CRITIQUE: ATTRIBUTES OF A TRUSTED DIGITAL REPOSITORY

H.M. Gladney

HMG Consulting
20044 Glen Brae Drive
Saratoga, CA 95070

10th October 2001

Overview

This draft report [RLG 01] deals well with a timely topic; however it seems to me to pay too little attention to recent technology and economic trends. Skillful application of some ideas and methods developed for other applications (such as e-commerce and banking) **might achieve your desired results more economically** than straightforward extensions of current library roles and procedures.

Economic trends might change how libraries and archives evolve. However, [RLG 01] draws primarily on the experience and methodology of professional librarians, making little use of recent progress in engineering and other disciplines. Your draft report overlooks trends that can be exploited for more cost-effective solutions and **more satisfactory roles for librarians—focusing on knowledge search and delivery** rather than on library services that can be automated.

Specific Comments

TCB Ideas Risk Inducing Wasted Expenditures

You build forward from a careful analysis (your pp. 6-17) of the "Trusted Computing Base (TCB)." But TCB was designed for defense applications, not for delivery of mostly public information. Starting with TCB ideas might lead the library community into adopting systems, infrastructure and internal methodology **far more expensive than needed to achieve your objectives**.

In addition to being designed for military intelligence work, a TCB is intended for **reliable** execution of arbitrary programs whose results cannot be independently validated except by further expensive calculations. In contrast, archives have only two critical kinds of output: reproductions of the documents stored and search results. Well-known technologies, such as message authentication codes, would allow users and auditors to validate such outputs inexpensively. [Schneier 96] I.e., **TCB is expensive overkill.**

What archive users want is trustworthy information. They do not particularly care that any particular library service is trustworthy. (This does not imply that librarians and other information service agents can ignore repository trustworthiness, because it might reflect on the reputation of their institutions.) **Proof of digital document authenticity and provenance can be achieved by message authentication codes signed by trusted institutions** and included in the metadata integral to each document delivery. E.g., while I would value a **reliable** British Library signature certifying the authenticity and provenance of a document received, I won't much care whether the document came from a British Library repository or not.

Certification vs. Testable Properties

Your report recommends certification of repository trustworthiness. I believe that **shifting focus from reliable repositories to reliable information deliveries will not only save money, but also improve the quality of your product (the information delivered).**

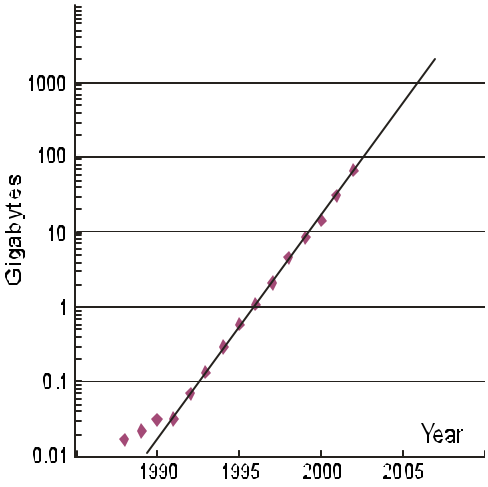
Certification and periodic auditing are expensive in skilled human resources. Structuring documents and metadata can be mostly automated, with the human labor required reduced to clerical steps that can easily be integrated into conventional library procedures.

The issue of reliability is more subtle—avoidance of systematic flaws that are adopted by all libraries because they were too trusting that the designers considered all potential failures and implemented correctly. (This issue parallels one that has worried scientists using simulation programs. If two scientists use the same program, which can happen without their noticing it, they might mistakenly think that similar results validate each other's work.) Presumably, the certifiers for one repository would be staff members of other repositories. Unfortunately, the research library and archive community is too small to elicit confidence in such a process, because the entire community might accept flawed standards or tools.

Furthermore, your report does not hold out hope for objective certification criteria—only subjective criteria. Standard engineering practice—particularly a rich history of digital security—teaches how to be objective. Instead of focusing on positive qualities, we should identify the important risks and failure possibilities and what mechanism is in place to minimize the frequency and impact of each. If this is done, **outside auditors will be able to test each assertion of quality**. (Such auditors will not need special knowledge of kinds available only in the library/archive community. This would make available a much large and less in-bred pool of auditors.)

In fact, having reliably signed documents is better than having unsigned documents delivered from reliable sources because the former reduces chances for mischief in the delivery network. I.e., **by shifting from concern with the trustworthiness of repositories to trustworthiness of information delivered, you can do the job both better and less expensively!**

Disruptive Trends in Digital Technology



Current trends in digital technology will continue for at least five years, and with other factors alluded to in this note, might enable quite different distribution of network tasks than are implied in your draft report.

For instance, consider magnetic disk storage technology (see graph illustrating the storage capacity that $200 will buy). Within 4 years, a terabyte of storage will be affordable for home personal computers.[1] Presuming availability of metadata along the lines proposed by several current XML-dependent initiatives, [Cover 01] it will be easy and inexpensive for any researcher to accumulate his own catalog and also to keep a copy of every document of interest to enable full text search.[2]

Such market trends combine favorably with replication in the Internet to achieve reliability for document preservation. See, for instance [Reich 01] and [Waldman 00]). You should consider how such ideas might be exploited to achieve your stated objectives. For instance, archives could buy storage service from industrial repositories or user institutions in order to avoid training their own specialized personnel. **It is not clear that libraries should be the only digital repositories, or even digital repositories at all.**

Enlarging on such notions, you should consider what [Odlyzko] suggests. **Marketplace trends can be exploited to improve how libraries and scholars operate.**

Finding Information

Today's most effective information discovery infrastructure seems to be the catalogs and search engines of the World Wide Web—not the catalogues of research libraries. (Of course, a combination of WWW search tools and library catalogs is more effective than either alternative used alone. E.g., when I search, I first exhaust Google, then use the University of California Melvyl catalog, and finally the on-line catalogs

[1] Similar improvements in other critical technologies—communications bandwidth, switching technology, processor speeds and costs, and so on—are expected, but not as easily quantified as those for magnetic disks.

[2] Intellectual property rules will influence what is allowed. Because those rules are in flux it is difficult to project their effects.

of the Library of Congress and similar sources.) However, the draft report does not consider effective combinations of traditional library mechanisms and infrastructure from the private sector.

Furthermore, given the technology cost trends and improving database technology, individual researchers (or, more realistically, small interest groups) will find it easy and affordable to construct search databases better suited to their particular interests than those libraries provide. Automatic means could keep such databases up-to-date. This suggests possible restructuring of how and where information functionality is laid out in the Internet. In a decade or two, **libraries might be neither the most used repositories nor the preferred search providers; they will still have a critical role in scholarly activities and preservation of the cultural record, but it will be different than it is today.**

What Corpus Will be Chosen for Preservation?

Your report should start to estimate how much "stuff" people might reasonably feel needs to be saved, or what kinds of digital documents might be wanted, or what the information sources might be. It seems to assume that the corpus is scholarly publications, but the validity of this is far from obvious. (Consider the holdings of the big research libraries; scholarly publications are only a small fraction of these. Recall the importance of unpublished material, e.g., the bills of lading of 16th century Spanish ships held in the Archive of the Indies in Seville, Spain. Another example is the Library of Congress holdings of many years of fire insurance records for Philadelphia.)

What will eventually be saved will be more than scholarly writings and cultural output such as folk music recordings, and much less than everything. Of course, estimates of quantities and sources of content made now will prove to be erroneous. Nevertheless, they will suggest changes of how collections should be managed, helping us **avoid organizational responsibilities and practices made unaffordable by the numbers and complexities of digital documents.**

Conclusions

I respectfully suggest that the authors of this draft report consider delaying "official" release until they are confident of good answers to the issues raised. It is my reluctant opinion is that RLG endorsement of the current draft would not be a service to the communities that care.

Bibliography

[Cover 01] Robin Cover, The XML Cover Pages, **http://www.oasis-open.org/cover/sgml-xml.html**, 2001.

[Odlyzko] A.M. Odlyzko, *Tragic Loss or Good Riddance? The Impending Demise of Traditional Scholarly Journals*, short version to be published in Notices of the American Mathematical Society, (Jan. 1995). http://www.research.att.com/amo/doc/eworld.html Also Odlyzko, *The history of communications and its implications for the Internet*, available at http://www.research.att.com/amoi.

[Reich 01] *Vicky Reich and David S.H. Rosenthal, LOCKSS: A Permanent Web Publishing and Access System*, D-Lib Magazine, June 2001.

[RLG 01] RLG-OCLC Working Group, *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources*, Draft for Public Comment, August 2001.[

[Schneier 96] B. Schneier, *Applied Cryptography: Protocols, Algorithms, and Source Code in C*, 2nd Edition, ISBN 0471117099, John Wiley & Sons, New York, NY, 1996

[Waldman 00] Marc Waldman, Aviel D. Rubin and Lorrie Faith Cranor, *Publius: A robust, tamper-evident, censorship-resistant, web publishing system* , Proc. 9th USENIX Security Symposium}, 59--72, 2000.

Appendix C of [RLG 01]: Roster of the RLG/OCLC Working Group

Neil Beagrie	Joint Information Systems Committee
Marianne Doerr	Bayerische Staatsbibliothek
Margaret Hedstrom	University of Michigan
Anne Kenney	Cornell University
Catherine Lupovici	Bibliothèque nationale de France
Kelly Russell	Cedars Project (CURL exemplars in digital archives)
Colin Webb	National Library of Australia
RLG Liaison: Robin Dale	Member Programs & Initiatives
OCLC Liaison: Meg Bellinger	Preservation Resources