Preserving Digital Media

Mary Baker and Mehul Shah
HP Labs, Palo Alto
mgbaker@hp.com, mehul.shah@hp.com

Abstract

The motion picture and broadcast television industries are in the midst of an unprecedented transformation from physical to virtual assets. While this conversion to digital workflows and end product has many advantages, one topic remains a puzzle: how do we preserve digital media assets with at least the same success as we have preserved physical, analog assets? In this paper we explain why digital assets cannot be preserved simply by storing them in current digital storage systems; we describe why physical curatorial processes do not work well for preserving digital assets; and we cover good practices, as we currently understand them, for preserving digital information.

1. Introduction

Digital preservation means the ability to access and use high-integrity digital assets many decades after they are stored. All three aspects are important: the access, the usability and the integrity. If the materials are stored but we cannot access them, they are useless. If the materials are accessible but no longer in a format we can use or understand, we might as well not have stored them. And if the materials are accessible and in a reasonable format but are significantly damaged, we still may not be able to use them. It is even worse if their integrity has been maliciously undermined and we do not detect it; in that case we may be in a lot of trouble, potentially releasing illegal or inappropriate material.

To preserve digital assets (rather than analog copies of the digital assets, which is a different problem), we need reliable long-term storage where the material stored remains accessible. At first glance we have two paths to follow. We could purchase an existing highly reliable digital storage solution, or we could use existing physical curatorial practices and apply them to the digital media.

Unfortunately, neither of these paths leads to a great solution. Most existing reliable storage systems are designed with other goals than longevity in mind and do not address the threats peculiar to long-term digital assets. Techniques for curation of physical assets also fail to handle some of the threats peculiar to digital rather than analog materials.

In the next sections we describe these threats and explain why neither current storage systems nor our traditional physical preservation processes are a complete solution for digital preservation. We then describe our current understanding of good practices for archival storage of digital assets and conclude with a description of open problems in the area of digital preservation.

2. Long-term versus Short-term Threats

Traditional enterprise storage systems are designed to provide very high availability and performance. Availability means that the system should almost never go down, and if it does, it should be available again in a matter of seconds. High performance means the system should be fast at handling a challenging workload. For instance, an enterprise storage system might have to handle a high rate of low-latency updates to an e-commerce database. Such systems are not easy to design and they can cost a lot per byte of storage.

In contrast, a digital preservation system often does not have to exhibit very high availability or performance. For many repositories it is okay if the system is down sometimes. It needs to come back up in a reasonable amount of time, and it must not have lost any data when it does so, but we can tolerate short service interruptions. A digital preservation system may also have a more relaxed workload to handle. Depending on the application, the end-user workload may consist mostly of reading stored materials, not updating them, and ingestion of new materials to the stored collections can often be scheduled so that it does not require extreme performance engineering. These differences provide an opportunity to reduce the overall engineering cost of a digital preservation system.

These advantages do not mean digital preservation is without its challenges. Digital preservation must address a set of threats only a few of which are considered by short-term storage systems. Here we categorize storage threats according to how specific they are to long-term storage. Some of the threats are peculiar to long-term storage: hardware and media obsolescence, software and format obsolescence, and loss of context with which to make sense of the stored data. Some of the threats affect both long-term and short-term storage but are harder to address in long-term storage. These include attack, economic problems, external dependencies, and organizational problems. Finally, some of the threats affect both long-term and short-term storage, but the longer the desired lifetime of a digital asset, the more likely the threat will be realized during the lifetime of the asset. These include large-scale disaster, human error, component faults, and storage media damage. Here is more detail about these problems, starting with threats peculiar to long-term storage.

Hardware/media obsolescence: Computer systems continue to evolve. Any storage hardware or media we choose now will later be replaced by something faster, denser, cheaper or easier to use and manage. Nine-track tape and 12-inch video laser discs are examples of storage media that are now largely unusable due to the lack of hardware with which to access the media. This means that the answer to digital preservation is not an exotic storage medium! We are more likely to be able to access common commodity storage media for longer, since more readers will be available.

Software/format obsolescence: In a similar manner, applications, data structures and storage formats continue to evolve, and new applications are not always compatible with old data formats. This can happen quite rapidly. For instance, photographers have already had to tackle obsolescence of various RAW formats for digital cameras [1]. Commonly-used formats are less vulnerable, as are formats based on open standards

that are well-documented so that the rudiments of the applications to interpret them can be recreated if necessary.

Loss of context: Sometimes information about the digital assets is needed to make full use of them. Examples of such information include decryption keys for encrypted assets, format information, or even information about which applications must be run in which order using the assets as input. Loss of this context can reduce the value of the assets or render them useless. We need to pay as much attention to preserving context as to preserving the data itself.

The following threats are not limited to long-term storage, but they are harder to address over long time periods.

Attack: We generally worry about securing systems from short-term intense attacks that render the system unusable or that destroy, modify or steal data. Over long time periods, though, we need to worry about slowly subversive attacks that gradually damage the data, reduce the effectiveness of the system, or cause it to misbehave. Traditional repositories such as libraries have been subject to such attacks, and online systems are also likely to be targeted, especially through platform attacks. If a security hole exists in a particular software platform (a particular operating system and its corresponding software tools), then any system with that platform may be breached. Replicating assets on systems with different platforms provides some protection against some of these attacks.

Economic problems: It is generally easier for an organization to justify using its funds to create new assets such as new movies than it is to budget for saving old assets. Further, it's hard to know how much money is required to save any particular asset, since we do not have adequate tools for understanding the long-term costs of preservation. Over many years, ongoing costs for power, cooling, rental space, system administration and hardware renewal can add up to much more than the original outlay cost for the system. We need to reduce the overall price/byte/year and not just the initial price/byte of these systems, and we need better long-term cost models.

External dependencies: There are external dependencies in systems that may fail over long time periods. In twenty years, particular certificate authorities and 3rd-party license servers may no longer exist. Embedded URLs and domain names may fail to resolve. Commodity storage solutions often do not consider these dependencies as threats over the short-term, but long-term systems need to be designed with an understanding of their external dependencies.

Organizational problems: Collections of digital assets are often most vulnerable when their sponsoring organizations or caretakers go through changes such as mergers, moves, bankruptcy, or ownership changes. Digital assets are more invisible than physical ones and can fall through the cracks during these changes. Organizations can also make mistakes regarding the care of digital assets, so reliance on single services is dangerous. We can avoid some of these problems if we can avoid dependence on a single service or organization. Putting in place "data exit strategies" at the time collections are formed, and ensuring they continue to function, helps make it possible to identify and move assets quickly when organizational changes occur.

Finally, the following threats are not peculiar to long-term storage, but the likelihood of their affecting particular digital assets during their lifetimes is higher the longer we want those assets to live.

Large-scale disasters: Floods, fires, earthquakes and acts of war can all destroy entire storage sites. Storing all copies of an important asset in one place leaves it vulnerable to these problems, so replication of assets across geographies is essential.

Human errors: Human mistakes are increasingly the cause of computer system failures [2]. System administrators can accidentally overwrite or delete precious data, or they can purposefully delete data only to find later it has become valuable. Any single storage system is vulnerable to human error. Clearer user interfaces and carefully established processes can help reduce error, but mistakes still happen. Replicating digital assets in systems administered separately can help avoid data loss in the event of an administrator's error.

Component faults: Any component in a storage system may fail, including software, required network services, and hardware such as disks, network interfaces, power supplies and fans.

Media damage: All affordable digital storage media are subject to degradation over time, surviving perhaps 30 years, but sometimes as little as 5 years. Magnetic tape, disks, CDs and other media must be refreshed eventually. Detecting corruption before all copies of the data are damaged is important.

Since many of these threats are not well addressed in short-term storage systems, we need to look to other solutions for long-term preservation. How about traditional curatorial processes for physical assets? We turn to this question in the next section.

3. Physical Curation of Digital Assets

The entertainment industry has long experience with preservation of physical, analog assets. For example, placing black and white separation negatives in climate controlled environments has successfully preserved motion pictures for many decades. The general rule is to find a medium that does not degrade much over time, to store your assets on that medium, and to place the medium out of harm's way. The stored asset should be pulled out of storage only when essential, so as not to damage it. Even if a small amount of degradation occurs, it might be tolerable. While it might reduce the quality of the asset, it does not render it useless. So why can't we just take the same approach with digital materials?

Unfortunately, the same preservation instincts do not help as much with digital materials, for several reasons: sheer amount, new expectations, and many of the threats listed above. Cost-effective capture and creation on digital media has led to huge amounts of material, much of which we might like to save and perhaps repurpose over time. Since any individual copy does not take much space, we do not feel a need to be as selective about what gets saved and we can end up with a lot to save. Digital media has also enabled us to make bit-per-bit accurate copies of materials, leading to the expectation that it will not be hard to preserve these materials as accurately as we can copy them. But threats such as media degradation, hardware/media obsolescence, software/format

obsolescence, and loss of context all interfere with our expectations. Affordable digital storage media degrade over time and often digital formats do not survive even small amounts of degradation. For instance, losing a single byte of information in some compression formats may make it impossible to decompress the material. It can also be harder to maintain the accessibility and usability of digital media. For example, finding a reader for a nine-track tape is now very hard. Even finding an application to interpret an old format can be challenging, and if you have a copy of the application, it might not run on today's hardware and software platforms.

To preserve digital materials in digital form we need to take a different, if unintuitive approach. Instead of avoiding touching the materials, we should regularly monitor or audit them for their integrity and continued usability. If we find problems, we should repair them before it is too late to do so. This means we need to have multiple copies of the assets in good condition from which to make these repairs. Taking this approach leads us to the set of suggested practices described in the next section.

4. Good Rules of Thumb

Addressing the threats we have identified, while exploiting the nature of an archival workload leads us to a basic set of rules: make copies of the assets, monitor and maintain their condition, and attempt to make the copies as independent of each other in the face of failure as possible. Further, do this at low cost. In more detail:

Replicate content and break failure correlations between replicas: To survive large-scale disaster, we can replicate archival content in different geographic locations. We should administer these replicas independently to survive operator (human) errors and we should provide platform independence to survive platform attacks.

Audit replicas proactively to detect damage: At large scale and over long time periods, damage accrues in stored information. Unfortunately, many kinds of damage may be undetected at the time of occurrence, including successful attack, human error, and media damage [3]. This damage must be found and fixed before all replicas of the information have failed or before failures in some replicas are repaired using data from other replicas with undetected corruption. This audit process must be inexpensive and itself cause little wear and tear on the archive. *This implies that data must be easily accessible for audit and that backup to high-latency offline media is not a complete archival solution*.

Automate content migration to maintain usability: Over time content must be migrated to new hardware, new software formats, new compression algorithms, new encryption keys, and so forth. Automating these processes as much as possible helps reduce management costs. Emulation on new hardware and software of old hardware and software platforms for continued use of old formats may also be a possible solution but is not yet broadly available and applicable at low cost.

Avoid external dependencies: The system should be built without dependencies on external facilities that have not proven likely to last for long periods of time, such as 3rd-party license servers. Where dependencies exist they must be clearly understood so that automatic migration away from them is possible. Having a good "data exit" strategy – the ability to move your data off of one system or service and onto another – also

helps reduce dependencies on particular systems or services. In the case where sponsoring organizations or services can suddenly cease to exist, it is best to avoid dependence on a single such organization or service.

Reduce preservation costs: The need for affordability combined with relatively relaxed workload requirements suggests that easily evolvable solutions built from commodity hardware are more likely to succeed over long time periods than specialized hardware solutions. It is important to automate management tasks wherever possible to reduce on-going costs of the archive.

Viewing these practices together, it seems it is impossible to define a single digital archival storage solution that will last indefinitely. Instead, it is important for preservation systems to be highly evolvable and dynamic, so that they can scale and adapt over time. Even the practices suggested above will surely change as we gain more experience.

5. Current Status of Digital Preservation

The bad news is that as of this writing, we know of no commercially available affordable solutions that employ all of the practices described above. Organizations such as the British Library have had to implement their own solutions that match many of these practices [4]. The good news is that organizations such as HP Labs are working on improving digital preservation processes and technologies. As part of this effort there are still several areas of open research to address before we can be completely comfortable with any solution. Here we list some of the open research questions, although there are many others.

Audit strategies: What are the best audit strategies? For instance, in a given system, what is the minimum frequency with which we must check our stored data to reduce the potential for undetected corruption to an acceptable level? Where discrepancies are found, what is the best way to resolve them? What is the most cost-effective method for auditing? What are the tradeoffs between cost of replication and cost of auditing?

Coping with heterogeneity: How can we bring down the cost of replication, especially if we should administer the replicas independently and house them on heterogeneous platforms? Extra copies of the material cost money, and common wisdom in the IT industry suggests that consolidation of operations and homogeneity of platforms reduces costs. These costs, though, are based on more expensive storage than we perhaps need to use for digital preservation. Can we reduce the cost of replication by deploying cheaper, less reliable replicas, knowing that we'll need to audit and repair across them to increase the overall reliability? Are there scheduling and deployment cost savings in allowing separate procurement and administration of the different replicas? Can virtualization technologies already used to reduce the appearance of heterogeneity in data centers be deployed to reduce the cost of preservation solutions?

Dynamic assets: What is the best way to preserve assets that are not just static media files? For instance, online games require applications and the associated software and hardware platforms to support those applications. Can we emulate these platforms at low cost? As another example, repurposing rendered animations may require preserving not just the rendered frames but the entire workflow and software environment in which

they were rendered, especially if repurposing the material requires changes early in the workflow.

Degradation-tolerant formats and architectures: Can we use digital encoding formats that behave like analog formats in the face of increasing degradation? Such a scheme would continue to deliver a usable asset, merely one that is lower in fidelity. At a higher level, can we build cost-effective preservation systems that fail gradually, providing useful but reduced functionality?

6. Conclusions

As creative industries such as the motion picture and television industries move from analog to all-digital workflows, it is important that we develop the corresponding tools to manage and preserve the digital assets that result. The computer industry has not yet caught up to this task, but we are beginning to understand the requirements of digital preservation solutions and to put these solutions to work.

7. Acknowledgments

We gratefully acknowledge the help, advice and ideas of the HP Labs Pharaoh project team and the LOCKSS project team (www.lockss.org).

8. References

[1] OpenRAW, www.openraw.org. Digital Image Preservation through Open Documentation, July 2006.

[2] J. Reason, “Human Error.” *Cambridge University Press*, 1990.

[3] M. Baker, M. Shah, D.S.H. Rosenthal, M. Roussopoulos, P. Maniatis, TJ Giuli, P. Bungale, “A Fresh Look at the Reliability of Long-term Digital Storage.” *EuroSys 2006*, Belgium, April 2006.

[4] The British Library Digital Object Management Programme, http://www.bl.uk/about/policies/dom/, July 2006.