Rip – Mix – Burn: Why Content Repurposing Is the Answer to Digital Preservation

Andreas Stanescu, Judith Cobb, Taylor Surface, OCLC Online Computer Library Center, Inc., Columbus, Ohio, USA

Abstract

Traditional preservation arguments focused on two technological options, emulation [4, 5] and format conversion [1, 2, 3] or a combination of both [6]. All provide solutions to some preservation scenarios but also list limitations of scalability [3] and unsustainable loss [7], respectively, a common belief which led us to employ a procedural hybrid focused on access and usage, incorporating migration on access advantages described by LOCKSS [3].

This strategy is based on the philosophy that any preservation action made directly to the original digital objects must be postponed until absolutely necessary, arguably forever. However, access is given immediately by employing current-era rendering programs and translating to current-era digital formats which are sufficiently-featured to fully represent the intellectual content in need of preservation. Therefore, we will argue that access, to mix and remix digital content, generates sufficient and necessary incentives [8, 15] to extend the life of digital objects, and indirectly that of digital formats, for as long as they are relevant to some user subgroup.

Additionally, we will introduce assessment and measurement processes complementing the Rosenthal's LOCKSS model. We will develop the main argument by showing the relationships between the three distinct classes of costs incurred by any preservation program. We will show how INFORM methodology [10] is used to time the selection of new programs and formats. Lastly, we will show how translation loss can be measured, managed and validated, by employing a process built on the project at the Library of Congress to analyze the sustainability of digital formats [11].

Introduction

Millions of digital objects are now created by billions of people without even a second thought: digital audio interviews are offered on hundreds of news and government sites, podcasts are used regularly by college professors, video footage covering every facet of breaking news is instantly published, bloggers add investigative reports every day on dozen of sites and everyone can create and exchange digital photos for an initial investment of less than $50. But all these millions of objects are encoded in only a handful of formats, increasingly homogenous, unlike 20 years ago [9].

Graphically, the relationship between digital formats and objects encoded in those formats is represented below:

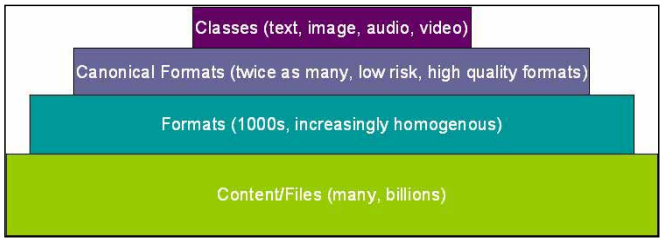

Figure 1: Content Tiers

All the digital content out there can be classified in four basic classes: text, still image, audio and video. This classification explicitly excludes executable programs and scientific datasets which are so diverse it is almost impossible to categorize into generic families.

Canonical formats [12] are digital formats considered best in their class, high quality and low risk. Eventually, each class will have one or two canonical formats identified, which will be capable of encoding all (or most of) the features of the class, regardless of format specification.

The next level of magnitude includes all the different digital file formats: PDF, HTML, XML, RealAudio, MPEG, etc. Since 1980, there were roughly 1000 formats and distinct versions [9] created to encode digital objects. While this analysis shows a set of formats quite diverse, many formats were very short-lived. Recent history appears to show that the actual number of used formats has been reduced to a much smaller number, trending increasingly homogenous.

At the bottom of the stack are the actual digital objects, encoded in one or many digital formats. The number of these objects is increasingly larger and larger, some archives reporting millions and even possibly billions of objects in their custody [14].

Preservation Cost Factors

One of the main challenges in digital preservation is cost. It can be argued that the (unknown) cost of preservation is the biggest impediment to preservation, especially recurring costs. In order to begin determining preservation cost factors, we created a cost model based on the digital content structure discussed above. This cost structure, as related to classes, formats and objects, is represented graphically below:

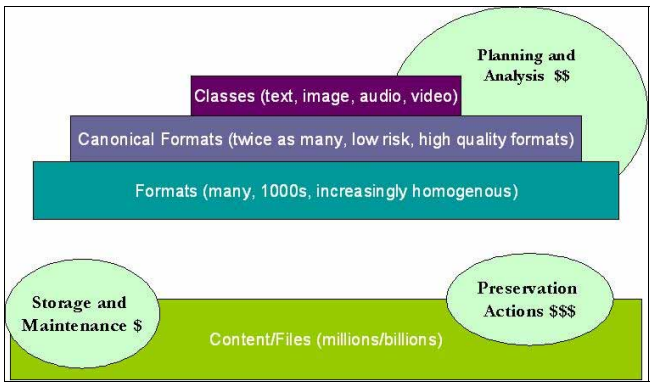

Figure 2: Cost Structure

The cost of preservation can be split into three different components (also see [13]):

1. Storage and Maintenance – the cost of storage plus the hardware replacement costs over time as hardware deteriorates and the cost to periodically verify the integrity of the stored bits. Storage costs are determined **per object**, because all objects require this preservation service.
2. Analysis – all formats held by an archive are identified and their risk level is monitored. The cost **per format** can be spread over **all collections** in the archive. Analysis is ongoing and includes: recurring costs to analyze, test and certify formats (very seldom, 4+ years); recurring costs to monitor technology changes (more often, yearly); and verification of the accuracy of a representative sample of objects (relatively less often, 1-2 years).
3. Action – preservation actions must be taken at the digital object level. The cost **per object** will include the acquisition of the tools necessary to perform the action (either emulation or conversion), verification of the action, certification of the authenticity of the results done both electronically via computer heuristics and human intervention, and documentation of all the actions performed.

The successful preservation program reduces or eliminates **per object** costs because any process or procedure applied to millions or billions of objects cannot be scaled cheaply or predictably. Therefore, in addition to the arguments already shown by Rosenthal [3], the cost classes above clearly point to the need to postpone action until absolutely necessary and instead use a process of analysis **per format**.

To put this in perspective, consider how this cost structure relates to the physical holdings in today's libraries:

1. First, shelves are bought, delivered and installed to hold books. The cost is assessed one-time, when the shelf is purchased and it is directly proportional to the number and size of the books capable of holding. In other words, the cost is **per object**.
2. Secondly, recurring maintenance costs are incurred to hold those materials on the shelf: electricity for air conditioning, gas for heating, shelf and building repairs, etc. These costs are not directly related to the number or size of the materials, but are incurred every day, for as long as the library stays in business. In this case, the cost is for **all collections**.
3. Third, preservation actions, such as removing staples or deacidifying the printed paper, incur costs on an item by item basis. Typically, these costs are not incurred regularly, but still occur from time to time with a one time cost. Clearly, the cost is **per object**.

So, to reduce the cost of preservation, the curator must put in place activities that **do not occur per object** until it is absolutely necessary. Ideally then, a digital preservation strategy will implement a process similar to deacidifying paper collections in mass, avoiding handling each item.

Some analysis activities, such as inspecting the bitstream to recognize the format and extracting preservation metadata, will be done for each object, but the expense is not anywhere near as high as when objects are migrated: the impact on the content must be addressed and this action is much more expensive than just simply inspecting (reading) the bitstream. Again, to put it in perspective, removing staples or neutralizing acid in paper is an expensive preservation action that is often postponed until absolutely necessary, whereas removing the book from the shelf, flipping through the pages and placing it back on the shelf is a day-to-day routine action that libraries often rely on users to do (and inform them of preservation concerns). Even though both actions are applied per object, one is significantly more intricate and expensive than the other.

Rip – Mix – Burn: The Answer to Digital Preservation

As discussed by Rosenthal [3], creating access derivatives on access while allowing various levels of fidelity, offers a number of significant benefits. To wit: originals are saved, hence eliminating loss; derivatives are created by most recent, and presumably best, technology available; only accessed content is transformed, therefore lowering the costs per object; only analysis processes are regularly executed to identify the most comparable current-era format, hence reducing operational costs to the minimum.

Our approach extends these concepts by addressing a much more basic issue: who would pay for all the effort and why would they do it? We assert that when the digital objects in question have enough commercial value to their owners, there will be sufficient incentives to address the following four mandatory preservation aspects:

- pay for the "preservation tax" needed to regularly perform the format analysis
- pay for pre-created access derivatives for more and more content
- extend the life of existing digital formats by creating necessary rendering software for current-era computers
- demonstrate access derivatives accuracy using feedback from the actual users

If commercial value is increased by exposing the digital content to more potential users in more creative ways, then ultimately the goal is to increase access, hence *Rip – Mix – Burn*.

Rip – Mix – Burn: An Inside Look

According to Gladwell [8] and Lavoie [15], when digital objects will significantly increase the bottom line of publishers and authors alike, they will then have all the necessary incentives to create more digital content and to make sure existing content is not lost. If our goal is to increase access, we can do so by creating tools and services capable of mining commercial value out of any and all digital objects, regardless of their initial purpose, responding to the needs of future users, whenever they need it, wherever they need it.

Furthermore, when collections are exposed to larger and larger groups of people, they are bound to create interest niches, as The Long Tail theory [17] discusses. And, since interest groups have a way of connecting and creating social networks, exposure becomes a function of sharing – more exposure is brought on by more sharing – but the latter is only possible when collection exchange is fully interoperable and employing a common and simple licensing scheme.

The possibilities are endless. To wit:

- Google's Book Search Library Project is only the beginning of making all previously non-digital content available in the digital world. While few argue its benefits, a simple licensing scheme through which copyright holders are paid could alleviate infringement concerns.
- TeachersDomain.org is a value-add service dedicated to reformatting multimedia resources for K-12 classroom-ready use, conformant to different state requirements, as well as sensitive to learning and comprehension levels. Such a service would remix digital objects into new learning objects, therefore creating knowledge from knowledge and improving education at the same time.
- Mashups are defined by Wikipedia as the "combination (usually by digital means) of the music from one song with the a cappella from another" but in the context of the new Semantic web, it means "combining content from more than one source into an integrated experience". For example, HousingMaps.com combines craigslist rentals with Google's map service; Chicagocrime.org overlays local crime stats onto Google Maps so one can see what crimes were committed recently in a neighborhood. When copyrighted digital content is allowed to live, people find innovative solutions to problems nobody knew existed, and we can be assured that high quality content preserved in digital archives will especially find numerous interests all over the world.
- Digital reproductions of priceless museum collections and library materials can be commercially reused in advertisement, entertainment and edutainment. For example, video game writers can place a new action game in 5000 BC Egypt by using high-quality master images of Egyptian collections all over the world while at the same providing funding for the institutions owning the rights to those images.
- Even public information can have incredible potential for the 21st century political games we play. In a world of 24-hour news networks and 24/7 political strategy, decades old letters, reports, statements or voting records can prove decisive in a Supreme Court nomination, public office election or cabinet appointment.
- Professors and researchers all over the world can find and reuse each others' results and materials to develop drugs faster, to respond to health emergencies more rapidly and decisively, to invent new things or to advance the state-of-the-art of technology. When data is cataloged and easily found, similar experiments don't have to be redone, conclusions already reached can just be used as stepping stones and progress is accelerated for everyone's benefit.

Impact on Preservation

An active digital content exchange market will alleviate the need to preserve conversion programs. As discussed by the genial computer scientist Eric S. Raymond [16]:

"Every good work of software starts by scratching a developer's personal itch."

In this context, that means that when someone believes that a certain collection is worth keeping around, that someone will spend all the necessary resources to make it happen. This someone can be an individual, an interest group, a library or a corporation. The Long Tail theory applies in the preservation context as well because everyone has the power to make a difference, an individual or an organization. One would expect a library or university to have greater resources and therefore protect a larger share of digital content, but it is feasible that even one individual can build that one rendering program necessary to salvage a unique collection.

If and when collections start losing commercial value (or for collections that had little commercial value in the first place), social and moral incentives will exert sufficient pressures to ensure public's access to information, to "do the right thing" and save culturally valuable collections or to keep memory alive.

Notice that it is the collection that is at the heart of the argument, not the formats. People and institutions find collections valuable, without giving a second thought to formats. However, since most digital collections are encoded in a small number of digital formats, the effort spent on one collection's formats automatically extends to all the other collections using the same digital format. By exposing as many collections as possible, we would be exposing as many digital formats as possible, therefore increasing the chance that each and every one continues to be supported in the years to come. Lastly, actual users of all stripes will inform how the objects are accessed, therefore precisely validating the accuracy of the transformation process because they will actually exercise what they believe to be significant properties of the intellectual content [18].

A manageable strategy

A preserved object must give its users the assurance that all of its intellectual attributes, or significant properties [18], are indeed the same now as they were when the object was first created. Rendering a preserved object will involve at least a change in environment and at most a change in format or bitstream, so it is mandatory that the accuracy and authenticity of the object is proven and documented.

Other preservation strategies, such as durable encoding [6], suggested preserving both the programs (algorithms) used to understand and render the objects, as well as the proof that the document is indeed original. Since we believe rendering programs will continue to exist, we only need to focus on documenting how

well those programs stay true to the original's significant properties, a task that is done automatically via computer heuristics on representative samples and manually by content owners and users alike. We believe this approach reduces the overall size of the preservation problem to only measurements, risk mitigation and documentation.

Access derivatives will be created directly from originals and can be implemented at various levels of fidelity. Fidelity can vary by the type of user, the network bandwidth available, the device used to request it and for many other reasons. For example, we can create low resolution images or lo-fi audio and video for public access or for mobile network access, while at the same time have master quality images and hi-fi audio and video streams available for authenticated users over broadband connections.

Preservation Process

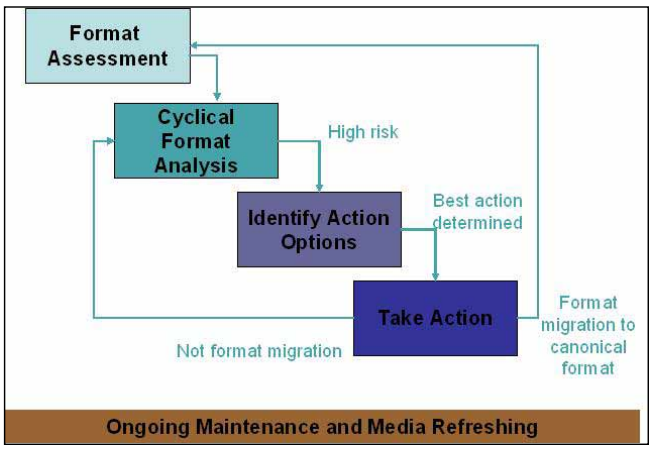

Figure 3: Preservation Process

The process we propose to implement is:

1. Assess all the formats by format class – the four classes of objects are text, still images, audio and video
   - acquire the mapping of format characteristics to the class abstract feature set [11], to determine its relative quality as defined and documented by experts in each field
   - define, test and certify the bitstream identification, analysis & rendering platforms, the Current Era Platform and the Reference Platform, for each format in custody
   - measure the format's risk by applying the INFORM risk assessment methodology [10]
   - measure access derivation translation "delta" and determine if this "delta" preserves accuracy and integrity of the "preserved copy"
   - synthesize the result of the analysis into metadata and apply comparison algorithms on this resulting metadata to learn the relative strengths and weaknesses of each format

2. Implement a cyclical analysis process:
   - a thorough technology watch process, focused on those formats and the software & hardware systems needed to render them
   - re-assessment every 18 months of all the formats currently known

3. As risks increase, determine best course of action, first by manipulating aspects related to the formats, and, as a last resort, individual objects:
   - by understanding that access and re-purposing drives preservation, and by measuring and documenting the needs of the designated communities
   - by using the definition of significant properties [18] for the format to determine the best path of action for specific collections and objects, if such action is necessary
   - by identifying, assessing, implementing and testing a new computing platform[1] capable of rendering the objects in their current form (i.e.: a new rendering program, a virtual machine, another operating system or hardware system)
   - by identifying, assessing, implementing and testing one or more authenticated derivative, based on the consumers' needs at that time
   - by recognizing that postponement of object-level actions are indeed necessary and migrate all objects to one of the canonical formats

4. Take action and document the reasons of the decision as well as the roles of those people who made the decision, including the archive managers and collection curators:
   - if risks are not considered high enough, document the new threshold and instruct the next occurrence of the cyclical analysis step
   - if a change is made, first allow collection curators to re-assess their collections, based on existing usage and other community needs, and determined which content will be acted upon and which will be destroyed
   - if a platform change is made or new derivatives are created, certify and document the accuracy of the new platform or derivatives, in order to maintain a proper trail of provenance and authenticity metadata:
   - create a verification algorithm to precisely determine and document loss, if any, by using the format analysis above and the class abstract feature set
   - test the proposed changes by applying the verification algorithm against a sampling of objects from the archive

Process Details

The process is designed to reduce complexities introduced by any changes in the make up or the environment of the object, including format migration and emulation. Each step builds on the work of many, including the Library of Congress and our own OCLC Digital Archive and Office of Research.

Canonical formats

[1] Two platforms are actually defined, one that considers the software and hardware investments already deployed by the beneficiary communities and another that uses software and hardware fallback solutions based on open source and open licenses, when current technologies become obsolete or too difficult to maintain. See section below on software and hardware.

Each class of content – text, images, audio and video – all have a very small number of preferred digital formats, called canonical formats. We will use the tables defined by the Library of Congress [11] to pick one or two formats to become our preferred formats.

These canonical formats will be suggested for use to our constituents in an effort to not only improve the quality of the master copies stored in the archive but also to reduce the cost of ongoing preservation activities because they are believed to have the longest lifespan potential. This belief will be tested by the risk measurement step detailed below.

Class abstract feature set

The class abstract feature set holds descriptions of capabilities of all the formats in the class: for example image size, color spectrum or clarity for images, margins, fonts and layout for text. Library of Congress used the most up-to-date knowledge to create an abstract set of characteristics, applicable to all past and future formats in that class. This step is of particular importance because it uses expertise not readily available to any and all digital archives to translate in-depth expertise to easily understood terms [11].

Software & hardware platforms

For every format and version, we will identify a current era platform and one or more reference platforms. If possible, at least one reference platform for each format will be an open source solution.

Current era platform (CEP) is a defined environment consisting of applications, operating systems and libraries, and hardware that can be used with the Digital Archive system for one or more formats. The current era platform is current technology, available to end users and probably already running on their PCs. OCLC will define and test various environments for all the formats in its DA.

Reference platform (RP) is an environment consisting of primarily open source applications, operating systems and libraries, and hardware. The belief is that the open source components will be less volatile than the current era platforms. However, in reality, it will be impossible to have all reference platforms be completely open. Instead there will be a range of open source (open license, open specifications) and proprietary components. The RP may be used as a fallback solution when the CEP reaches end-of-life.

Risk measurement

The INFORM methodology [10] offers a systematic and objective way to measure the viability of a digital format. The digital format and the candidate software & hardware platforms are assessed to determine the preservation viability of this solution.

The following steps must be taken to assess the preservation viability:

1. Identify any other format specification on which the source specification may be dependent – select from the classes of potential candidates listed in section *Risks of the digital format*
2. Identify any software dependencies on which the source specification is dependent – select from the classes of potential candidates listed in section *Risks of required software*
3. Identify any hardware dependencies on which the source specification is dependent – select from the classes of potential candidates listed in section *Risks of required hardware*
4. Identify organizational dependencies introduced by each software and hardware dependencies and identify any other potential classes of dependent organizations – select from the classes of potential candidates listed in section *Risks of associated organizations*
5. Assess the current implementation of the Digital Archive – use the format presented in section *Risks of the Digital Archive*
6. If the candidate solution uses a migration strategy, assess the migration processes – use the format presented in section *Risks of preservation plans based on migration*
7. Obtain the preservation factor of each preservation solution by combining the preservation factors of each category, for each format specification, for each software and hardware dependency, for each associated organization, for the Digital Archive and the possible migration assessment, as described in section *How to apply the risk model.*

Conclusion

The OCLC Digital Archive is developing this preservation strategy based on technological, policy, and economic forces. This document, and our work on the preservation strategy, represents an ongoing process that builds on the work of others while attempting to create a viable, affordable, implementable preservation strategy for digital materials.

References

[1] HESLOP, H., DAVIS, S., AND WILSON, A. National Archives Green Paper: An Approach to the Preservation of Digital Records. http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper%25.pdf, 2002.

[2] THE FLORIDA CENTER FOR LIBRARY AUTOMATION. DAITSS Overview. http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf, 2004.

[3] ROSENTHAL, D. S. H., LIPKIS, T., ROBERTSON, T. S., AND MORABITO, S. Transparent format migration of preserved web content. D-Lib Magazine 11, 1 (Jan. 2005). doi:10.1045/january2005-rosenthal.

[4] LORIE, R. Preserving Digital Documents for the Long-Term. In IS&T Archiving Conf. (San Antonio, TX, USA, 2004), pp. 88-92

[5] ROTHENBERG, J. Ensuring the Longevity of Digital Documents. Scientific American 272, 1 (1995). http://www.clir.org/pubs/archives/ensuring.pdf.

[6] GLADNEY, H.M. Trustworthy 100-Year Digital Objects: Durable Encoding for When It's Too Late to Ask . ERPAePRINTS (2004) http://eprints.erpanet.org/7/ .

[7] van DIESSEN, R., STEENBAKKERS, J. The Long Term Preservation Study of the DNEP Project. (2002). ISBN 90-6259-154-X, pp. 4.

[8] GLADWELL, M. The Tipping Point: How Little Things Can Make a Big Difference. Back Bay Books (2002). ISBN 0316346624.

[9] Wotsit's format: the programmer's resource. Available at http://www.wotsit.org/

[10] STANESCU, A. M. Assessing the durability of formats in a digital preservation environment: the INFORM methodology. OCLC

Systems and Services, vol. 21 no. 1. 2005. Emerald Group Publishing Limited.
[11] Library of Congress. Sustainability of Digital Formats: Planning for Library of Congress Collections. March 2005. http://www.digitalpreservation.gov/formats/content/content_categories.shtml
[12] LYNCH, C. Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. D-Lib Magazine 5, 9 (Sept. 1999). doi:10.1045/september99-lynch.
[13] ROSENTHAL, D. S. H., ROBERTSON, T. S., LIPKIS, T., AND MORABITO, S. Requirements for Digital Preservation Systems: A Bottom-Up Approach. D-Lib Magazine 11, 11 (Nov. 2005). doi:10.1045/november2005-rosenthal.
[14] TALBOT, D. The Fading Memory of the State. Technology Review. July 2005. http://www.technologyreview.com/articles/05/07/issue/feature_memory.asp
[15] LAVOIE, B. F. The incentives to preserve digital materials: roles, scenarios and economic decision-making. OCLC Research. 2003. http://www.oclc.org/research/projects/digipres/incentives-dp.pdf
[16] RAYMOND, E. S. The Cathedral and the Bazaar. 1997-2000. http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/
[17] ANDERSON, C. The Long Tail. Wired Magazine 12, 10 (October 2004). http://www.wired.com/wired/archive/12.10/tail.html
[18] HEDSTROM, M., LEE, Christopher A Lee. Significant Properties of Digital Objects: Definitions, Applications, Implications. 2002. Copy from Internet Archive Wayback Machine: http://web.archive.org/web/20030114060912/http://www.dlmforum2002.org/download/margaret_hedstrom.PDF

Author Biography

Primary Author

Andreas Stanescu is Software Architect for OCLC's Digital Archive. He is developing and prototyping processes to create preservation plans for documents ingested by the Digital Archive, including a method to identify and measure changes in the supporting IT environment. As technical lead, Mr. Stanescu focuses on the system architecture for the OCLC Digital Archive and optimizing it for preservation. Prior to joining OCLC, Andreas developed a software program that secured access to system services and implemented strong cryptographic solutions to protect data integrity. His BS and MS in Computer Science degrees are from Franklin University, Columbus, Ohio.

Contributing Authors

Judith Cobb, Senior Product Specialist, OCLC Online Computer Library Center, Inc.

Taylor Surface, Global Product Manager, OCLC Online Computer Library Center, Inc.