THE IMPORTANCE OF TRUST IN DISTRIBUTED DIGITAL PRESERVATION: A CASE STUDY FROM THE METAARCHIVE COOPERATIVE

Matt Schultz
Educopia Institute
1230 Peachtree Street, Suite 1900
Atlanta, GA 30309

Emily B. Gore
Clemson University Libraries
CB 343001
Clemson, SC 29634-0001

ABSTRACT

Distributed digital preservation is a maturing and appealing solution to the pressing problem of ensuring the survivability of digital content. Like all other digital preservation efforts, distributed digital preservation solutions must communicate trust to their Designated Communities as they continue to mature. The following paper discusses the importance of establishing this trust, retraces the development of TRAC as a reliable tool for evaluating trustworthy repositories, and details the process of the MetaArchive Cooperative's application of TRAC to its distributed digital preservation solution. This process revealed that the current metrics for gauging trust in digital preservation could be readily applied to distributed solutions with great effect. However, because these metrics often presume a more centralized approach to preservation, the process also revealed the need to apply them carefully and with great thought. To underscore this need, three organizational and technical comparisons are made between the MetaArchive's distributed preservation activities and the more centralized model assumed by TRAC and the OAIS Reference Model. The paper concludes with the question as to whether distributed digital preservation needs to be better defined within existing models such as OAIS or through the creation of a new reference model for distributed digital preservation.

INTRODUCTION

Distributed digital preservation is a maturing solution to the pressing problem of ensuring that future generations will have access to digital content of scholarly, cultural, political, and scientific value. As framed in the recently published *A Guide to Distributed Digital Preservation*, "...a growing number of cultural memory organizations have now come to believe that the most effective digital preservation efforts in practice succeed through some strategy for distributing copies of content in secure, distributed locations over time." [13]

Indeed, many projects and service models are actively addressing the need for digital preservation in this geographically distributed fashion. Among these are LOCKSS (Lots of Copies Keep Stuff Safe) and Private LOCKSS Networks (PLNs) such as the MetaArchive Cooperative, ADPNet, PeDALS, and Data-PASS (to name just a few); data grid solutions such as Chronopolis; and cloud-based initiatives such as DuraCloud. These projects and services represent a strong approach that ensures that digital assets can survive well into the future in the face of such threats as natural disasters, human error, and technological obsolescence.

Just like the more centralized institutional or shared repository solutions that have comprised some of the early foundational efforts in the field of digital preservation at large, these distributed digital preservation efforts must focus attention on the issue of communicating trust to their Designated Communities as they continue to mature.

IMPORTANCE OF TRUST

Trust is defined as the "reliance on the integrity, strength, ability, and surety of a person or thing." [6] When establishing a preservation service model, especially one with a distributed membership, like the MetaArchive Cooperative and other distributed digital preservation efforts, it is important that trust be at the center. Members need to trust each other, trust the leadership, and trust the preservation system itself. Establishing and maintaining trust can be a daunting task even when colleagues and peers, as opposed to vendors, control, manage and maintain the network. A cooperative model is designed to be much like a democracy, where members take ownership and voice concerns, opinions and shape future directions.

In an interview published in 2000 in *RLG DigiNews*, Kevin Guthrie, then-President of JSTOR, indicates that establishing trust in 3rd party vendors is "important because the goal is to be able to establish a relationship whereby a library can rely on a third party to provide a service that has been a core function of a library; that is, archiving." [8] The MetaArchive Cooperative supports that belief and arguably enhances it by philosophically and practically striving to enable libraries to work collaboratively to archive their own materials in a trustworthy manner. The MetaArchive Cooperative (www.metaarchive.org) is a community-based network

that coordinates low-cost, high-impact distributed digital preservation services among cultural memory organizations, including libraries, research centres, and museums.

Cooperative, distributed digital preservation relationships may be favorable to individual institutions due to both the cost-effectiveness of the approach, which capitalizes on the existing infrastructures of cultural memory organizations rather than requiring the establishment of external services, and the implied sustainability of an alliance of institutions working together. If nothing else, the current economic situation has forced libraries to realize that content in "silo" repositories could be at greater risk as institutional priorities, funding streams, and the greater economy fluctuates. There is greater trust in at least the medium-term sustainability of collaborative efforts than in local efforts where the reduction or elimination of funding for one year can have dire consequences. In collaborative relationships, economic crises at one or two institutions have less of an impact on the collaboration as a whole.

When prospective members consider joining an organization like the MetaArchive Cooperative, trust is arguably the main element they are looking for – they are asking if they can trust the organization, the partners and the technology with the critical assets they are charged to manage for the long-term. In the paper *Creating Trust Relationships for Distributed Digital Preservation*, Walters and McDonald state that, "the concept of trust and its manifestation between institutions as an essential element in designing digital preservation systems − both technical and organizational − is critical and appears in the organizational level needs of the *CRL/RLG-NARA Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist*." [7]

TRAC

The origin of TRAC itself is in trust relationships and alliances among key organizations. The call for a "network of trusted archives" initially drove the creation of the trusted digital repositories concept as well as influenced the development of the *Reference Model for an Open Archival Information System* (OAIS). [2] As an OAIS-approved follow-on activity, TRAC and the actual metrics development also evolved through these same relationships. The RLG-NARA Task Force on Digital Repository Certification obtained valuable alliances with the then-new Digital Curation Centre, as well as colleagues in Germany directing the *nestor* project. A critical alliance with the Center for Research Libraries (CRL) also emerged. In 2005, the Center for Research Libraries was awarded a grant by the Andrew W. Mellon Foundation to develop the procedures and activities required to audit and certify digital archives. The CRL Certification of Digital Archives Project worked closely together with the RLG-NARA task force to redevelop the audit metrics and provided critical opportunities to develop and test the audit process itself. This practical testing, along with the DCC test audits that led to the development of DRAMBORA, contributed greatly to filling the gaps identified in the earlier draft, *Audit Checklist for the Certification of Trusted Digital Repositories*.

The final version of TRAC was published in February 2007 with 84 criteria broken out into three main sections: Organizational infrastructure; Digital object management; and Technologies, technical infrastructure, and security. It provides tools for the audit, assessment, and potential certification of digital repositories; establishes the documentation requirements for audit; delineates a process for certification; and establishes appropriate methodologies for determining the soundness and sustainability of digital repositories.

It currently serves as a de facto standard for repository audit and is being actively used by organizations as both a planning and self-assessment tool. Additionally, it continues to serve as the basis of further audit and certification work, including the National Science Foundation-funded CRL project, Long-Lived Digital Collections. [5]

METAARCHIVE COOPERATIVE SELF AUDIT

A recent effort has detailed for the larger community (including prospective and non-members) the organizational and technological trust foundations of one successful and growing distributed digital preservation solution. Between June and December 2009, the MetaArchive Cooperative worked with an outside evaluator to conduct a self-audit using the *Trusted Repositories Audit & Certification: Criteria & Checklist* (TRAC). [1] The Cooperative makes use of the LOCKSS (www.lockss.org) open source software to dark archive multi-format digital collections. Collections being preserved in the MetaArchive network include electronic theses, digitized photographs and manuscripts, websites, oral histories, and many others. This content is available to the content contributor alone in the event of catastrophic loss of its original content—thus enabling the retention and preservation of the many important works that cannot be openly shared at this time due to intellectual property and other concerns.

Self Audit Results

The results of the MetaArchive Cooperative's self-audit revealed that the MetaArchive conformed to and addressed the concerns of each of the 84 criteria specified within TRAC. As importantly the assessment helped to identify and prioritize at least 15 activities to be reviewed and/or enhanced over the course of 2010 and 2011. [12] The success of this process made it clear that current metrics for gauging trust in digital preservation could be readily applied to distributed solutions, it also underscored the need to apply them carefully and with great thought.

DISTRIBUTED SELF AUDIT METHODS

Assessing the MetaArchive Cooperative revealed that an evaluator at work in this distributed digital preservation

environment must be willing to invest a fair amount of time engaging with repository staff through a careful and synthesized analysis in at least three ways:

- The first of these involves systematically coming to grips with the design solution of the repository. This can be done through extensive reading of internal and published documentation and conducting multiple interviews with repository staff. Specifically, an evaluator must ask questions regarding how the repository is organized to effect preservation, and how the underlying technology both facilitates and constrains that organization appropriately.
- The second area of analysis involves comparing and contrasting this overview of the repository with the OAIS Reference Model, and its functional recommendations for building a trustworthy repository.
- Finally, the evaluator must grapple with the concerns embedded in TRAC itself, and ensure that in pursuing the objective of applying OAIS frameworks and definitions to a repository's activities, the evaluation fairly accomplishes its core goal: that of gauging genuine degrees of trust and best practice within the repository.

Though OAIS seeks to apply its functional elements in responsible ways to diffuse models such as those of federated repository endeavours, a centralized model for preservation is largely at focus in OAIS and TRAC. [4] This is no doubt because most digital preservation initiatives, even those, such as the Hathi Trust (http://www.hathitrust.org) that have pursued trustworthy federated approaches have tended to situate each of the OAIS functional elements and roles within single repository spaces for various administrative and technical reasons. For reasons of this precedent an evaluator of a distributed digital preservation network may be required to extrapolate out some of the OAIS Reference Model's elements when necessary and look for their representation across diffuse locations and multiple roles.

Drawing Fair Comparisons

Three examples that demonstrate the need for such extrapolations stand out from the MetaArchive Cooperative's self-audit.

- *Central vs. Distributed Infrastructure*: this first example sheds light on the importance of being able to draw some proper distinctions between a distributed digital preservation effort's network server environment and its web-like representation of a "repository", in contrast to the more unified and centrally housed infrastructure that tends to be standard to many other digital preservation solutions.
- *Push vs. Pull on Ingest*: this second example highlights the behaviour of the LOCKSS software and its "pull" scheme of ingesting submission information packages (SIPs), and constructively contrasting this with the typical "push" scheme facilitated by many repositories (electronic ETD submissions for institutional repositories as one example).
- *Dark Archiving & Designated Communities*: the third example involves properly addressing the OAIS Reference Model's notions of Access and Designated Communities (Producers/Consumers) in light of the MetaArchive's dark archive approach to bit-preservation and the format agnostic designations of LOCKSS.

Central vs. Distributed Infrastructure

Though the OAIS Reference Model and TRAC both acknowledge that there are multiple ways to organize a repository's infrastructure, the documents themselves overwhelmingly have related a more centralized approach to designing and operating a digital preservation solution. The MetaArchive Cooperative (along with other PLNs, Chronopolis, and other initiatives) has established a distributed network of linked servers that cooperate to mutually store, manage and refresh contributed content at the bit-level. This methodology holds that replications of content that are geographically distributed and maintained on multiple servers in highly secure networks stand the greatest chance of meeting the integrity and longevity standards that the cultural memory field must strive to achieve.

During the course of researching the organizational and functional design of the Cooperative for self-auditing purposes, it became clear that the conceptions of the more stationary and routine operations of a traditional archival "repository" in TRAC had to be mapped to an understanding of the more dynamic and automated changes of state that are inherent to the software operations of LOCKSS. Clarifying this distinction allowed for a proper response to a central concern within OAIS and TRAC: the fixity or integrity of the content.

LOCKSS, for example, engages in a vigilant, and automated process of verifying that the geographically dispersed copies that have been ingested from a content contributor's source are consistent with that source and with one another. It handles this through the use of a voting and polling scheme between the linked servers with mutual copies of content, and relies on temporary checksum comparisons. Indeed, LOCKSS distinguishes itself from perhaps more static repositories by actively anticipating the potential for corruptibility and has developed a recovery scheme in the face of such eventuality by first of all refusing to rely on long-term validation through the maintenance of checksums – which are themselves easily corruptible. [10] Rather it leverages the validation power of a network of redundant servers, and maintains an open re-ingest stream to the authoritative source, once corruption of a copy is detected.

This is quite different than running digests on a single copy of an ingested digital object as it resides or is migrated on disk/tape and then comparing its hash value to a previously generated checksum, which requires its

own set of long-term curatorial data management. This, latter scheme is encouraged by OAIS and TRAC in its prescriptions for content fixity, and is implemented and relied upon by many centralized repositories. Though the concern is one for the content's integrity, in and of itself, this approach often only alerts to the occurrence of file corruption, rather than going beyond this to trigger an automated assist in its diagnosis or recovery.

As an evaluator applying TRAC to the Cooperative, while at the same time trying to genuinely address the concern for the content's integrity that resides around this issue of fixity, it became clear through this careful comparison that the emphasis for this LOCKSS-based network needed to be directed differently. The emphasis needed to be placed less on managing and reporting on the veracity of the fixity data itself (though not unimportant), and more so on being able to report on the rate and nature of content repair and re-ingest, so that any disruptions to network activity could be more properly diagnosed and mitigated. To this end the central staff and membership of the MetaArchive Cooperative have begun experimenting with the rich information handling of the LOCKSS daemon in order to provide timely and actionable reports on the status of the network's operations. Progress on this front is being accomplished with great effect through integrations between the LOCKSS daemon and in-house data reporting tools developed by MetaArchive.

Push vs. Pull on Ingest

In many centralized repositories a content contributor is provided a submission pathway whereby they are charged with handing their digital object(s) off to repository specialists. This hand-off typically occurs in a format that can be easily managed or migrated by the repository for the sake of long-term preservation. Occasionally this places the content contributor in front of an access interface that will accept various user-generated metadata concerning the digital object(s), and a mechanism for uploading these objects, as Submission Information Packages (SIPs). At that point the repository takes over and shepherds the digital object(s) through a series of processes to prepare the objects for long-term storage, management, and dissemination. The pathway is thus a process of "pushing" content into an archive, which aligns quite comfortably with our unquestioned protocols for donating artefacts to traditional archives. It is also the process most visibly detailed within the OAIS Reference Model [3]—and even more so, in the cultural memory community's use and discussion of this model.

Distributed digital preservation solutions have often taken a "pull" approach that differs somewhat from this paradigm. The MetaArchive Cooperative (via LOCKSS and its web-crawl based ingest mechanism), and Chronopolis (via the use of "holey" BagIt files as one of several ingest mechanisms) are both examples of repositories that can be said to be using a "pull" scheme for obtaining digital objects.

Specifically for the MetaArchive this has meant that central repository staff must work in a coordinated fashion with content contributors to ensure that they have prepared their content in structured ways (referred to as 'data wrangling') to ensure a successful and on-going "pull" of their content into the preservation network. Once the content has been prepared this "pull" process is finalized by having a content contributor construct an XML plugin that enforces any inclusion/exclusion rules necessary to identify collection files as they reside on an active web server directory. This plugin is then used by the LOCKSS software to guide a web crawl and perform a harvest of the collection.

An evaluator applying the OAIS Reference Model and TRAC to this arrangement has to recognize and account for the way that various functional elements that would typically be reserved only for repository staff operating under a "push" system, namely the preparing of a SIP to become an Archival Information Package (AIP), need to be looked for in various ways on the side of the content contributors within a "pull" environment. This is because the content contributors take responsibility for preparing their own content for its ultimate preservation state by engaging in the "data wrangling" and defining of their collections for harvest. In the MetaArchive context, this has led to the development of documentation that more explicitly describes the MetaArchive network's expectations regarding content organization and the ingest procedures that contributors follow. This documentation is thus working to better define the functional point at which a SIP becomes an AIP, and the roles on both sides of the Cooperative community that bear the responsiblility for such transformations.

Dark Archiving & Designated Communities

Though the majority of digital preservation initiatives have linked the priorities of preservation and access quite closely, as in the case of institutional repositories, there are several examples of use cases that make immediate access to preserved materials a secondary priority. Dark archiving, which involves preserving materials for future use with no direct means of access from the repository, is an approach that has been attractive to those with content that needs to be preserved but that is not immediately or openly available for access. This has multiple permutations.

In the case of CLOCKSS (http://www.clockss.org/clockss/Home), publishers and libraries agree that a publisher should retain the authority to provide access to their electronic publications, but that libraries can assume this role under certain conditions. This requires that libraries preserve a copy and restrict access until such a defined "trigger event" has occurred – loss of a publisher or a title no longer being offered for example. Through the use of proxy mechanisms, the *end-user* of a journal's Designated Community may not even notice that the publisher's hosting has switched to that of the library

because LOCKSS caches at a library site collect and preserve the original journal content exactly as it was served from the publisher. The switch in many cases appears seamless.

The MetaArchive Cooperative has found the use of the LOCKSS software to be similarly useful for the dark archiving and bit-level preservation of their members' digital collections. As mentioned, a member can construct an XML plugin that enforces any inclusion/exclusion rules necessary to crawl collection files as they reside on an active web server directory. This plugin can then be used by the LOCKSS software to guide a web crawl, perform a harvest, and dark archive the collection on a separate, geographically dispersed server. For such members a copy is thus preserved in the event that the originating web server is unable to provide access to a content contributor for their own institutional purposes.

In the case of the MetaArchive, however, on-going and immediate access for a member's *end-user* Designated Community need not be the ultimate guiding priority. The MetaArchive has taken the de-prioritization of access a step further by avoiding the requirement that members select collections that are "dissemination worthy," or that lend themselves to any foreseeable exhibition and use. In fact, members have broad rights of selection when seeking to preserve their collections in the MetaArchive network. Not only is LOCKSS well designed for preserving content in reserve for future *end-user* access scenarios, but it is also format agnostic. This means that members can not only preserve normalized and derivative files that lend themselves nicely to our current notions of future 'readability' and 'understandabilty', but the original bit stream data, and even high quality master files that can be used for any future, as yet unknown, migration or emulation requirements. Under these terms the MetaArchive Cooperative has empowered its members to assume the curatorial responsibility for the decision-making surrounding the preservation of their collections, rather than requiring them to contribute only highly vetted, access-oriented collections in formats that are considered "manageable" by the repository.

A MetaArchive member enters into agreement with other members to mutually preserve one another's collections to guard against the all to too real threats of natural disaster, human error, and technological obsolescence. These are the "trigger events", and when they occur, a member may recover their collection intact from the network, where it has been both technically and legally shielded from any dissemination chain (including to those institutions that hold replicated copies of the content for preservation purposes). Under these terms, a MetaArchive member is the *end-user* for all intents and purposes, and is in a sense both a Producer and a Consumer in OAIS Reference Model terms.

When assessing such repository arrangements with auditing tools like TRAC it is vital that an evaluator be able to de-couple the notion of a Designated Community of Producers and Consumers from the OAIS Reference Model's emphasis on access and use. Though MetaArchive members may not hypothetically choose to preserve files and formats that satisfy our current notions of maintaining future 'readability' and 'understandability', they have been provided a preservation solution that grants them the flexibility to engage their collections on terms that are appropriate to their institutional priorities – which cannot be underestimated in a time when many cultural memory organizations find themselves contending with short-term limited resources but a desire to avoid outsourcing to multiple third party services, in the hopes that they can gradually build expertise and capacity in preservation.

Nevertheless, the concern with useable formats is a natural one, for which LOCKSS has sought to engage for the possible, but by no means impending, approach of widespread obsolescence. [9] [11] Nor is the MetaArchive opposed to monitoring the current and foreseeable usability of its members' collections. The Cooperative's members and central staff remain open to the potential long-term usefulness of the Unified Digital Formats Registry, and if called upon by its members, to exploring integrations with JHOVE2 and DROID, especially as tools that could enable the MetaArchive to communicate broadly to its membership the number and types of formats being preserved in the network, thereby further empowering them with the information they might need to effect preservation and access for their own Designated Communities as they define them.

CONCLUSION

In much the same way that centralized repositories have worked assiduously to prioritize trust as a guiding principle for design and management of their preservation solutions, the maturing field of distributed digital preservation must also communicate the trust relationships that are foundational for a responsible network. When using current tools to accomplish this aim--such as OAIS, TRAC, and successor tools such as the *Metrics for Digital Repository Audit & Certification* being prepared for ISO standardization–distributed digital preservation solutions must make clear the ways that they differ, both organizationally and technically, from more centralized solutions.

The MetaArchive Cooperative has started this process by engaging in a self-audit with these existing tools. The MetaArchive's ability to actively conform to and address the concerns of each of the 84 criteria within TRAC successfully and to use this audit tool to help it schedule 15 items for review and enhancement, demonstrate that TRAC can be a valuable tool for distributed solutions. However, it is important for evaluators to engage in a careful and synthesized analysis of the repository, the standards, and the audit metrics in order to sincerely address concerns and identify new implementations that are compatible with the distinctive activities that are unique to this growing set of distributed preservation endeavours.

It is also worth questioning whether distributed digital preservation needs to be better defined by its community of practice. Abstracted principles that enable discussion, foster understanding, and provide a foundation for assessment are necessary elements in our growing digital preservation arena. It may be time to explore the efficacy of either better defining a distributed digital preservation network within the existing OAIS framework or creating a reference model that explicitly addresses the technological and organizational issues that arise in the distributed preservation network context.

REFERENCES

[1] Center for Research Libraries; Online Computer Library Center. *Trusted Repositories Audit & Certification: Criteria & Checklist Version 1.0*. Center for Research Libraries, Chicago, IL, 2007, Available at: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[2] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): ISO 14721:*2003, CCSDS Secretariat, Washington, D.C., 2002. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

[3] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Pink Book, August 2009*, CCSDS Secretariat, Washington, D.C., 2009, pg. 4-49 – 4-51. Available at:http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[4] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Pink Book, August 2009*, CCSDS Secretariat, Washington, D.C., 2009, pgs. 6-4 - 6-9. Available at:http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[5] Dale, Robin L. and Emily B. Gore. *Forthcoming* – "Process Models and the Development of Trustworthy Digital Repositories". Information Standards Quarterly, Spring 2010

[6] Dictionary.com. 2010. Definition of *trust*. Accessed on 1st May, 2010. Available at: http://dictionary.reference.com/browse/trust

[7] McDonald, Robert H. and Tyler O. Walters. "Restoring Trust Relationships within the Framework of Collaborative Digital Preservation Federations," Journal of Digital Information, Vol. 11, No. 1, 2010

[8] Research Libraries Group. "Developing a Digital Preservation Strategy for JSTOR, an interview with Kevin Guthrie," *RLG DigiNews* 4, no. 4 (August 15, 2000). Available at: http://worldcat.org:80/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file476.html#feature1

[9] Rosenthal, David. "Format Obsolescence: Assessing the Threats and the Defenses", Library Hi-Tech, Vol. 28, Issue 2, 2010, pgs. 195-200. Available at: http://www.emeraldinsight.com/journals.htm?issn=0737-8831&volume=28&issue=2

[10] Rosenthal, David; Robertson, Thomas; Lipkis, Tom; Reich, Vicky; Morabito, Seth. "Requirements for Digital Preservation Systems: A Bottom-Up Approach", D-Lib Magazine, Vol. 11, No. 11, 2005. Available at: http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

[11] Rosenthal, David; Robertson, Thomas; Lipkis, Tom; Morabito, Seth. "Transparent Format Migration of Preserved Web Content", D-Lib Magazine, Vol. 11, No. 1, 2005. Available at: http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

[12] Schultz, Matt. "MetaArchive Cooperative TRAC Audit Checklist", Atlanta, GA, 2010. Available at: http://www.metaarchive.org/sites/default/files/MetaArchive_TRAC_Checklist.pdf.

[13] Skinner, Katherine and Matt Schultz Eds. *A Guide to Distributed Digital Preservation*, Educopia Institute, Atlanta, GA, 2010, pg. 6. Available at: http://www.metaarchive.org/GDDP