IJDC | *Conference Pre-print*

Mutually Assured Preservation: Fostering Active Preservation Practice Through 'Fire Drills'

Bradley J. Daigle
APTrust / University of Virginia

Abstract

Sound preservation practice is a series of active engagements with the content one hopes to preserve. In many cases, this has not always been the case. Both institutions and services—while not actively encouraging passive preservation—neglect the key components in the stewardship of our historical record. In other words, there is much more to preservation than simply choosing a storage solution and placing one's content there. The materials need to be verified, checked, and tested against expectations within the service. This is accepted practice for many. However, very few services provide the necessary assurance to test both its own user expectations as well as the depositors' themselves. Creating a methodology for both depositor and service to be assured that preservation meets expectations is critical. This is happening in very select ways. This paper discusses one such dialogue and its function.

.

Submitted 16 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Bradley Daigle, University of Virginia Library , P.O. Box – 400109 2450 Old Ivy Rd. Charlottesville, VA 22903. Email: bjd2b@virginia.edu

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

While at a PASIG (*Preservation and Archives Special Interest Group*) conference, in 2019 in Mexico City, there was an extended conversation about the role of preservation transparency and responsibility. The primary questions were: what role did the preservation service play with respect to the depositor; and what standards should the depositor employ to ensure the service is meeting those expectations? These are highly appropriate questions directly related to preservation stewardship responsibility. The responses were—as to be expected—thorough and thoughtful. However, upon deeper reflection, perhaps they reveal an implicit bias in approaching preservation. The preservation community has been confronted with a myriad of preservation options: in house; external, not-for-profit; external, for profit; and a combination of all three. Though many of these approaches take an almost paternalistic approach to preservation. In other words, they encourage the depositing of content and, in fact, do all they can to lower the threshold for that practice. Cloud services and many vendor products reflect this philosophy. Uploading and storing of content deploys business models that encourage content transfer but discourage content withdrawal. Robust read/write actions by the depositors are, for the most part, quietly discouraged through the levying of significant fees. The end result is that the service predisposes an almost passive approach to stewardship on the part of the content owner—forcing them (logistically and economically) to rely on internal and available reporting of preservation actions externally undertaken on one's content. Hence the question referenced above about assurance.

In most cases, fixity checking is the primary concept seized upon to prove the efficacy of the preservation service. In the latest ITHAKA report on the state of digital preservation, Oya Rieger notes "The key to digital preservation is sustaining interactivity and variability to support future uses in addition to considering the core archival principles such as authenticity, fixity, and integrity." (Reiger, 2018) These key elements can be observed and guided by various certification bodies and standards.[1] There are, of course, many other means to track what preservation actions are executed on a depositor's content. The Library of Congress's Preservation Metadata Standard (PREMIS) is a particular example.[2] From a critical perspective, these activities take an enormous amount of resources: from standards awareness and compliance; service reporting; and depositors logging and auditing components such as PREMIS events. Often one considers the burden of preservation to be borne by whichever service a depositor uses. Much of the apparatus of preservation relates directly to the need for transparency, clarity, and systemic responsiveness.

Given that these activities are the fully realized mission for any preservation service/depositor relationship is there anything that could be missing in this picture?

Perhaps another way to phrase the question is what is the nature of the method of transparency between service and depositor? The published survey results from the Digital Preservation Network (DPN) describes a level of frustration with access to and lack of reporting on one's content.[3] DPN's content model was based on a twenty-year window of deposit—a model that was greeted with a certain degree of scepticism from depositors. The wariness was not related to the service but rather the depositor's ability to interpret content that had been submitted decades previous. This latter point leads to the missing element in the scenario outline in the previous paragraph: namely that no technical transparency, reporting, and responsiveness can replace the simply question of whether or not a depositing organization knows what to do

[1] See, for example, TRAC or TDR certification [e.g. CRL: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/iso16363] —which are based on the ISO 16363 standard. Also, more recently, CoreTrustSeal [https://www.coretrustseal.org] has emerged as a viable certification body.

[2] See https://www.loc.gov/standards/premis/

[3] Digital Preservation Network Final Report, (2018): 18, https://osf.io/3p9jq/

with its own preserved content once it needs it back. In other words, there is a means—a very clear means—to test a preservation service's assurance by pulling down one's content. Did the depositor receive what was expected? Does the content match the original transport manifest? These all are metrics of assurance of any service. As stated earlier, this data point is mitigated by the read/write costs associated with retrieving one's content. Therefore, if one finds a solution to the costs of retrieval how can one test the depositor's resilience in interpreting the results? If they cannot, then mutual assurance is very difficult. How does this impasse get solved?

The purpose of this paper is to explain from both a content and technical perspective a singular approach that provides assurance of both a preservation service and the depositing entity. We at Academic Preservation Trust (aptrust.org) seek to provide a rounded approach to preservation responsibility through a means of mutual assurance. This paper will describe what the membership designates as "fire drills"—which in technical terms refers to as random test restores of a depositor's content. Briefly stated, these drills provide the depositor with random samples of content they have deposited. This happens at random intervals as well and is based on a technical algorithm that allows us pull down content at intervals that do not accrue egress fees – thus providing the means to test durability at no additional cost. The first part of the paper will outline the basic strategy and the second half the specific technical implementation of that strategy.

Restoration Testing as a Means of Assurance

Active and Passive Preservation – the question of fault

When it comes to taking responsibility for preservation activity, the duty can be assigned in many ways. It can be assigned to the individual within an organization, the organization as a whole, or the preservation service if it is an external entity. Experience over the last few decades has shown that responsibility for what preservation actually means can frequently shift depending on one's organization and perspective—and this shift can often be imperceptible, going unnoticed until a crisis occurs. This responsibility can extend beyond the standard indemnification clauses couched in various agreements, MOUs, and contracts that point to specific culpability and filter down within an organization to a single department or staff member. Fault can also be tied to the "why and how" we are preserving our cultural heritage as a profession. These can be both content strategies [4] as well as broader, environmental studies.[5] Preservation needs to be an active set of actions rather than passive—and this set should be guided by organizational alignment with preservation policy. Preservation stewardship is active in the sense that it is an ongoing activity—one that is never truly finished, an asymptote in its purest form. Preservation stewardship is cannot be passive in that preservation is not just storage. The act of putting collections into storage should not be construed as true preservation without the active management of those materials (fixity, versioning, etc.) and that management clearly assigned to various staff within an organization.

[4] See, for example the Jisc study by Neil Beagrie: "What to Keep: A Jisc Research Data Study" 2019, https://repository.jisc.ac.uk/7262/1/JR0100_WHAT_RESEARCH_DATA_TO_KEEP_FEB2019_v5_WEB.pdf

[5] New, non-content specific metrics are coming into play for this overview of collecting strategies. For example, figuring in the carbon footprint of the various data centers needed to store content is becoming an increasing concern. See James Glanz, "Data Centers Waste Vast Amounts of Energy, Belying Industry Image." The New York Times, September 22, 2012, sec. Technology. http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html ; Keith L. Pendergrass, Walker Sampson, Tim Walsh, and Laura Alagna (*2019*) Toward Environmentally Sustainable Digital Preservation. The American Archivist: Spring/Summer 2019, Vol. 82, No. 1, pp. 165-206.

The Assigning of Responsibility in Preservation

Organizational responsibility

There are some clear indicators that comprise a comprehensive organizational approach to preservation. Creating and managing a strategy that is keyed to an organization's mission and sustainable resources plays a major role in the process. There is no shortage of primers on the subject. However, a basic starting point could be the following:

1. Collection Development/Curatorial Policy [what do you collect?];
2. Preservation Policy [how does your organization define preservation?];
3. Technical pathway for preservation (e.g. the Levels of Preservation[6]) [how is a preservation strategy applied?];
4. Clearly assigned roles for updating / managing each of these components [who does what and with what kind of accountability?] and finally;
5. Review schedule [how often do you need to revise this process?]

By no means a comprehensive list, these steps are a good place to start for any organization. These are some key elements that comprise an organization's approach and pathway to preserving content. As noted earlier, there are standards—both international and community that can provide initial signposts for preservation. It is the responsibility of the organization to have these elements in place—particularly if active preservation is a core goal.

Responsibility of the service

There are some basic elements that are hallmarks of a successful preservation service. These can range from the publicly detailed compliance with ISO 16363 to a more customized solution for a given organization. Many of the basic principles noted above still apply. Steps 1-2 take place locally but perhaps the third step is undertaken by an external service. That would make numbers 4 and 5 shared between organization and service. Clarity among all the roles and specific responsibility for each action is perhaps more important when it is undertaken by an external service. If that service is either unable or unwilling to provide specific details on what that means for what you need, that should be seen as a warning sign. That is because the obvious element that runs through all of these components is transparency. Without the means to clearly verify what is being expected and undertaken (by any agent) then accountability disappears. In particular, when an organization relies on a third-party service, that transparency and accountability rise to the top of any set of functional requirements.

With Great Responsibility…

What role does transparency play in assurance? In some cases, vendors do not provide the level of transparency commensurate to their preservation responsibility. Or, if they do, it is not widely shareable due to some form of non-disclosure agreement. This can be for reasons that are quite sensible—some of these solutions fall under the rubric of trade secrets, the sharing of which would put that service's existence at risk. As understandable as this reality is, how do we, as a profession, continue to absorb responsibility for preservation if we cannot know how it is being done or cannot easily share our stories about its success and failures? Amazon Web Services is a singular example in this regard with their (now famous) assurance statistics of the "11 9's).[7] This is where assurance comes in. The drive to obfuscate certain technical details within preservation solutions to protect a for-profit service's investment is the nature of working with vendors. Understanding that is part of the choice when outsourcing preservation services. We, as a

[6] The NDSA Levels of Preservation have been recently updated after remaining static since 2013. More information can be found here: https://ndsa.org/activities/levels-of-digital-preservation/
[7] https://docs.aws.amazon.com/whitepapers/latest/aws-overview/storage-services.html

profession dedicated to preservation, need to understand more fully how to work with services that limit the sharing of critical information. Preservation practitioners need to develop assurance metrics—metrics that can be clear and verifiable but not wholly revelatory of trade secrets. It is up to us to do this because the private sector will not. If we do not, then we need to think long and hard about using services that do not allow for open sharing of how preservation takes place.

Active Preservation: "Fire Drills" as Content Restoration Tests

One possible solution to testing assurance this paper refers to as "fire drills". Also known as content restoration spot tests, these are periodic requirements of a depositor to verify the restoration some of the content that they have submitted to that service. This has a trifold purpose. The first is that it provides a verifiable action that the content is what it is expected to be - thus giving the depositor greater assurance that the materials are being preserved properly and according to their organizational metrics. This is a critical test of the service itself. A depositor should always get they expect. Anything else compromises the assurance of the service. The second purpose, of perhaps even greater value is the test these restorations provides for the organization that owns the material. In other words not only should the depositor get back what was expected but they need to be able to understand and identify their own materials—particularly if they were deposited by previous staff or from a different department. That understanding is as much a test of the depositor's active preservation practices as it is of the service that holds their content. Finally, combining the two elements of mutual assurance, these "drills" function as a robust test of a system's ability to do restorations easily and reliably as part of a suite of best practices. No one plans for a disaster or emergency. Knowing that the system or methods one uses are durable and tested frequently forestalls against many possible vagaries. Therefore, it can be shown that these fire drills underpin an active preservation strategy by testing the durability, resilience, and application of both the service and organization.

After some initial testing, APTrust is working with its members to determine a frequency that may allow for some configuration of these restorations. For example, a new member might request a greater number of fire drills initially to test both its own workflows as well as the assurance of the service. Then over time, the frequency may decrease as confidence rises. There is also a need to balance the costs of moving materials around on both sides (staff time and I/O costs) and the environments in which content is deposited so the right balance must be struck.[8] Initial results have also provided some insight into how depositors bag and submit their collections. Obviously, if a collections package is several terabytes this would not necessarily be something that should be restored frequently. Therefore, taking into consideration how and in what ways these restorations will test a depositor's local preservation management will be part of the equation. As APTrust moves forward with these innovations, we encourage other services to explore the means by which we, as a community, can align around fire drills as a best practice. We will report out more as we delve deeper into this engagement with our members.

Cause for Danger: Real Fires

As if the need for preservation fail safes and transparency weren't clear enough, the recent investigations of the "Universal Studio Fire" that happened in 2008, the extent of which was effectively suppressed, is only now becoming truly known.[9] How many more examples like this need to be brought to light? In many ways, the preservation community still functions like the special collections and archives community of decades past—where theft and disaster lessons

[8] I have frequently stated in presentations that preservation is not a technical problem but a behavioural and cultural one; David Rosenthal argues that it is also an economic one: https://blog.dshr.org/2017/08/preservation-is-not-technical-problem.html

[9] Jody Rosen, "Here are Hundreds More Artists Whose Tapes were Destroyed in the UMG Fire," New York Times, 25 June 2019, https://nyti.ms/2ZMP7oq

were not widely shared due to fears of reputation loss and donor concerns. How could the scope of such a disaster as the Universal fire have been mitigated if the community had known earlier? This is why we need to encourage and support conversations around preservation disasters/mishaps/snafus in a way that we can learn and grow as a community. Organizations like the *Digital Preservation Coalition* (dpconline.org)[10] work to create those spaces and opportunities and continued effort along those lines needs to become ongoing, normalized practice across the wider profession.

The work that APTrust has begun with its members and the broader community of distributed digital preservation services is just one of the efforts to increase assurance and transparency of practice. Expanding this effort across other not-for-profit services might prompt the for-profits to engage in similar conversations. This would be of benefit to the entire community at large. Approaches and outcomes from ideas such as random test restorations require ongoing effort and feedback from the community and could become critical components of a shared approach to digital stewardship. Digital preservation is predicated on active engagement with our historical record. Restoration "fire drills" provide the means by which we can test both a service and one's organization's preservation readiness.

Restoration Tests: Technical Details

To understand more fully and in technical detail what is meant by these restoration tests, this section will outline the details of how this is undertaken. In February, 2019, APTrust began automated monthly restoration spot tests. APTrust randomly restores one bag from each depositor institution and then emails administrators at that institution to let them know that the system has automatically selected some of their content to be restored. The depositor's responsibility is to examine the bag to ensure it is complete and they can make sense of its payload. Through spring of 2019, depositors have been responding by email to confirm that their bags are complete. In addition to that work, there is a plan underway to capture member responses as part of the repository's core metadata. This will provide an audit trail of past restorations and their results over time. APTrust uses Amazon Web Services (AWS) as its preservation storage backend. However, it has built its own web services on top of that storage layer to provide robust preservation reporting for its members.

Restoration Considerations

APTrust receives materials in a verified BagIt format.[11] Once received, the system then unpacks the bags and stores its contents (both payload files and tag files) as individual files. The files in a bag usually constitute a single intellectual object, though some larger objects may be split across multiple bags. A database registry, separate from preservation storage, keeps track of which files are logically grouped into which intellectual objects. This registry would also allow for the recreation of intellectual objects and their metadata in case of an organization's catastrophic loss of its content management system.

During the restoration process, APTrust reassembles all of the files that constitute an intellectual object into one or more BagIt bags and moves the bag(s) into an AWS S3 restoration bucket from which the depositor can download their content. Although both the SIP[12] the depositor originally submitted and the DIP that the APTrust system returns to them use BagIt format, the DIP that is restored in this process is guaranteed not to be identical to the SIP the depositor submitted for the following reasons:

[10] See also the *Digital Preservation Coalition's* suite of advocacy tools for preservation practitioners of all levels: https://dpconline.org/knowledge-base/advocacy
[11] Based on the Library of Congress: http://www.digitalpreservation.gov/documents/bagitspec.pdf
[12] Based on the OAIS reference model: https://public.ccsds.org/Pubs/650x0m2.pdf

1. Depositors may have deleted individual files from the preserved intellectual object.

2. Depositors may have uploaded newer versions of the object, containing new or altered files.

3. Depositors typically submit bags with either md5 of sha256 manifests, while APTrust restores bags with both manifests.

4. Depositors generally omit tag manifests in the SIP, but APTrust includes them in the DIP.

5. APTrust includes a JSON file in the restored bag describing all PREMIS events for the intellectual object and each of its constituent files.

6. Because an object's files may have been added, deleted, or replaced with new versions after initial ingest, the PREMIS events JSON file is essential for the depositor to understand why the restored bag differs from the originally submitted bag. This file contains a record of all events affecting the object since ingest, including deletion of files, addition of files, and re-ingest of files, with a full history of all ingest checksums and fixity checks.

When verifying the contents of a restored bag, depositors typically must validate the bag and ensure it contains everything they expect it to contain (including tag files that may hold information useful for re-importing content into a local system).

As of spring 2019, all depositors have confirmed the validity of all restored bags. This early success is an indication that the process can be an effective means to tweak local practice and preservation expectations of both the service and depositor.

Restoration Spot Test Process

APTrust designed the spot test process to mimic the normal depositor-initiated restoration process as closely as possible. Depositors restore an object by clicking the Restore button in the APTrust web user interface management web service called, Pharos or by sending an API request to Pharos. Either of those actions creates an entry in the restoration work queue with the object's identifier.

The restoration spot test algorithm selects one object belonging to each APTrust depositor and creates an entry in the restoration queue with the object's identifier. The criteria for choosing objects are:

1. The object must not have been restored in the past 180 days. This is to ensure sufficient randomness of the depositor's materials.

2. The object must be 50 GB or less in size. (This is a current criterion so as to not burden depositors with large downloads.)

Once the object identifier is in the restoration queue, the process is the same regardless of how it got there. The restoration tests are meant to encourage the depositor's local agents to verify the content but not overwhelm them. To that end, the depositor has a given window of time to verify the content restoration otherwise the materials will simply be removed from the restoration bucket. This allows for some flexibility in a member's current state. However, a policy is being written within the membership to require a set number of annual restorations to be validated as part of a suite of best practices. APTrust is also working on a configurable sliding scale of restoration "windows" that a depositor can choose that best works within the policy stipulations.

Restoration Process

For both depositor-initiated and system-initiated restorations, APTrust does the following:

1. Gets a list of all files that constitute the object from Pharos (which includes the APTrust metadata registry). This list includes both payload and tag files.
2. Copies all files from preservation storage to a local staging area.
3. Ensures all files present, and that all checksums match what's in the registry.
4. Bags all of the payload and tag files, creating md5 and sha256 manifests.
5. Adds to the bag a JSON file containing all PREMIS events related to the object and all of its files. (This is added as a tag file, not a payload file.)
6. Creates md5 and sha256 tag manifests.
7. Validates the bag.
8. Copies the bag to the depositor's S3 receiving bucket.
9. Sends an email to the depositor saying the restored bag is available for download from the receiving bucket.

Post-Restoration

APTrust automatically deletes the restored bag from the depositor's restoration bucket after 14 days to avoid incurring unnecessary costs. If the depositor did not retrieve it in the 14-day window, they have to initiate a new restoration. For spot restoration tests, APTrust asks depositors to respond via email to say whether they received what they expected and whether they were able to make sense of the restored content. As stated above, APTrust is considering implementing some policy requirements in the future in order to capture depositor responses in the APTrust metadata registry.

Conclusion

The concept of "fire drills" is meant to test both the preservation service and the depositor's durability as part of an active preservation stewardship practice. It is not the only means by which this can be accomplished but in light of the need for assurance across the preservation landscape it appears to be a good start. As a means of mutual assurance, it also avoids the "black box" approach of a given service—allowing the organization (whether it uses internal or external solutions) to test that service as well as its own ability to approximate a preservation emergency.

Ongoing testing and iteration are required to investigate more fully the repercussions of these restorations as well as a broader conversation with the preservation community to help articulate the success metrics of such an approach. As a proof of concept with over a dozen test organizations we have begun to test the waters for assurance metrics that span any service. Perhaps at a future date, these metrics will become part of a standard that will help any organization articulate and its approach to an active digital preservation stewardship program. This work is ongoing and the responses and feedback from the community are critical components of this work. Digital preservation is predicated on active engagement with our

historical record. Restoration tests provide the means by which we can test both a service and our own preservation readiness.

References

Reiger, O. (2018) The State of Digital Preservation in 2018: A Snapshot of Challenges and Gaps, 2018, https://sr.ithaka.org/wp-content/uploads/2018/10/SR-Issue-Brief-State-Digital-Preservation-20181022.pdf