The APTrust Story

a collaborative model for digital preservation

The Academic Preservation Trust (APTrust) is an innovative consortium committed to the creation and management of a sustainable environment for digital preservation and aggregate repository services for academic and research content. This paper responds to a number of questions that were voiced when APTrust was established: What led to the formation of APTrust? What does APTrust hope to accomplish? Most important, why should it be a top priority for academic libraries? After considering these questions and more, founding partners of APTrust agreed that building a robust and sustainable solution for digital preservation is the greatest challenge facing research libraries and their parent institutions today. This is the story of a collaboration that intends to rise to that challenge.

Martha Sites
Deputy University Librarian, U.Va.
Executive Lead, APTrust

4/22/13

The Problem

"The scholarship that is being produced today, both print and digital, is at serious risk of being lost *forever* to future generations." James Hilton, Vice President & Chief Information Office at the University of Virginia (U.Va.), has often made this bold statement. The current lack of a robust and sustainable solution for digital preservation makes this the greatest challenge facing research libraries and their parent institutions today. Hilton's straightforward and simple statement is a stark reality, and it should be particularly unsettling to librarians who believe that preservation of information published in all media and formats is central to the mission of libraries. So, while all information may not be worth preserving, I believe we must ensure that important, rare, and unique scholarly content is accessible in the future.

What We Know

Currently, universities don't have a cost-effective, fail-safe way to preserve the scholarly record or the content needed to produce new scholarship. Many institutions do have preservation repositories, but they are at risk of being single points of failure.

Early producers of digital scholarship and research at U.Va. began generating scholarship in digital formats of all types as far back as 1992. Now, some twenty years later, when the importance of preserving this content is crucial, few institutions have adequate long-term preservation strategies in place, and academic institutions especially struggle with how to sustain complex digital works. In recent years special collections libraries have begun to face new challenges as they acquire born-digital "papers" – the unedited video files that eventually form the re-election commercial of a senator, the progression of an author's novel as revealed by multiple versions of word processing files, or the time-based media projects of artists. These collections of contemporary scholarly, literary, and political figures and organizations need a preservation solution. Alongside these special collections, libraries want to acquire significant and wide-ranging academic content that is produced in digital formats: born-digital publications, electronic theses & dissertations, electronic records, social media, on-line course modules, complex mixed-media scholarship, and more recently, "big data." All these mediums are bringing additional complexity to preservation strategies and result in important original content that has no path to preservation. The explosion of "all things digital" makes it impossible to continue to ignore this common problem. It begs the question: how will this content be preserved?

Possible Answers

Many people, including most of the scholars who create and use these resources, assume that someone is already solving the digital preservation issues for academic content. After all, there are some well-known players in this space that appear to have things under control. What role do they have in this ecosystem?

CLOCKSS and Portico? Both CLOCKSS and Portico have well-established preservation services and thus it is reasonable to expect that all ejournals are preserved. However, a joint study of ejournal preservation coverage in 2010-11 conducted by Cornell and Columbia University Libraries determined that only 15-20% of their ejournals were preserved by the joint services of CLOCKSS and Portico.[1] A similar look at U.Va.'s ejournal subscriptions revealed that only 23% of U.Va.'s titles are actually preserved. So, while both CLOCKSS and Portico are valuable players in the preservation space, neither can yet guarantee that the majority of ejournals will be preserved long-term. We applaud the efforts of these entities to provide early solutions to preservation challenges for journals, and we need to work with them to expand ejournal preservation going forward.

Publishers? Some people assume that publishers themselves archive their holdings and preserve them forever. Relatively few publishers see preservation as part of their mission. In March 12, 2012 Robert Boissy posted a comment to a blog post on *The Scholarly Kitchen* in which he said "We [Springer] have taken the time and effort necessary to digitize our journal content back to volume 1, issue 1 wherever possible, and are currently undertaking high quality digitization of as much of our book content going back to 1840 as we can find."[2] Springer has one of the more progressive attitudes about digital preservation of journals, and they have participated in digital preservation efforts with LOCKSS, CLOCKSS, and Portico from their beginnings. While some publishers have become better stewards of their content, Boissy goes on to suggest that partnerships with libraries for this complex task is the most appropriate path moving forward: "We relied solely on libraries to store, preserve, and circulate our publications in the long age of print, but publishers like Springer have become archive partners with libraries, and perhaps better stewards, in the digital era."[3] Publishers are important producers of scholarly content, and we need to continue to engage them in long-term preservation strategies. They cannot and do not plan to do it alone.

Google? Google attempts to organize the world's information and make it universally accessible and useful. This mission is clearly stated and repeated throughout their site (see http://www.google.com/about/company/). While they are actively engaged in preservation activities such as the Endangered Languages Project, they make it clear that the long-term goal is for true experts (in this case, in the field of language preservation) to take the lead. Other similar efforts may emerge, but Google's primary mission is not likely to include preservation. They will be invaluable in providing technological solutions to esoteric, endangered types of content, and it is prudent for libraries to continue to join Google in exploring common solutions as another potential contributor in this space.

Other Obstacles and Vulnerabilities

While technological progress has enabled the explosion of digital resources and is mature enough to support digital preservation, there are still significant barriers that add to the complexity of fashioning digital preservation solutions. They include the non-trivial legal issues of intellectual property ownership, rights management, and risk management. The ongoing tension around ownership of publications and orphan works, in particular, came to a head in 2012 when the Authors Guild filed a lawsuit to protect authors' rights. In response to the lawsuit, HathiTrust staunchly defended its mission to preserve digitized content and make it accessible, and fortunately for libraries, the court ruled in HathiTrust's favor.[4] Many legal challenges continue to abound even though scholars and libraries wish for simple answers to what can be digitized and made available under current laws. The noble act of making copies for preservation purposes alone will still come into question by those with monetary and ownership concerns, and the differing perspectives of authors and libraries will likely be a cause of conflict for some time to come.

Natural disasters bring another kind of serious threat. In recent history, many watched with dismay as Katrina devastated the New Orleans area in 2005. Among the ruins was the devastation of libraries (and their collections) in Katrina's path.[5] Add to the physical damage the loss of hardware and software that provided access to the digital content of those libraries, and you can begin to appreciate the inability to completely recover from situations like this. Recovery is made even more difficult because the loss of library content must rightly be weighed against more immediate needs caused by the destruction of much of the city and surrounding areas.

An equally unpredictable but serious threat comes from potential political upheaval. A well-known political vulnerability was described by Preston Chesser in *The Burning of the Library of*

Alexandria where he wrote, “The loss of the ancient world's single greatest archive of knowledge, the Library of Alexandria, has been lamented for ages.”[6] That catastrophe occurred in the third century, and people assumed it would never happen again. Unfortunately, in January 2011, the Bibliotheca Alexandrina faced another major threat during the Egyptian Revolution when they helped provide young people with access to computers and social media.[7] As a result, access to the Library was completely shut off for days, and it was not clear whether there would be another mass destruction of this historic Library. Fortunately, another catastrophe did not occur, but in a world of unrest and impending wars, there is no certainty that it won’t.

Less dramatic situations than natural or political upheaval can also put durability and sustainability of digital content at risk. Consider the inordinate number of failures among “dot-com” companies and other hardware and software producers over the past decade. Such failures especially impact digitally produced scholarship that is dependent on specific hardware or software. While those dependencies create challenges for digital preservation in the first place, the disappearance of the companies that created and supported the hardware and software require other solutions to be found for the delivery of that scholarship. In some cases, it is impossible to find a solution.

Rising to the Challenge

The current state of preparedness for digital preservation is a recipe for disaster. When attempting to answer, “what can be done to change that?” it is quickly apparent there are no easy answers. Although libraries are stewards of academic content for the academy, and librarians have well-honed procedures and processes for preserving physical material, it is not so clear that librarians have mastered techniques required for the preservation of digital material. Individual libraries cannot solve these problems by themselves. Libraries must involve other libraries as well as technologists, lawyers, forensics experts, and scholars, to name a few. The challenge is to reach beyond our own organizations, to corral the smart minds at all our institutions, and to commit to finding common solutions.

In the Beginning

At U.Va., James Hilton, VP & CIO, and Karin Wittenborg, Dean of Libraries and University

Librarian, have discussed these issues for some time. They concluded that a national/international preservation effort would only be successful if it were a collaborative effort. In considering with others how libraries could respond to the digital preservation challenge, two initiatives emerged: Digital Preservation Network (DPN), and Academic Preservation Trust (APTrust). The focus of this paper is on APTrust, although the relationship between the two initiatives is important to a full understanding of digital preservation ecosystem that is evolving.

In August 2011, Hilton and Wittenborg convened a meeting at U.Va. to test whether the digital preservation challenge resonated with other academic libraries. The deans of libraries from Duke University, Emory University, Johns Hopkins University, University of Maryland, University of North Carolina, North Carolina State University, and University of Virginia came together to discuss ways in which we might align digital preservation efforts in order to

- digitally preserve the scholarly record in ways that would allow us to treat the digital copy as the copy of record;
- provide sustained funding for digital preservation of the scholarly record;
- collectively accomplish more than we might accomplish either alone or in smaller collaborations.

The group quickly coalesced around a common interest: the preservation of academic content. The use of a community approach to pursue this shared goal led to an agreement to create a consortium (APTrust) committed to creation and management of an aggregated preservation repository. Limiting this consortium to regional participation was thought to be shortsighted, so the initial group of seven partners decided to add five more like-minded partner institutions: Columbia, Michigan, Notre Dame, Stanford, and Syracuse. In order to remain nimble and ensure the ability to move quickly, the group decided to wait until an initial set of production services was in place before adding additional partners. To date APTrust has twelve institutional partners and will consider expanding membership by Summer 2013.

An important realization came from these initial discussions in that partners recognized the value of leveraging joint resources and dedicating time and energy to define common goals. We believed that building a community model would demonstrate how much could be accomplished by working together. Thus, with these principles in mind, the Academic Preservation Trust (APTrust) was born.

The Value of APTrust

In discussing what would make this initiative worth a commitment and fit with individual institutional priorities, the founding partners zeroed in on a number of potential benefits, including:

Long-term preservation: For all the reasons listed earlier in this paper, this is a no-brainer. Long-term preservation solutions are a top priority and are clearly identified as the core purpose for APTrust.

Aggregated content: The possibilities for future benefits are great if multiple institutions put their content into a single repository, make it accessible, and identify collaborative efforts around collections or formats to fashion specialized access or delivery methods.

Disaster Recovery: Ideally a disaster recovery solution would ensure not only that content is preserved but also offer a delivery service to ensure access to content until the affected institution had recovered capacity. Those interested in this service would like to see APTrust provide a hosting service option.

Community-building: The consortium is committed to collaboration that is more than simple cooperation. If we can bring our best collective minds together to solve problems, agree to work toward the common good, and make compromises to that end, there will be tremendous value added through scalable solutions, new functionality, and enriched services.

Economies of scale: Where it was once possible to define the value of a library by its individual investment in collections and total expenditures, those measures of success are being re-evaluated. As all academic libraries deal with escalating costs of journal subscriptions and growing financial demands for services, it is clear that the best solutions require more investment than most libraries can afford individually. Moreover, the depth of technical expertise required to address these challenges is beyond the ability of most libraries. By combining expertise and resources, we can build robust, scalable solutions together.

Articulating the value of this community endeavor was an exercise that demonstrated the solid alignment of partners and resulted in the APTrust value statement:

The Academic Preservation Trust (APTrust) consortium is committed to the creation and management of a preservation repository that will aggregate academic and research

> *content from many institutions. Solutions will be based on respected open-source technologies that are scalable, sustainable, and provide audit functionality.*
>
> *As part of a national strategy for long-term preservation, the APTrust repository will serve as a replicating node for the Digital Preservation Network (DPN). At the local level, APTrust will provide a preservation environment for participating members, including disaster recovery services. By leveraging the expertise and resources of multiple institutions, APTrust will realize economies of scale and increase value for all members.*
>
> *The consortium will work together to determine the shape of future services and best practices as they align around solutions for the common good. Ultimately, APTrust will enable academic libraries to protect the scholarship produced by the academy, a value that will transcend us all.*

The APTrust Organization

APTrust currently has 12 academic research partner libraries from across the country (see Figure 1: APTrust Founding Partners), representing both public and private institutions. An advisory group led by Pat Steele, Dean of Libraries at the University of Maryland, meets quarterly to determine priorities and set direction. U.Va. is leading the incubation of APTrust and is funding two software engineers and a Program Director until March 2014. In addition, other U.Va. staff are also contributing to the effort: Donna Tolson is interim Program Director, and several software developers are assisting with software development in the incubation phase of the project. In December 2012, Scott Turnbull became the Technical Lead for APTrust, and, as of February 2013, searches are underway for two software developers as well as a full-time APTrust Program Director.

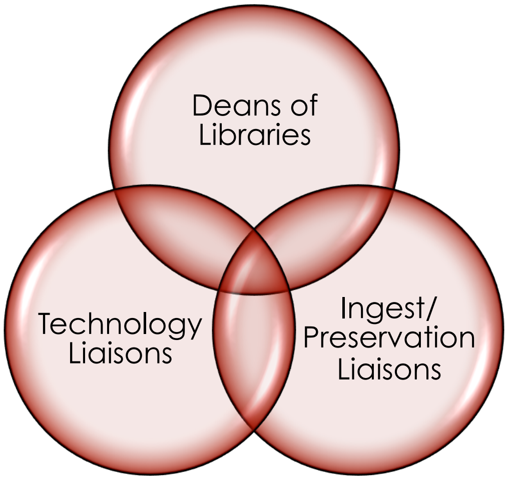
One of the valuable aspects of APTrust is a unique collaboration (see Figure 2: A Unique Collaboration) between Deans of Libraries, technologists (including CIO's), and ingest/preservation specialists. Each partner institution has identified liaisons for each of these roles, and they are invited to actively participate in shaping APTrust. The technology framework, for example, is being built so that partners will, in future, be able to contribute enhanced technology services where that is desirable. Partners recommend policy and content considerations.

The cost model for APTrust is currently built around a single annual fee of $20,000 per institution. Existing partners understand that their investment is for building a set of basic production preservation services. That single fee model will be used for at least another 1-2 years. As production services begin in January 2014, data will be gathered and used to determine the best ongoing pricing model.

Figure 1: APTrust Founding Partners

Columbia University
Duke University
Emory University
Johns Hopkins University
University of Maryland
University of Michigan
University of North Carolina
University of Notre Dame
North Carolina State University
Stanford University
Syracuse University
University of Virginia

Figure 2: A Unique Collaboration



APTrust Services

The University of Virginia Library is currently leading the development of APTrust services: 1) an aggregate repository for long-term preservation of content in all formats; 2) a replicating node for the Digital Preservation Network (DPN); and 3) future services such as advanced disaster recovery, format migration, access services, or hosted repository services.

1. Aggregate Repository

APTrust's commitment is to provide an aggregate repository that is comprehensive, resilient, retrievable, and scalable. The architecture (see Figure 3: AGGREGATE REPOSITORY ARCHITECTURE) was designed by the U.Va. technologists along with DuraSpace, who joined the U.Va. team in 2012. The Fedora-based aggregate repository will collect all forms of content, support administrative access, augment the preservation strategies of individual institutions, and provide a foundation for exploring future access services. The infrastructure is built in the Cloud using DuraCloud technologies that include Cloudsync and ReplicationTaskSuite to stage content from Fedora and DSpace instances, respectively. While the repository itself is Fedora-based, content can be ingested from other repository types. So far, ingest of Fedora and DSpace content has been tested, and tests for ingest of general file space content will soon follow. Other content types will be accommodated as demand for them is identified. To date, the majority of effort has been spent on creating the infrastructure and workflows for the aggregate repository with an expectation that fully supported production services will be available by January 2014. Ownership of the APTrust repository will ultimately reside with APTrust.

Each institution will determine its own content strategy and will be responsible for assuring permissions for sharing or viewing the content. APTrust repository services will provide rights management at the object level for permissions, and for identifying whether the content is destined for long-term storage in the DPN replicating node.

In order to understand how APTrust fits into the preservation ecosystem, it is helpful to consider that content moving through the scholarly process from conception to long-term preservation, is naturally winnowed. For librarians who curate these collections, different levels of preservation are available for different content (see Figure 4: Winnowing of Content). APTrust will provide a secure preservation environment that connects easily to institutional repositories and to DPN.

There may be widely different philosophies about how to make content selection choices, especially as processes are formalized; what is important to one institution may not be as important to another. Therefore, partner ingest/preservation liaisons will determine best practices for content decisions, and individual institutions will ultimately decide how their content will be handled. Once stable production services are in place, APTrust will be able to connect interested partners who want to leverage common interest in aggregated content.

FIGURE 3: AGGREGATE REPOSITORY ARCHITECTURE

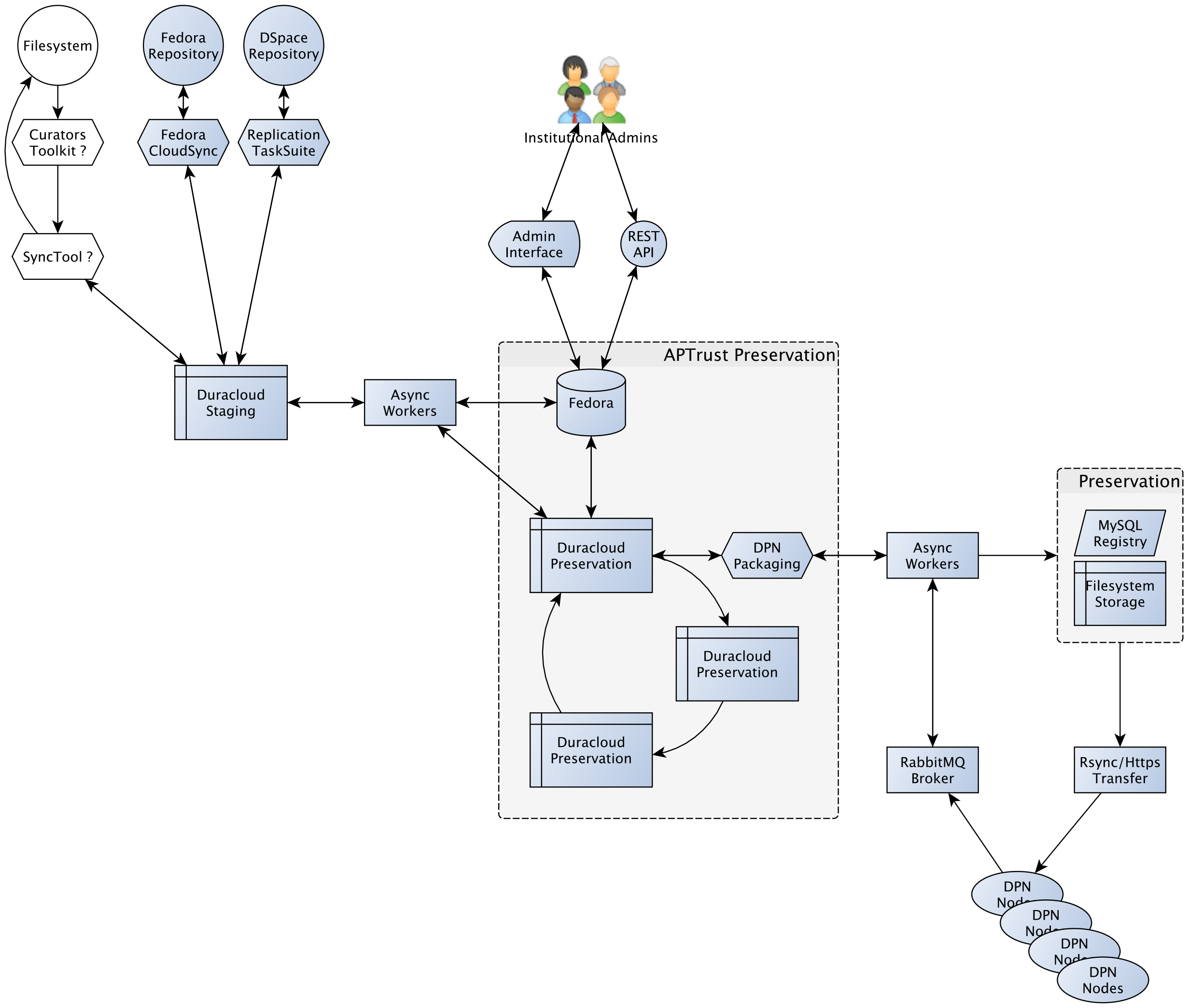
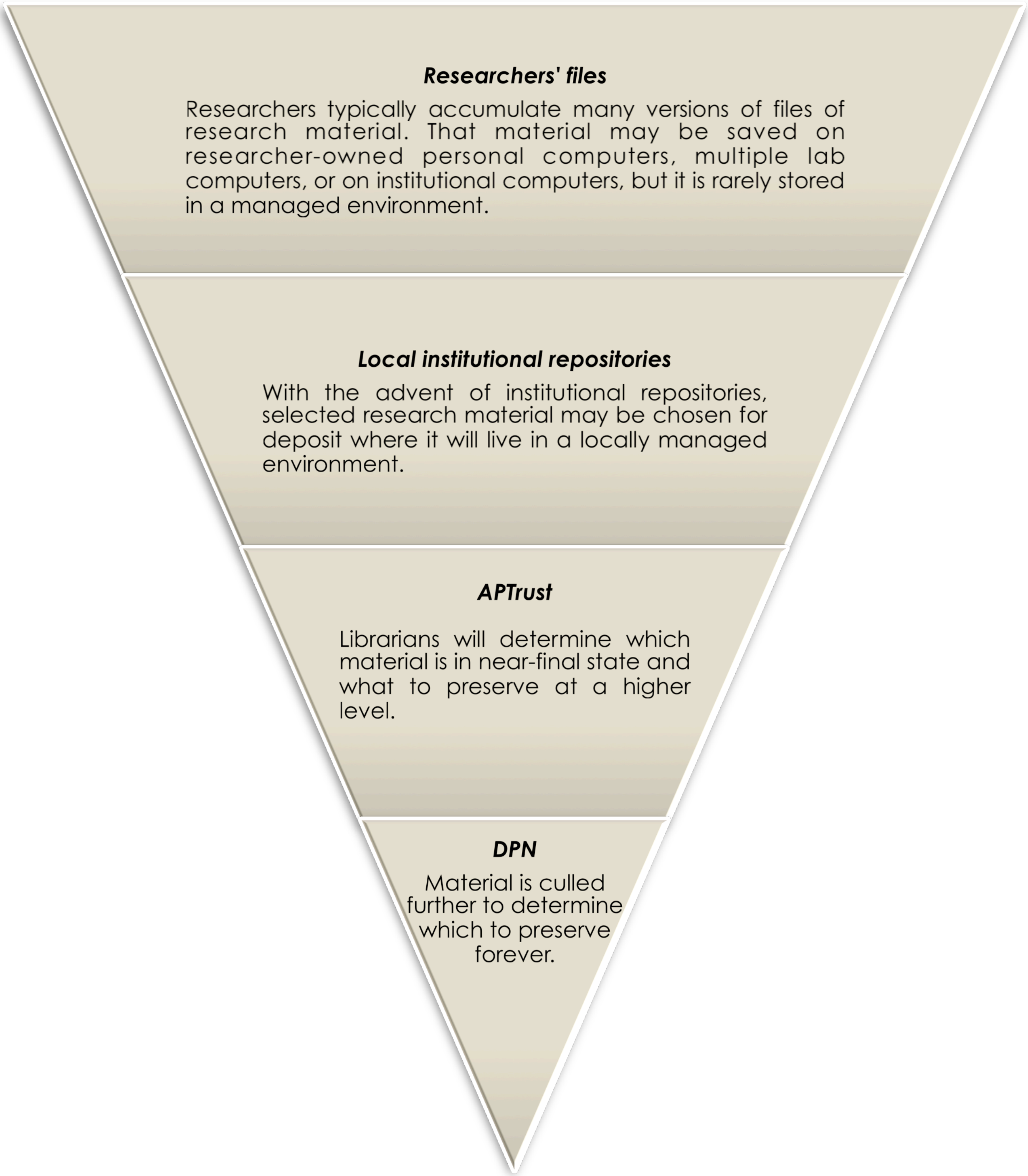


Figure 4: Winnowing of Content



Researchers' files

Researchers typically accumulate many versions of files of research material. That material may be saved on researcher-owned personal computers, multiple lab computers, or on institutional computers, but it is rarely stored in a managed environment.

Local institutional repositories

With the advent of institutional repositories, selected research material may be chosen for deposit where it will live in a locally managed environment.

APTrust

Librarians will determine which material is in near-final state and what to preserve at a higher level.

DPN

Material is culled further to determine which to preserve forever.

2. DPN Replicating Node

APTrust's commitment is to provide a Digital Preservation Network (DPN) replicating node that is part of a federated network that is flexible, redundant, and secure. APTrust will adhere to the principle aspirations of DPN by providing a path for content destined for long-term preservation and by replicating content from other DPN nodes. Currently, DPN is a framework of five nodes that will replicate content across a minimum of three of those nodes. DPN's federated solution ensures that academic content is preserved even if one or more of the nodes fail.

To help people understand the difference between APTrust and DPN, we have identified three areas with distinctive characteristics: purpose, organizational models, and type of protection.[8]

First, while both are preservation environments, they differ in purpose. While DPN provides permanent "dark archive" storage, APTrust allows content to be altered and removed. Two simple analogies have proven useful in describing the difference in purpose between DPN and APTrust.[9] Each of these is useful for different audiences:

> THE INSURANCE ANALOGY APTrust is like having insurance on your house or car; if it has to be used, the benefit comes to you and you use it. It could also be likened to health insurance in that you use it to ensure the overall health of the content while it is being actively curated. DPN is like life insurance; you don't expect to receive the benefit, but you expect others to receive it (and make use of it). Your ancestors will benefit. Both APTrust and DPN help create peace of mind, but in very different ways.
>
> THE KITCHEN ANALOGY The institutional repository is the shelf; things on the shelf are freely available, but vulnerable to loss. APTrust is the refrigerator; items stored there are preserved for a long time, but not forever. You can take things in and out, but it's more trouble than if it's on the shelf. DPN is the deep freeze; there, the shelf life of items is preserved forever, or as close as we can come to that. You take it out ONLY when it exists nowhere else.

Second, there are different organizational models:

> APTrust allows (and encourages) direct interaction with contributing institutions; partner institutions will have administrative access to their content and the potential for more, depending on service development. Partners will also choose and affect the direction of development.

DPN allows interaction ONLY through the nodes. At this time institutions cannot directly access DPN, and it is not a contributory design. Partners support the concept and influence governance, but only designated technical leads contribute to design development.

Third, APTrust and DPN protect against different kinds of threats:

APTrust guards content against something unintended, or unforeseen happening in the world around us (e.g., natural disaster, human error, degradation of access content, technology failure, etc.) - situations where the content owner wants to, and eventually can, restore the content so that it can be used. Also, APTrust allows content to be removed if the content owner wants to withdraw and no longer pay for the service.

In contrast, DPN guards content against a negative (whether intended or not) consequence that may directly impact the content owner (e.g., hostile takeover, political upheaval, apocalyptic scenarios). It protects against situations where the content owner no longer wants to make it available to the world, or is permanently wiped out and thus unable to restore content. Once submitted to DPN, content cannot be withdrawn, even if the content owner no longer pays.

3. Future proposed services

APTrust's commitment is to provide future services that are innovative, creative, collaborative, and inclusive. APTrust will leverage the rich, aggregated collections of partners with services that utilize a reliable, flexible, and scalable infrastructure and are developed collaboratively by APTrust staff and the entire APTrust partner community. Discussions about priorities for future services have just begun, and decisions about those priorities have not been made yet. A glimpse at the range of discussions includes primary interest in access and delivery of aggregated content, format migration, disaster recovery, and hosted repositories. None of these services is trivial to implement; therefore, it is not feasible to offer them all immediately. We will, however, develop further functional requirements and timetables for moving the top priority services into production. The APTrust Advisory group and the APTrust staff will make decisions, based on partner feedback and business case considerations.

What's in it for academic libraries?

In a recent EDUCAUSE Review article entitled *The Game Has Changed,* Chuck Henry and Brad Wheeler wrote: "It is becoming increasingly clear that neither the *challenges* that confront colleges and universities nor the *solutions* to those challenges are unique to each institution."[10] Long-term preservation is one of those *challenges*. APTrust is one possible *solution* to that challenge.

APTrust is a unique collaboration that brings together Deans of libraries, technology liaisons, and ingest/preservation liaisons with other stakeholders from within and outside our institutions. The value of bringing these stakeholders together cannot be overstated. Many of us have come to understand the enormity of the challenge that fully scoped preservation services bring. While individual libraries may solve some of these challenges satisfactorily for the short term, APTrust partners believe that individual solutions will not scale and will not provide the level of assurance for long-term preservation that is ultimately needed. Engaging stakeholders who bring different knowledge and perspectives from across institutions benefits individual institutions by informing their policy, financial, and technical requirements as well as effecting changes that will help shape the future of scholarship produced by the academy.

Conclusion

It is impressive to know that U.Va.'s founder, Thomas Jefferson, spoke about preservation as early as February 18, 1791, when he wrote: "… let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident."[11] It is a reminder that preservation is not a new problem. The past informs the future. We don't have all the answers; we don't know all the problems; and we certainly don't know what we will encounter along the way. Nevertheless, we are committed to embarking on a path that will begin to solve the problems associated with preserving digital research & scholarship and the content needed to produce it. Ultimately, we believe that both APTrust and DPN will enable academic libraries to protect the scholarship produced by the academy, and that is a value that will transcend us all. Our hope is to leverage the benefit of doing this together.

References

1. 2 CUL LOCKSS Assessment Team. 2CUL LOCKSS Assessment Study, p. 16. (public release Oct 2011) Retrieved Feb 10, 2013 from http://2cul.org/sites/default/files/2CULLOCKSSFinalReport.pdf

2. Boissy, Robert. *The Scholarly Kitchen.* (Mar12, 2012). Retrieved Feb 10, 2013 from blog post comment in response to a post by Rick Anderson, *E-journal Preservation and Archiving: Whether, How, Who, Which, Where, and When?* (Mar 7, 2012) http://scholarlykitchen.sspnet.org/2012/03/07/e-journal-preservation-and-archiving-whether-how-who-which-where-and-when/

3. Boissy, R. *The Scholarly Kitchen* (see reference 2.)

4. *HathiTrust Statement on Authors Guild v. HathiTrust Ruling,* Press release (Oct 12, 2012). Retrieved Mar 5, 2013 from http://www.hathitrust.org/authors_guild_lawsuit_ruling

5. Corrigan, Andy, Associate Dean of Libraries. *Hurricane Katrina and the Library's Collections.* Retrieved on Apr 22, 2013 *from* http://library.tulane.edu/collections/katrina_recovery

6. Chesser, Preston. *The Burning of the Library of Alexandria* (Jun 1, 2002). Retrieved Feb 10, 2013 from http://ehistory.osu.edu/world/articles/articleview.cfm?aid=9

7. Serageldin, Ismail, Librarian of Alexandria, Director of the Bibliotheca Alexandrina. *To all Our Friends Around the World:18 Days that Shook the World.* Press Release (Feb 12, 2012). Retrieved Jun 2, 2012 from http://www.bibalex.org/News/NewsDetails_en.aspx?id=3133

8. Tolson, Donna. (personal communication, Dec 7, 2012)

9. Soroka, Adam. (personal communication, Dec 7, 2012)

10. Henry, Chuck & Wheeler, Brad. *The Game Has Changed, EDUCAUSE Review,* (March/April 2012), pp 58-59

11. Jefferson, Thomas. *Letter To Ebenezer Hazard* Philadelphia Feb. 18. 1791. Retrieved Apr 22, 2013 from the Papers of Thomas Jefferson, Digital Edition http://rotunda.upress.virginia.edu/founders/default.xqy?keys=TSJN-search-1-1&expandNote=on - match