Government Archives and the Digital Repository Audit Checklist

Bruce Ambacher

Modern Records Program

National Archives and Records Administration*

College Park, MD 20760

301-405-2043

bambache@umd.edu

ABSTRACT

This article examines the RLG/NARA draft *Audit Checklist for the Certification of Trusted Digital Repositories* from the perspective of publicly funded repositories, especially government archives. It reviews the historical origins of the checklist, the comments received from government archives on the metrics in the draft document and the task force's adjudication of those comments. Finally it addresses some unresolved issues.

Categories and Subject Descriptors: H.3.7 [Digital Libraries]: Collections, Dissemination, Standards, H.3.6 [Library Automation]: Large Text Archives, D.2.9 [Management]: Life cycle

General Terms: Management, Documentation, Verification, Standardization

Keywords: Policy expression, audit, assessment criteria, OAIS, Designated Community, Digital Archives

1. The Path to the Digital Repository Audit Checklist

The Research Libraries Group (RLG)/ U.S National Archives and Records Administration (NARA) *Draft Audit Checklist for the Certification of Trusted Digital Repositories* (RLG and NARA, 2005), which represents version 1, that circulated for comment between late August 2005 and January 2006 advances the worldwide search to manage and preserve digital records. For a few pioneering repositories the search began in the late 1960s (Ambacher, 2003). For the vast majority the search began just a decade ago.

The draft Audit Checklist's goal is to develop criteria to "identify digital repositories capable of reliably storing, migrating, and providing access to digital collections." (RLG and NARA, 2005). Certification also will instill confidence in data creators, resource allocators, and users that the repository - if it is a certified repository - meets recognized standards and can fulfill its preservation and access mission. This paper examines the draft Audit Checklist from the perspective of archives, especially government archives. This perspective is one enriched by more than a decade of working toward this goal through participation on, or leadership of, various efforts such as the *Reference Model for an Open Archival Information System*; the Digital Archives Directions workshop; the Archival Workshop on Ingest, Identification and Certification; the RLG/NARA task force; and federal government standards development.

Government digital repositories around the world were in the vanguard in the development of the *Reference Model for an Open Archival Information System*. Led by the data repositories of the national space agencies through the Consultative Committee for Space Data Systems (CCSDS) (http://public.ccsds.org/default.aspx), the developmental group, - dominated by government data managers, with other archivists and librarians - developed the reference model that became ISO Standard 14721:2003.

Reality checks throughout the eight years from conception in 1995 to international standard in 2003 reinforced the development effort. U.S. and international working group meetings were always open to anyone who wished to participate, minutes were published on the CCSDS website, and each draft OAIS RM was available for public comment.

The introduction of *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, commissioned by the Commission on Preservation and Access and RLG and released in May 1996 (RLG, 1996), also provided both a reality check and some archival grounding for the development of OAIS. It introduced the concepts of Content, Context, Fixity, Reference and Provenance to the OAIS community and helped shape their model. It also brought the two communities together as the OAIS was developed and the open forums described below were held.

The OAIS developmental group sought additional, focused input from the digital repository community through two open workshops – the Digital Archives Directions (DADS) workshop in 1998 and the Archival Workshop on Ingest, Identification and Certification Standards (AWIICS) in 1999. Both were hosted by NARA at its College Park, MD facility. At the AWIICS workshop the certification panel called for the development of a certification process for digital repositories as "a method by which an Archives' customers could gain confidence in the authenticity, quality, and usefulness of digitally archived materials." The group also believed certification would help "ensure management that an archives was fulfilling its role of long term preservation." Finally, the group noted that all four areas of traditional certification – individual, program, processes, and data – should be addressed to some degree in a digital certification program (AWIICS, 1999). The draft Audit Checklist does that.

The next major steps were the public release of the OAIS reference model in 2002 and its movement into the ISO standards process. In that same year RLG and OCLC released the *Trusted Digital Repositories: Attributes and Responsibilities* (RLG, 2002). This report provided a more comprehensive look at the organizational context for a digital preservation program and made a direct call for the development of a digital certification program.

The following summer RLG and NARA established the joint Digital Repository Certification Task Force with membership from the U.S., U.K., France, and the Netherlands representing multiple domains including archives, libraries, research laboratories, and data centers from government, academic, non-profit, e-science, and professional organizations.

The task force worked two years developing the draft Audit Checklist. The process consisted of background research, multiple periods of weekly teleconferences, exchanges of draft texts, and a multi-day meeting at NARA in Washington, DC. The draft version was released for public comment in August 2005 with the comment period extending through mid-January 2006.

2. Converging Traditions

The draft Audit Checklist recognizes the convergence of three digital data traditions – archives, libraries, and data centers. While information professionals in these institutions are converging and are using much the same terminology, their interpretation of the terms is influenced by the type of repository in which they work. One clear indication of this is title to the data in their custody. Archival repositories, including government archives, hold unique records and strive to obtain an unrestricted deed of gift or other legal instrument that transfers full ownership and physical custody to the archives.

Digital libraries, holding mixes of data in which the majority may be copies of data, such as e-journals also held in other repositories, more often operate on the basis of a deposit agreement. Under deposit agreements the donor may retain more rights and control over the data, including who may access the data, the costs imposed for access, and the right to withdraw the data.

Data centers often contain a mix of materials. Data centers usually reflect the mandate, collections, and cooperation of a designated community. The majority of their holdings are unique archival materials and data on deposit, materials for which they have clear title through deeds of gift or explicit deposit agreements. Some data centers, especially those which conduct web harvests, possibly without the knowledge of the data creator, may not have clear title to those data holdings. Again, these practices exist in traditional archival and library settings but have been brought to the forefront in the draft Audit Checklist.

3. The draft Audit Checklist

The public draft represents a major step toward certification. It is based on the premise that self-assessment is the essential first step in the development of a repository's certification program. By using the draft Audit Checklist as one major focus in a self assessment of their digital preservation programs, organizations can measure their established priorities and goals against the Checklist's metrics. When self assessment is blended with other organizational measures, the repository has made major strides in understanding digital preservation requirements, in determining what can be done to improve its programs, and in building the metrics to support a future external audit. The series of test audits conducted during the public comment period benefited from the results of the self assessment and corrective actions those repositories took as a preliminary to the external audit. Those same auditors noted, however, that self-assessments often failed to critically evaluate the repository's degree of success or the factors required for success.

The draft Audit Checklist is organized into four sections: Organization covers governance, staffing, policies and procedures, financial sustainability and contracts and other obligations. Program functions addresses the whole range of repository preservation responsibilities including ingest (accessioning), archival storage, description, metadata, access, and preservation strategies. The Designated Community section focuses on both the records creators and users and the ability of the repository to meet their needs. The Technologies and technical infrastructure section concentrates on security, software and hardware, and similar issues that enable digital preservation.

The draft Audit Checklist was designed to be used and adapted by a variety of digital preservation programs including archives, museums, libraries, cultural heritage organizations, e-science programs, and data centers. The task force discussed the need for case studies based on application of the draft Audit Checklist. Such studies could provide a body of experience, example and guidance for other digital repositories as they begin self assessment and external audit. Version 1 of the draft Audit Checklist does not address its implementation – Who will be eligible to be audited? Who will conduct the audits? What will it cost? Version 1.0, *Trustworthy Repositories Audit & Certification: Criteria and Checklist*, scheduled for release in March 2007, reflecting adjudication of the comments received and additional revision by the task force, will be used as one basis for development of a draft international standard on digital repository certification.

4. Public Comments on the draft Audit Checklist

During the six month public comment period twenty-three institutions and individuals provided comments on the draft Audit Checklist. This was a wide, varied and representative set of responses. The major themes and concerns focused on the relevancy of the domain and the audience, the difficulty and deficiencies of the metrics, the logistics of the process, and the degree to which specific communities would commit to the process, and compliments.

The fact that both archivists and librarians thought the draft Audit Checklist was too specific to the other's domain and therefore less relevant to their domain indicates, in my view, that the task force successfully avoided being domain specific in its language. As repositories use the checklist and possibly post the results of their self assessments, other repositories will more readily understand how the checklist applies to their domain.

The draft Audit Checklist places several traditional archival concepts and practices in a new context, bringing them under sharp scrutiny. These include the mission statement as it pertains to the focus of the collection and who will use the records, the nature of ownership, the commitment to permanence, the loss of holdings, preservation including disaster prevention and preparedness, and the degree to which a repository must measure up to profession-wide standards in order to be accepted as a trusted repository.

A. Metrics for Certification

Some reviewers thought the metrics for certification were too difficult, while others thought they were too easy. To some extent this reflects an ongoing dichotomy within the task force whether certification represents a pass/fail process or a graduated multi-level process reflecting graduated degrees of competency and confidence. The task force is clear that certification should not a one time, forever accomplishment. Rather, it should be an ongoing process that must be renewed periodically to reaffirm to the designated communities served by the repository that it continues to be a trustworthy repository. The task force also understood version 1 and version 1.0 are preliminary steps in a broader process to develop an international certification standard.

A second major theme concerning the metrics was how they would be measured; i.e., what evidence could a repository provide to prove it had met the criteria of that metric. The task force accepted and amplified the "evidence" criteria developed for each metric by the Digital Curation Centre. These evidence statements can be seen in version 1.0.

The task force also agreed that the process will work best if it is a two stage process in which repository self assessment precedes any external audit. This two stage process, and the types of evidence that a repository can use to demonstrate meeting any metric, has been elaborated in Version 1.0. Many repositories may choose to stop after self assessment, to correct deficiencies, and to await the development of the certification process. Other repositories, because they are confident they do not need further action to meet the needs of their designated community, may not seek certification.

The comments received contained numerous broad, abstract or conceptual ideas and issue development about the certification process. They also contained numerous very specific edits, suggestions and concerns. The task force reviewed and adjudicated each comment.

B. NARA Comments

One of the major set of comments from a government archives came from NARA. The NARA staff who participated in the review of the draft Audit Checklist characterized it as

a "useful guide for evaluating a digital repository." The core of the issue, in their view, was whether the checklist could become more than an evaluation tool. The answer depends on the development of a certification infrastructure. The task force fully understood this issue. The working group currently beginning the ISO certification standard process must address some mechanism to provide competent audit teams within a valid certification infrastructure.

The NARA review team also pointed out that the draft Audit Checklist would impose de facto standards. Measuring up to such standards could impose additional procedures and obligations on digital repositories. The NARA reviewers focused on the ingest phase and noted the metrics could impose additional verification and metadata development requirements on a repository which could translate into additional resource obligations. The same issue could be raised about other aspects of the archival life cycle, especially archival processing and preservation storage. While valid, those comments address the very purpose of the draft Audit Checklist, the audits conducted to date, and the initial work on an international standard - to develop measurable criteria that will demonstrate that a digital repository has taken the necessary steps to ensure ongoing preservation of, and access to, its digital assets.

C. Succession Planning

Section A1.2, Succession Planning, proved vexing to both public and private repositories. It calls for repositories to have a succession plan or escrow arrangements in place in case the repository ceases to operate. Government archives noted their mandate is to act as the repository for their government for as long as that government exists. Private repositories questioned the potential impact on the confidence level of possible donors if the repository has a defined plan for its demise in place at the same time it is soliciting new collections. The task force reaffirmed that all digital repositories should address this issue.

D. Financial Sustainability

One recurring concern in the government archives' comments on the draft Audit Checklist was Section A.4, Financial Sustainability. This metric calls for a digital repository to "prove its financial sustainability over time." Many government institutions noted that they operate as part of an overall government budget, not as part of a business plan with short and long term financial planning cycles and the ability to adjust programs to budgets or to develop operating reserves. The task force understood the nature of the government budget cycle and viewed the existence of a government mandate and a number of years of past funding as a reasonable premise for the repository to expect relatively stable funding and to undertake long range planning.

E. Contracts Licenses and Liabilities

Government archives were critical of Section A5, Contracts, Licenses, and Liabilities. They viewed it as being too weak on requiring a repository to have clear title to its

collections and a corresponding clear statement of responsibility and duties. Their position was quite distinct from contributor based repositories which may rely more upon deposit agreements, or web "archives" that harvest their holdings from the internet, possibly without prior knowledge or permission of the creators. The task force purposely allowed some latitude on this to accommodate multiple deposit arrangements while encouraging the most binding arrangement possible.

F. The Role of Government Archives

The government archives that evaluated the draft Audit Checklist included those with well established, respected digital records programs. These institutions were confident they could become certified repositories even though such criteria have not been established. One even volunteered to serve as an example to assist in the further development of the comprehensive process. Another government agency noted that it had used the checklist as a valuable guide as it planned the Request for Proposals to initiate development of its digital repository program.

Two things should be noted here: First, digital certification probably will not affect the relationship of a government archives to its sponsoring government or to the government agencies that create its collections. Those relationships and obligations usually are established by law and regulation. Second, in the short term, government archives, if they choose, can ignore the draft Audit Checklist and certification process and continue their archival activities with undiminished status.

The responding archives, however, did not take that view. Instead, they recognized an obligation to provide leadership on the issue, to use the audit as an opportunity to evaluate and improve their programs. They also saw it as an opportunity to help other archives understand the audit process and achieve certification. They recognized preservation and customer service as clear parts of their mission worthy of re-examination from a new perspective.

G. Designated Community

The concept of the Designated Community, which relies heavily upon concepts and terminology in the *Reference Model for an Open Archival Information System*, generated comments from both public and private repositories. While using more direct language than traditional archival program mission statements might use, there is little difference between the designated community concept and the collecting policy of a labor archives, the scheduling and appraisal of institutional records in a university archives or a private corporation, the mission statement of a donor-based collecting archives, or the holdings of a digital repository established to serve a designated community. Government archives pointed out that, as public institutions, their definition of the designated community they serve, must be quite broad. At NARA, for example, the legally defined user community is anyone with an interest in records of the U.S. Government or the information they contain, for whatever reason. A user need only provide proof of identity and age (they must be fourteen years of age or more), and state their research purpose. Others who

commented on the draft were concerned by the implications of the term designated community and by the strident nature of the definitions and examples of designated community, fearing they were too restrictive and would inhibit the more casual researcher. It certainly brings into clear focus a long accepted practice of repositories defining their collections and, therefore, their probable users. Viewed alone it seems strident; viewed as a refinement of decades of practice, it is more acceptable. In version 1.0 the designated community requirements are integrated into other sections of the checklist to further place them in context.

5. Test Audits

At the same time that the draft Audit Checklist was undergoing public review and comment, it also was undergoing practical field testing. The Center for Research Libraries, with funding from the Andrew W. Mellon Foundation, undertook a project to test the Audit Checklist in a variety of digital repositories. The repositories audited included the Koninklijke Bibliotheek (KB), the Interuniversity Consortium for Political and Social Research (ICPSR), Portico, and LOCKSS (http://www.crl.edu/content.asp?l1=13&l2=58&l3=142&l4=71 accessed 10/20/06). This project was headed by Robin Dale, RLG's Audit Checklist co-chair. This ensured compatibility between the projects and the ability to use the field test results in combination with the public comments in developing *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (Version 1.0 of the Audit Checklist) scheduled for release in March 2007. Version 1.0 also benefits from digital preservation work being done at the Digital Curation Centre (http://www.dcc.ac.uk/ accessed 10/20/06) and from the certification work of the Nestor project (Network of Expertise in Long-Term Storage of Digital Resources) in Germany (http://www.langzeitarchivierung.de/index.php?newlang=eng accessed 10/20/06.)

6. Unresolved Issues

A. Data Loss

The nature of digital ownership can affect the repository's commitment to preservation. The draft Audit Checklist, by evaluating various aspects of preservation, can make the repository's position more transparent to potential donors and users if the results are made available in some form. One aspect of digital preservation that the checklist brings into clear focus is the acceptance of data loss. While no repository establishes a program with the intention to lose a portion of its holdings, the checklist, reflecting the volatile nature of data, recognizes data loss may be inevitable and encourages digital repositories to determine and state the amount of data loss that is acceptable to them. The draft Audit Checklist contains the figure .0001, one-one-thousandth of one percent. If traditional repositories were required to state what percentage of their holdings they had lost, or expected to lose, to negligence, theft, disaster, or natural forces within the documents, what would their figure be? Should the fact that the information is also held in another digital repository, for example an e-journal, affect a repository's approach to preservation?

The task force chose to highlight this issue to reinforce the need for early decisions regarding the value of digital objects and for continuous maintenance of the objects in order to ensure their longevity. This stands in sharp contrast to traditional repositories with textual collections where breaks in preservation activity, a period of benign neglect, may not result in loss of the collections.

B. Maintaining Accessibility

A second aspect of digital preservation that may be affected by the nature of ownership is the commitment to maintaining the data in a viable form through time, across potentially multiple hardware and software platforms and a declining user community understanding of the data and the accompanying metadata and documentation. This commitment may be affected by the uniqueness of the data and the type of ownership.

C. Disaster Planning

A third aspect of preservation evaluated by the draft Audit Checklist is the adequacy of a repository's disaster planning. It examines elements unique to digital objects such as duplicate copies housed offsite, access to alternate process systems, and universally understood external and internal data file labels. Again, the aim is to instill best practices into a repository's daily operating procedures.

Collectively the draft Audit Checklist's evaluation of the repository's commitment to preservation begins to assess it against accepted norms and thereby allow donors and users to compare the repository with other data repositories.

7. Precedents for External Evaluation

Another area in which the draft Audit Checklist extends traditional archival principles and practices beyond the casual to the formal is the basic concept of the need to be formally evaluated and assessed. Archival organizations, including the Society of American Archivists, have had an institutional evaluation checklist for decades (Society of American Archivists, 1982). It is voluntary, relatively brief, little used, and rarely publicly cited. The task force's intention for the draft Audit Checklist is that the audit and certification process becomes a two stage process, that it becomes so respected that it is virtually universally used, and that it becomes the intellectual foundation for donors and users to easily identify quality digital repositories.

The first stage would be a repository self evaluation, possibly with assistance from experienced outside mentors. In the self evaluation phase the repository would gather the evidence and supporting information required for a formal audit. This activity would allow the repository to determine which aspects of its operations were deficient and to undertake improvements. During this process the staff also would identify those aspects that required additional funding and seek that funding from its resource allocators. A repository should only seek digital certification once it has undergone self evaluation and has addressed all issues which arose from that self evaluation.

8. Achievability

The task force remains divided on the attainability of certification. There is a philosophical split on what the "pass-fail" rate should be. One position holds digital certification to be something to aspire for. It should be difficult to attain if it is to have any meaning or value. They support this with the fact that "data archiving" is relatively new and few repositories have sufficient expertise and infrastructure to meet the certification criteria.

The other position is that any repository should be able to be certified if it examines the criteria, evaluates its program and develops the appropriate infrastructure and procedures. This position does not hold that digital certification should be automatic or easy to attain but that it should be an attainable goal. They hold that digital certification can be both highly valued and widespread.

9. Conclusion

So, what is the role of government archives in this scenario? First, because they tend to be the older, more experienced, and better funded programs, they have an obvious leadership role. Government archives have, and some recognize they have, a professional obligation to lead the way. They should be among the repositories helping to determine the audit criteria – and they are. Once the Audit Checklist process leading to an international certification standard is complete, government archives should be among the first to undergo the audit process, beginning with self evaluation. They should assist other digital repositories as they begin to transform their programs to comply with the audit checklist. They can assist the process by providing staff to serve as auditors in other institutions and serve to train other to serve as auditors. They can help fund the audit and certification process mechanism in whatever form it will take. They can develop illustrative case studies to guide other repositories. Finally, government archives can promote the audit process with political bodies, regulatory boards, potential donors, and users.

REFERENCES

******Prior to his retirement in 2007, the author served as NARA's co-chair of the RLG-NARA Task Force to develop the Draft Audit Checklist for the Certification of Trusted Digital Repositories. In 2007 he became a Visiting Professor in the College of Information Studies, University of Maryland, College Park.*

Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS) http://nost.gsfc.nasa.gov/isoas/awiics/ accessed 10/20/06

Bruce I. Ambacher, *Thirty Years of Electronic Records*, Scarecrow Press, 2003

Commission on Preservation and Access and RLG, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information,* RLG, Mountain View, CA, May 1996.

Digital Archive Directions (DADs) Workshop
http://nost.gsfc.nasa.gov/isoas/dads/papers2.html accessed 10/20/06

Digital Curation Centre (http://www.dcc.ac.uk/ accessed 10/20/06

Nestor project, Network of Expertise in Long-Term Storage of Digital Resources (http://www.langzeitarchivierung.de/index.php?newlang=eng accessed 10/20/06

RLG and NARA, *An Audit Checklist for the Certification of Trusted Digital Repositories,* RLG, Mountain View, CA, 2005; pp.1.http://www.rlg.org/en/page.php?Page_ID=20769

RLG, *Trusted Digital Repositories: Attributes and Responsibilities, Mountain View,* CA May 2002. http://www.rlg.org/legacy/longterm/repositories.pdf. accessed 10/20/06

Society of American Archivists, *Evaluations of Archival Institutions: Services, Principles, and Guide to Self-Study,* Chicago, 1982