples of institutional guarantee include the introduction of symbolic financial instruments, such as paper banknotes. Underlying distributed computer networks is the fact that multiple distributed copies of important media holdings are evidently more secure than two or even three conventional tape archives. Because identical master copies of titles can be maintained securely on a geographically distributed computer network, individual installations might be attacked without impact upon the safety of actual collections. Security of title masters is guaranteed through the automated maintenance of archive mirrors at several locations, and in turn, institutions and curators of exhibitions gain new possibilities of collection sharing.

In a first demonstration of "institutional guarantee", two major institutions have joined forces utilizing ZKM's Digital Heritage system. By mid-2005, the Netherlands Media Art Institute, Montevideo/Time Based Arts and the Science Museum London will share cloned copies of parts of their collections.

7. PUBLIC-PRIVATE PARTNERSHIP

In recent years, the term "public-private partnership" has been used increasingly, with definitions varying from out-sourcing to outright privatization. Strictly speaking, a public-private partnership is an "agreement between a public agency and a for-profit corporation. Through this agreement, the skills and assets of each sector (public and private) are shared in delivering a service or facility for the use of the general public. In addition to sharing resources, each party shares in the risks and rewards potential in the delivery of the service and/or facility" [4]. Advantages of such partnerships include cost-reducing solutions that maintain the same or better levels of qual-

ity, and the successful leveraging of increasingly limited government resources to complete high-cost, high need projects. More importantly, the "application of sound business practices can provide substantial increases in productivity and efficiency", or enable the installation of projects, that would not have been possible in the context of a public institution. This is of particular concern in Germany, as Secretary of Commerce Wolfgang Clement often emphasizes. While Germany excels at inventing and holds a substantial number of patents, commercial exploitation often lags behind. This was, for example, the case of the mp3 format developed at the Fraunhofer Institute, which was exploited not in Germany, but in the USA by start-ups such as mp3.com, etc.

Phases of the Digital Heritage Program have already benefited from public-private partnership, notably between ZKM and Street Vision, as mentioned above. A further step in the knowledge transfer between public institution and private economic sector will be initiated in 2005. With the founding of a start-up in the state of Baden-Württemberg, some of the Digital Heritage Program technology and services will be offered through a private company. The results of this cooperation will the reported in our next paper.

8. REFERENCES

[1] Hart, P. and Ziming Liu. "Trust in the Preservation of Digital Information", *Communications of the ACM*, June 2003.

[2] Shread, P. (2002). "Google Experiments with Distributed Computing", *ITManagement*, http://itmanagement.earthweb.com/netsys/article.php/999551, posted 27-March-2002.

[3] Simpson, D. "iSCSI gains a toehold in SAN market", *i-tech. Intelligent Storage Solutions*, April 2003.

[4] "Using Public-Private Partnerships to Meet America's Essential Needs", *The National Council for Public-Private Partnerships*, Washington DC, 2002.