Archiving in the Networked World: Metrics for Testing

by Michael Seadle,
Editor, *Library Hi Tech*
Professor and Director,
Berlin School of Library and Information Science

Abstract

Type: Research article

Purpose: This column looks at how long term digital archiving systems are tested and what benchmarks and other metrics are necessary for that testing to produce data that the community can use to make decisions.

Methodology: The article reviews recent literature about digital archiving systems involving public and semi-public tests. It then looks specifically at the rules and metrics needed for doing public or semi-public testing for three specific issues: 1) triggering migration; 2) ingest rates; and 3) storage capacity measurement.

Findings: Important literature on testing exists but common metrics do not, and too little data is available at this point to establish them reliably. Metrics are needed to judge the quality and timeliness of an archive's migration services. Archives should offer benchmarks for the speed of ingest, but that will happen only once they come to agreement about starting and ending points. Storage capacity is another area where librarians are raising questions, but without proxy measures and agreement about data amounts, such testing cannot proceed.

Implications: Testing is necessary to develop useful metrics and benchmarks about performance. At present the archiving community has too little data on which to make decisions about long term digital archiving, and as long as that is the case, the decisions may well be flawed.

Introduction

Is testing a binary function -- i.e., a system works or not -- or a set of benchmarks or a continuous performance curve? Participants at the Aligning National Approaches to Digital Preservation (ANADP) conference [1] in Tallinn, Estonia, 23-25 May 2011, agreed that the binary approach was too simplistic. The fact that a program will compile and run against a sample set of data does not mean that it works in any meaningful way. Of course all publicly available archiving and preservation systems have undergone more extensive testing than that, but no one at the conference had a clear notion of what metrics should apply to make it possible to compare the effectiveness of one system with another. Nonetheless conference participants (including vendors) generally agreed that the digital archiving community needs to develop benchmarking and other forms of measurement to enable a fair and open comparison and to make it possible for libraries and other cultural heritage institutions to make rational choices for long term digital preservation.

This article builds on discussions that took place at the ANADP conference. Testing was a major leitmotiv on the Technical Alignment panel, which I chaired, and discussions about what to test and how to test continued throughout the conference and in follow-up discussions. The conference was ideal for this because it was small (about 100 participants) with experts who came by invitation to discuss technical, organizational, economic, legal, and educational issues. Testing matters well beyond purely technical functionality. It has clear economic and organizational consequences. Assumptions about legal issues need testing and educational testing is a well-established if problematic theme. This column will focus mainly on technical and economic issues.

What is Testing?

All digital preservation systems have presumably undergone some degree of internal testing both before the system went live and afterwards to watch for problems or to make sure that updates functioned properly. This level of testing can be taken as a given. The more important type of testing takes place in public. Let me suggest the following definitions to clarify the discussion:

> **Public test:** a public test is one that a) people not involved in the original can replicate and b) whose results have been published openly in public journals or websites.
> **Private test**: a private test is one that a) cannot be replicated without the assistance of the original partners and b) whose results are available only to a closed audience.
> **Semi-public test**: a semi-public test is one that a) can only be replicated with the assistance of the original partners and b) whose results have been published openly in public journals or websites.

The concept of a semi-public test is necessary since some digital archiving systems have only a single site or have proprietary restrictions that mean the vendor must always assist in any test. While this goes against the scientific principle of truly independent replication, it need not mean that the test is carried out with less rigor or that the results are false. The reality in the digital archiving world is that many systems are vendor driven. The chief exceptions are LOCKSS

(Lots of Copies Keep Stuff Safe from Stanford) [2] and DAITSS (Dark Archive in the Sunshine State from the Florida Center for Library Automation). [3]

Testing Literature

Public testing

Exactly how much public or semi-public testing has taken place is hard to measure, even for someone who follows the digital preservation and archiving literature relatively closely. This is because test information may exist almost invisibly in the context of some other topic. An example of this can be found in the paper by Stephen Abrams et al. (2009) on JHOVE2, where the authors strongly imply testing that certainly took place:

> “However, its [JHOVE’s] extensive use over the past four years has revealed a number of limitations imposed by idiosyncrasies of design and implementation.”

Perhaps the test results were not given because no formal test plan or scenario existed and the results grew from a conviction among those using JHOVE that improvement was needed. In the library and archiving culture this has been acceptable, but might not be in a stricter environment. Nonetheless it seems likely that others could replicate the tests and get similarly problematic results.

A plainer example of public testing comes from David Rosenthal et al. (2005) on transparent format migration where the test is explicitly described:

> “To confirm the feasibility of this design, a proof-of-concept was implemented and tested. We chose an "obsolete" format widely used in actual content collected by the production LOCKSS system, and a suitable "current" format to replace it. … We did not implement the full **Mime-Type** matching process, but rather a configuration option in the LOCKSS proxy Web server that prevented **image/gif** from matching any **Accept:** header. The mismatch triggered a GIF-to-PNG conversion directly, delivering the content converted to PNG at the original URL but with **Mime-Type=image/png**.”

Since LOCKSS is open source, anyone who wants to reproduce the experiment could set up a system and repeat these procedures. The process is fully transparent.

Several other public tests are available as well. Ex Libris conducted a scaling test for the Church of Jesus Christ of Latter-day Saints. They explained the methodology, which involved “synthetic” data in four sizes -- 10 KB, 100 KB, 10 MB and 500 MB -- and various tuning measures such as spreading the Rosetta shares “across multiple cabinet arrays to avoid read/write contention” and increasing the “application server’s JVM heap size” (Ex Libris, 2010). Selected results were presented in figure 3 in the Ex Libris paper. From the description it seems likely that other Rosetta customers could undertake a similar test.

In my last several “Archiving in the Networked World” columns (Seadle, 2011a & b), I also ran tests against publicly available data sets from LOCKSS, Portico, and DOAJ (Directory of Open Access Journals). In each of the columns I included the Perl programs used to analyse the data. Anyone who wants to can reproduce the results. This kind of open testing is possible, but librarians and others engaged in long term digital preservation need to demand it before it will become widespread.

Semi-Public testing

Auditing processes represent a form of semi-public testing in which the results are made public, but the audit process is not readily replicable without active assistance from the original partner. The Center for Research Libraries did an audit of Portico in 2010, the results of which are publicly available. The audit used criteria based on the TRAC (Trustworthy Repositories Audit and Certification) checklist and appears to have concentrated on documentary information rather than actual live tests of the system. Documentary evidence works well for some subjects, but probably does not actually test the capabilities of a computer-based system. The report does not say exactly what documents were examined or what further tests might have been done. In that sense, the audit is more like a private test that no one outside of Portico and the audit team could reproduce.

Another digitally more invasive audit-type analysis is the SafeArchive system that does automatic verification of TRAC-based criteria. The Digital Preservation Alliance for the Social Sciences (Data-PASS) developed the system “to create a virtual overlay network on top of a peer-to-peer replication network, to support provisioning, monitoring, and TRAC-based auditing.” (Altman, 2011). The developers have been working with the Stanford LOCKSS staff to test the system, which would then audit Private LOCKSS Networks (PLN). No test results are yet publicly available to the best of my knowledge, but LOCKSS has a long history of public testing and results likely will be published. Nonetheless an audit of a specific PLN could be done only with the cooperation of that PLN.

Metrics and Rules

Some of the library literature involving public and semi-public tests provides metrics that could be used for comparison. An excellent example is the scaling test from Ex Libris. Even proof of concept tests (for migration, for example) can offer quantifiable measures of success. While it is an important first step to know that an archive is capable of migrating data in format A to format B, it would be equally useful to know what rules or measures of availability triggered the migration, how often the need for migration was tested, and how long it took after the trigger event actually to implement the migration. In terms of metrics the preservation and archiving world is at the point of saying “the car runs” without being able to give mileage, maintenance, crash/safety data, ownership costs, or replacement statistics without which a well-informed consumer would not (should not) buy a car.

It is easy to ask for such statistics, of course, and much harder to provide them or even to know what the digital archiving equivalent to such consumer statistics might be. Looking at the Ex Libris scaling test gives some sense of the complexity of the issues. The company created its own test data and loaded the data from a closely attached storage device. These decisions made sense in that they help to standardize the analysis, but such data and circumstances may not (like the US automobile mileage tests) necessarily reflect real experience. It is also not certain that Ex Libris would make the data that they used available to other systems (though they might) – or even that these data would make sense for other tests.

Questions about how much data an archive system can handle come up regularly. These capacity questions have at least two key aspects: how long the ingest process will take and how much data can a system manage. Measures for these two aspects are quite different. Benchmark levels for the ingest process are theoretically possible to set in terms of hours per gigabyte, but benchmarks for data quantity are harder because the potential quantities of data and the upper limits of performance for storage systems are both moving targets, and because of economic issues involving large-scale tests.

The following sections will consider possible metrics and ways of doing public or semi-public testing for three specific issues: 1) triggering migration; 2) ingest rates; and 3) storage capacity measurement.

Triggering Migration

The ability to migrate digital content from one format to another is virtually synonymous with digital preservation in the minds of many librarians, but trigger events are rarely discussed. Nonetheless some extremes seem obvious. When a proprietary word processing system like Wordstar ceases to be sold and the operating environment in which it ran ceases to be used, the optimal trigger date for migration presumably lies in the past. This is a problem. The situation for older versions of Microsoft Word is more complicated, because newer versions of Word continue to open (and automatically migrate) older versions, though only for a few versions back. Open Office can open earlier versions of MS-Word than some newer releases of MS-Word. Should this retard a migration trigger date? The digital archiving world has no basis for answering.

Systems like Wordstar, MS-Word and Open Office are not publishing platforms in the sense of producing files designed for public sharing. They are text editing and preparation systems for genuine publishing formats such as (originally) paper and (increasingly today) PDF. A reasonable argument can be made that the word-processing versions are throw-away drafts, but for those who want to keep such drafts, a reasonable event for triggering migration could be the last date of sale for a particular version. That date is a publicly available and generally coincides with the release of a new version. The problem with this date as a trigger mechanism is that waiting does no immediate harm. If archives had data about the average duration of format usability after a version change, they could judge the window of opportunity for migration better.

A more urgent situation involves digital content published on the internet. A PDF file that is a single self-contained bitstream is the simplest example, though embedded links add complexity. The more problematic case involves content that may be marked up in one format (say HTML) and formatted visually in another (say CSS) with javascript commands and with inserts such as images or video files that may exist on another server and may themselves have various formats. The browser further complicates internet-based migration issues, because it is ultimately that piece of software's ability to render content that signals when migration is either unneeded or past due. Here the rules for determining when to migrate depend not on a single program reading a single format, but on a range of programs and extensions (e.g.plug-ins) to programs, as well as on the addressing system (the URLs) binding them all together.

Fortunately the HTTP protocol includes a list of acceptable mime-types that tells a server what formats the browser can accept. It should be possible for an archive to test common browser types regularly to find what mime-types are requested and to test how often servers fail to deliver them. Failures that affect formats in the archive could act as a signal for implementing migration, but failure in one browser might not matter immediately if other more popular browsers continued to work.

Browser testing is only the beginning. Plug-ins, apps, and other extensions need testing too, if an archive wants to be sure that content remains usable in its original way. The key metric may involve not a single event (such as a change of versions) but a chain of online tests that systematically check both browsers and their extensions to detect the first point of dysfunction. As long as everything works – and thus far in the http-based web virtually everything continues to work – then migration is unnecessary. Once common browsers and their extensions begin to fail, the need for migration will be greater. Again, data about mean-failure times is needed before benchmarks for migration urgency can be set.

Ingest Rates

Because of recent conversations with publishers, I am aware that some of them have noticed a problem with the ability of archiving systems to ingest their content in a timely manner. I wrote about the problem in my last column (Seadle 2011a), where I calculated that it would take 8 years for the eDepot to ingest all of the full text open access journals from the Directory of Open Access Journals. Some archives are commendably open about unprocessed content in their ingest queue: Portico, for example, is quite explicit in its titles list about what volumes remain to be added. The problem is highly variable. Really big publishersrarely have content in any immediate danger, so a few extra years for ingest may not matter for them. Works from small and economically less stable publishers may need more urgent archiving.

As the Ex Libris test showed, a software test with standard data is possible and an ingest rate in the form of the number of hours per gigabyte of data gives librarians metrics to compare systems, but a purely automated test of archiving software may tell less about the de facto ingest rate than statistics using the number of elapsed days from the point when a publisher makes content available to the point where the archive has actually stored the content – at least for any system whose processes are not 100% automated. This kind of beginning-to-end processing is what most publishers mean when they complain that ingest is slow, not just an automated segment in the middle.

Benchmarking using elapsed-days would require a clear common understanding about the starting point. Availability for LOCKSS, for example, might reasonably mean the point at which the publisher put the electronic manifest on their server to allow LOCKSS to crawl it for data, since the LOCKSS ingest process cannot begin before then and publishers are not uniformly prompt in mounting the manifest. But with LOCKSS some decisions about the structure of archival units and the preparation of plug-ins have already taken place before the manifest goes up. Publishers more typically deliver the contents on storage media and archives must delay comparable decisions until the delivery takes place. While agreement on comparable starting and ending points may not be simple, it should be doable with sufficient community pressure. Comparison data would then be available.

Storage Capacity

Storage capacity is an issue that comes up when librarians worry about the rate of growth of digital information and the ability of archiving systems to cope. The issue itself is in some sense puzzling because no major archiving system has had a problem with storage capacity. For librarians the ability to distribute the management of essentially unlimited amounts of data Google-style across a large number of servers seems not to count as a demonstration when applied to an architecturally similar system such as LOCKSS. The conceptual problem may come from their individual experience with computer storage filling up or a single server hitting a capacity limit.

The question of storage capacity came up recently in a private discussion with a capable and intelligent colleague who has, however, no computer science background, and wanted a demonstration that storage capacity of the German Private LOCKSS Network had no problematic limits. At the theoretical level such a demonstration could be a simple calculation of the maximum storage used for any LOCKSS box (server) times the number of possible internet-connected LOCKSS boxes divided by seven -- which is the number of servers needed to offer reasonable long term data integrity. (Note: some librarians believe that all data must be stored on each LOCKSS box, but that is true only in the condition that a network has just seven boxes.) The number of possible LOCKSS boxes in the world is limited by the number of available IP addresses, which is far from infinite but is certainly large enough to outstrip any immediately foreseeable demand for long term archival storage.

Unfortunately this theoretical answer seems never to satisfy librarians, who want a physical demonstration. But a demonstration of what amount of data? Hilbert and Lopez (2011) estimated that: “In 2007, humankind was able to store 2.9 × 10^{20} optimally compressed bytes...” (This is just under 300 exabytes.) Even though this number is doubtless a significant underestimate of the amount of data in the world today, it is well beyond the size for any reasonable actual test of an archiving system, but how much less is acceptable?

A physical test can be prohibitively expensive if the storage is not actually needed for production purposes. For capacity benchmarking to be testable, two conditions need to be met. The first is agreement on reasonable proxies for actual storage so that costs are manageable. The second is agreement on what reasonable size an archival system should be able to manage. The Library of Congress estimated in 2009 that “the approximate amount of our collections that are digitized and freely and publicly available on the Internet is about 74 terabytes. We can also say that we have about 15.3 million digital items online.” (Raymond, 2009) As a benchmark this could be a starting point.

Conclusion

Testing is key to making rational decisions about digital long term archiving systems, but establishing metrics and rules by which librarians can compare the results is far from from easy. Librarians should not, for example, judge the quality and reliability of an archive’s migration services merely on the fact that they exist, but on how effectively they function at the right time. An excellent migration tool that is implemented too late has failed. Many archives leave the migration decision to human judgment without any metrics to suggest when implementation should occur. Librarians think of themselves as reliable people, but even reliable people tend to put off activities that take time and effort and have nothing to show afterwards except that everything works as well as it did before.

Migration is only one example. This column looked at two other areas where issues about testing and metrics came up in recent months. Archives should be able to develop benchmarks for the speed of ingest. This requires agreement about common measures, but that will happen only when the community forces the various archive system providers to come to agreement, which they are unlikely to do so on their own, since they see themselves as competitors. Storage capacity is another area where librarians are raising questions, but testing storage capacity involves non-trivial economic demands. Without proxy measures and agreement about how much data is enough to certify sufficient capacity such testing cannot proceed.

Testing, benchmarking, and developing metrics are different than establishing standards. Standards tend to be ex cathedra agreements that often have relatively little data behind them. As one colleague at the ANADP conference said: we have too many standards. What he might also have said is that we have too little data on which to make decisions about long term digital archiving, and as long as that is the case, our decisions may well be flawed.

Notes

[1] Aligning National Approaches to Digital Preservation conference: http://www.educopia.org/events
[2] LOCKSS: http://www.lockss.org/lockss/Home
[3] DAITSS: http://fclaweb.fcla.edu/FDA_landing_page

References

Abrams, Stephen, Sheila Morrissey, and Tom Cramer (2009), “What? So What”: The Next-Generation JHOVE2 Architecture for Format-Aware Characterization, in *The International Journal of Digital Curation*, Vol. 4, no. 3. Availabe (June 2011): http://www.ijdc.net/index.php/ijdc/article/view/139

Altman, Micah, and Jonathan Crabtree (2011), “Using the SafeArchive System: TRAC-Based Auditing of LOCKSS”, Forthcoming in the proceedings of the Archiving Conference 2011. TO be available at: http://www.imaging.org/ist/conferences/archiving/

Center for Research Libraries (2010), Portico Audit Report. Available (June 2011): http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/portico

Hilbert, Martin and Priscila Lopez (2011), “The World’s Technological Capacity to Store, Communicate, and Compute Information”, Science, Vol. 332, No 6025, pp. 60-65.

Raymond, Matt (2009), “How 'big' is the Library of Congress?”, Library of Congress Blog. Available (June 2011): http://blogs.loc.gov/loc/2009/02/how-big-is-the-library-of-congress/

Rosenthal, David S. H., Thomas Lipkis, Thomas S. Robertson, and Seth Morabito, (2005), Transparent Format Migration of Preserved Web Content, in D-Lib Magazine, Vol. 11, No. 1, Available (June 2011): http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Seadle, Michael (2011a), Archiving in the Networked World: Open Access Journals, in Library Hi Tech, Vol. 29, No. 2. Available (June 2011): http://www.emeraldinsight.com/

Seadle, Michael (2011b), Archiving in the Networked World: By the Numbers, in Library Hi Tech, Vol. 29, No. 1. Available (June 2011): http://www.emeraldinsight.com/

Ex Libris (2010), The Ability to Preserve a Large Volume of Digital Assets: A Scaling Proof of Concept. Avialable (June 2011): http://www.exlibrisgroup.com/files/Products/Preservation/RosettaScalingProofofConcept.pdf