LONG-TERM DIGITAL PRESERVATION: WHY IS PROGRESS LAGGING?

H.M. Gladney

Saratoga, CA 95070

Abstract: An in-principle solution for every difficult challenge of long-term preservation had been published by 2007, with a request that the interested community criticize this work. Such criticism was deemed essential before investing to create required software and mount pilot installations that test and demonstrate the ideas' correctness and practicality. The author paused, waiting for reactions.

Over four years later, almost nobody has commented on this work, nothing distinctly different and workable has been published, and several annual preservation conferences and workshops seem remarkably similar to their counterparts of a decade ago. How could this happen?

The author suggests social flaws—specifically inattention of the archival community to work of software engineers. He sketches the 2007 solution and challenges the workshop participants to refute his technical and social views. Specifically, participants are invited to demonstrate that he is mistaken in asserting that his theoretical work is correct but ignored, that four years have passed without progress, and that these flaws are the consequence of well-known weaknesses of scholarly communities.

"DIGITAL ARCHIVING" AND "LONG-TERM DIGITAL PRESERVATION"

It might be trite to note that "digital archiving" and "digital preservation" should not be taken to mean the same thing. However many published works seem to confound or conflate these phrases.

Satisfactory digital library services, sometimes called digital archiving,[1] have been in wide-spread use for about two decades. The enterprises that provide them assiduously listen to their customers and provide improvements for whatever these customers identify as shortfalls. In contrast, digital preservation is commonly identified as a challenging research topic.

It is therefore prudent to choose "long-term digital preservation" to mean something beyond "digital archiving"—a set of activities that do not supplant "digital content management" functionality. Instead, what we mean by the former phrase is a set of upwards compatible extensions of digital archiving services as these are today widely construed.

In contexts suggested by Figure 1, "digital archiving" seems to be generally construed to describe a service in which any authorized Information Producer is provided service in which any suitably prepared digital document is handled according to a repository institution's published statement of service. Furthermore, this service commits to assist any authorized Information Consumer with finding and obtaining an authentically true copy of almost any of its catalogued holdings.

The Figure 1 Information Consumer can, perhaps by making suitable inquiries of an appropriately selected Information Producer or Archive Manager, ascertain that information copies he receives are what they purport to be. If this is not reliably possible, the Consumer understands and accepts the risks associated with perhaps being misled.

We interpret "digital preservation research needs" to include only challenges caused by deterioration of media, changes of digital representation, and fading human recall, and to exclude digital repository and information-finding needs that would occur even in a world free of information degradation over time. In this spirit, preservation is a collection of measures that, to the extent theoretically feasible, enable the eventual Information Consumer to interpret the content of a preserved object satisfactorily, to examine its embedded provenance description, and to test that all this delivered information is pre-

[1] Some authors distinguish "digital library" (a.k.a. "digital content management") from "digital archiving"—a distinction that I do not understand, but believe to be irrelevant in the context of the current article.

cisely what it purports to be. It enables all this even if no creator or custodian of his copy is available to answer any questions, perhaps because all erstwhile authorities have long ago died.

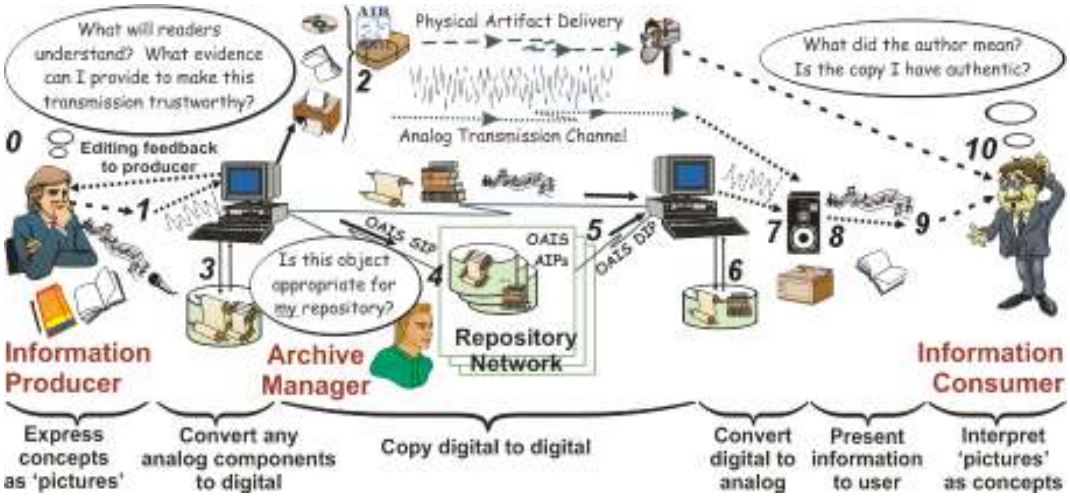

Figure 1: Documentary information interchange and repositories [from Gladney 2007], suggesting alternative pathways by which a document might move to its eventual consumer.

What would a digital preservation solution need to accomplish? It should:

- Ensure that a copy of every preserved document survives as long as it might interest someone;
- Ensure that authorized consumers can find and use any preserved document as its producers intended, avoiding errors introduced by third parties that include archivists, editors, and programmers;
- Ensure that any consumer has the information to decide whether information received is sufficiently trustworthy for his intended application;
- Hide information technology complexity from all its end users;
- Replace human effort by automatic procedures whenever doing so is feasible; and
- Empower authors, editors, and other information producers to package information so as to relieve overloading of professional cataloguers.

A practical answer to these challenges would allow conservators to protect millions of intellectual products from the ravages of technology obsolescence and fading human recall. It would also allow repository institutions and individual users to continue to use the tools they have already deployed and their evolving replacements without disruption, conforming to software interface standards and conventions that permit "mix and match" from competing providers—standards and conventions that, over time, will continue to be improved.

THE STATE OF AFFAIRS AND A CHALLENGE

Roughly a decade ago, the U.S. Congress funded and the Library of Congress led a multi-year study called the *National Digital Information Infrastructure Preservation Program (NDIIPP).* [Anon. 2004] This author participated in many of its committee meetings. About halfway through the allotted time, since his recommendations were seemingly ignored, he published opinions of what had to change for NDIIPP to be successful. [Gladney 2007] Little changed. The *NDIIPP* expenditures led to nothing that deserves to be called a national program.[2]

The author and a few mostly silent colleagues concurrently worked out a method that we claim provides an in-principle answer for every earlier technical question about how to preserve any digital ob-

[2] The Library of Congress publishes a newsletter reporting what remains of NDIIPP. Its content is similar to that of several annual European conferences and workshops. See, for instance, what is reported at http://www.digitalpreservation.gov/news/newsletter/201107.pdf.

ject reliably into the indefinitely distant future.[3] We published this in several venues, together with sincere invitations for public or private criticisms.[4] What was the reaction? Almost nothing! As far as we know, only a single published reaction appeared, and it expressed puzzlement. [Prom 2009]

This history stimulated my sending electronic mail to Livia Prediou, the organizer of the conference for which the current article was accepted. It expressed a questions and challenges for which colleagues and I would enjoy the most specific answers possible—questions whose essence is repeated below.

These questions originate in personal opinions—opinions that might be biased—that the last decade's publications and conferences have not, in fact, revealed technical ideas or significant sharable software implementations beyond what was published by 2007. Our challenge invites the conference participants to identify articles that refute my perceptions, with each refutation accompanied by a brief statement of its novel idea(s) and of how this idea or implementation has advanced the state of the art.

Our scope of interest is managing documents of academic and public interest, such as those of museums and national libraries. I.e., the author does not intend the current inquiry to include proprietary or private-sector work that is unsuitable or too expensive for public institutions. His questions and comments are further limited to technical aspects, alluding to social aspects only indirectly by way of formal requirements for technical components.

The perceptions for which refutations are invited are that, for the period 2007 to the present: [5]

(1) Presentations and publications describing digital preservation have been limited to "here's how we would do it if we could muster sufficient resources, which we have not accomplished".

(2) No significant new technical ideas for digital archive management or digital document preservation have emerged.[6] Here, "significant" means "providing ease of use and/or ease of implementation and/or cost reduction over what prior ideas promise".

(4) Much prior work on the CfP topics has conflated digital preservation with archiving. Both discussions and eventual solutions will be much simplified if archiving and preservation are treated as distinct topics that are lightly coupled.[7]

(5) The additional technology to support digital archiving and preservation of public sector content is a small increment over widely used affordable content management offerings and personal computer document management software.

(6) The design detailed in [Gladney 2006] and related publications, and sketched below, is close to optimal (close to best "bang for the buck") for the domain under discussion. (Respondents might be able to identify and describe improvements, doing so sufficiently articulately for publication. My colleagues and I would welcome such criticisms.)

PRESERVATION WITH TRUSTWORTHY DIGITAL OBJECTS (TDOS)

There follows a brief sketch of what we claim is, in principle, a complete and optimal preservation solution. Skeptical readers are referred to [Gladney 2006], published in the flagship ACM periodical, and to a later book. The book addresses preservation starting with epistemological foundations and ending with a software sketch. It is centered on two models: that of the many paths in which documentary information flows from its creators to its consumers (Figure 1) and that of a durably useful representation of each information parcel (Figure 2).

---

[3] Our work is limited to prescription of the technical component of long-term digital preservation, leaving careful treatment of other elements to other authors.

[4] Among software engineers, such openness and invitation for professional criticism is deemed merely prudent before continuing to the next steps—solution embodiment in program offerings that are tested in full-scale pilot installations. This is part of scientific and engineering practice taught in the 19th century. [Snyder 2011]

[5] This is the [Gladney 2006] publication date. Additionally the *Digital Document Quarterly* tracked other workers' activities between 2002 and 2009; see http://www.hgladney.com/ddq.htm.

[6] A few American workers might have created counter-examples. I have yet to search for recent work of David Rosenthal and Vicky Reich (Stanford University), of Herbert van De Sompel (Los Alamos National Laboratory), and of Carl Lagoze (Cornell University).

[7] In software implementations, such coupling becomes explicit module interfaces.

In ancient times, wax seals impressed with signet rings were affixed to documents as evidence of their authenticity. A digital counterpart is a cryptographic signature block embedded in each important document. Our objectives suggest other solution elements that can be almost independently addressed:

- Content servers that store packaged works, and that provide search and access services.
- Replication mechanisms that protect against the loss of the last remaining copy of any work.
- A method for packaging a work together with metadata that includes provenance assertion and reliable linking of related works, ontologies, rendering software, and links connecting package pieces with one another. [Figure 2]
- Topic-specific ontologies defined, standardized, and maintained by professional communities.
- A bit-string encoding scheme to represent each content piece in language insensitive to irrelevant and ephemeral aspects of its current environment.
- Some number of socially-communicated languages and standards for encoding starting points.

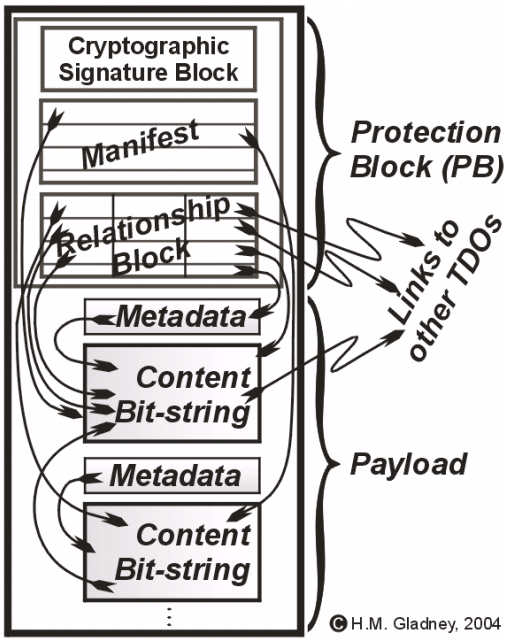

Figure 2: Structure of a Trustworthy Digital Object (TDO)

The TDO base (Figure 2) is record schema for long-term holdings. To prepare the object set that makes up a work, an author or editor causes conversion of each content bit-string into a durably intelligible representation and collects the results, together with standardized metadata, to become the payload of a new TDO. In addition to its payload, each TDO has a protection block into which a human editor loads metadata conforming to evolving standards and records relationships among parts of the new TDO and between it and other objects. The final construction step, executed at a human agent's command, is to seal this bundle with a *cryptographic signature block*. In a valid TDO representing some version of an object:

- The bit-string set that represents the version is XML-packaged with registered schema.
- These bit-strings and metadata are encoded to be platform-independent and durably intelligible.
- The metadata include identifiers for the version and for the set of versions of the work.
- Every critical link to another TDO is secured by a message authentication code.
- The package includes or links reliably to all metadata needed for interpretation and as evidence.
- All these contents are packaged as a single bit-string sealed using cryptographic certificates based on public key message authentication.
- Each cryptographic certificate is authenticated by a recursive certificate chain.

A Competing Preservation Proposal

In continuing "due diligence" search for challenges to and improvements of TDO technology, I recently encountered *Towards SIRF: Self-contained Information Retention Format* [Cohen 2011] and performed a careful comparative analysis.[8] A significance of this paper is that it is the most recent technical article within our domain of discourse, and should be expected to position its assertions vis-à-vis all the most competent prior literature.

The central notion of [Cohen 2011], built around its Figure 2 (copied below as Figure 3) seems to be a model for how preserved data should be structured within archival storage.

[8] This analysis was shared with the paper's authors. An earlier inquiry suggesting their making such a comparison had been sent them on 21st June. Discussion has started, but not finished.

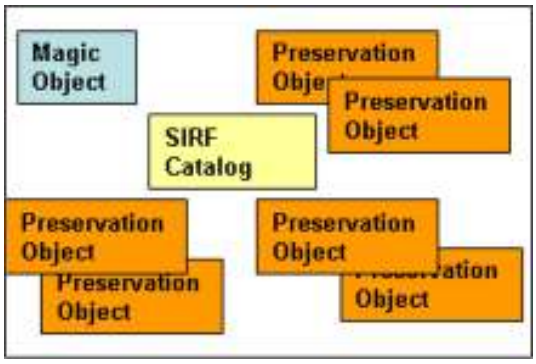

Figure 3: SIRF components [from Cohen 2011]

About this, my current view—open for amendment—is as follows.

- The SIRF description starts by asserting intent for OAIS compliance [OAIS 2009], but is even less detailed than OAIS. Furthermore, OAIS itself does not specify enough detail to ensure that compliant packages can be shared among machines managed by interested participants. [Egger 2006]
- The Figure 3 SIRF structure is a less-detailed version of the Figure 2 TDO structure. [Cohen 2011] does not provide enough detail for more insightful comparison.
- [Cohen 2011]'s focus on how data is structured within a repository is inappropriate. End users don't care how this is done, but instead only about information formats for repository ingestion (OAIS SIPs, as in Figure 1) and as delivered by repositories (OAIS DIPs). Moreover, a little thought centered on the Figure 1 communication model suggests that, for unimpaired information exchange, each DIP must be identical to the corresponding SIP.
- The paper hardly mentions widely-deployed digital content management offerings, and says nothing about how a SIRF realization would extend these without disrupting their users.
- [Cohen 2011] is deficient in not comparing itself to prior work. What is most disturbing about this is not the fact in itself, but rather that this weakness is common in digital preservation literature.

Summarizing this, the [Cohen 2011] claim that "SIRF is a significant advance" is unjustified. As far as this author has been able to determine, the paper offers nothing that was not already described in deeper detail in [Gladney 2006].

SUMMARY AND CONCLUSIONS

> "All this has been said before—but since nobody listened, it must be said again."
>
> Attributed to André Gide

Many authors confound or conflate long-term digital preservation with digital archiving. Today, many self-styled digital archives exist and serve without claiming or providing long-term digital preservation. And it is possible to preserve information reliably without involving an archiving institution.

There has been no significant advance toward digital preservation since about 2007. Instead, conference speakers repeat statements of requirements and wails already heard a decade ago and earlier.

This author perceives disturbing viewpoint differences over what constitutes a topic worthy of "digital preservation research". For computer scientists and engineers, research is required only for challenging questions for which plausible answers are not already available. Creation of tools to realize those ideas and promise what end users and archivists say they want is called "development". In contrast, archivist and librarians seem to regard as research-worthy any need for which they cannot acquire fully satisfactory tools within their (regrettably) straightened budgets.

The problem of the prior paragraph is, perhaps, partly a consequence of a truth that is rarely, if ever, mentioned. Archives and libraries have neither the skills [Prom 2009] nor appropriately structured budgeting practices for creating software tools that they obviously need.

Dysfunction across disciplinary boundaries pervades all these problems. Specifically, progress towards digital preservation at scales appropriate for the information at risk will not happen without productive collaboration between the archival community and the worlds of software engineers and

commercial information technology enterprises. Such dysfunction was eloquently described by [Snow 1969] and continues to thrive today. [Schemer 2011]

REFERENCES

[Anon. 2004] *The National Digital Information Infrastructure and Preservation Program (NDIIPP).* Library of Congress, 1975. See also *Update to the NDIIPP Architecture: Version 0.2*, 2004, §7.

[Cohen 2011] Rabinovici-Cohen, Simona, Mary G. Baker, Roger Cummings, Sam Fineberg, and John Marberg, *Towards SIRF: Self-contained Information Retention Format*, Proceedings of the Annual International Systems and Storage Conference, May 2011, Haifa, Israel.

[Egger 2006] A. Egger, *Shortcomings of the Reference Model for an Open Archival Information System (OAIS)*, TCDL Bulletin 2(2), 2006, http://www.ieee-tcdl.org/Bulletin/v2n2/egger/egger.html

[Gladney 2006] Gladney, H.M. *Principles for Digital Preservation,* Comm. ACM 49(2), 111-116, February 2006. This sketch has been elaborated in Gladney's *Preserving Digital Information*, Springer Verlag, 2007, ISBN 978-3-540-37886-0.

[Gladney 2007] Gladney, H.M. *Digital Preservation in a National Context: Questions and Views of an NDIIPP Outsider*, D-Lib Magazine 13(1/2), January 2007.

[OAIS 2009] *Reference Model for an Open Archival Information System* (OAIS), Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August 2009

[Prom 2009] Prom, Chris. *Trustworthy Digital Objects,* blog posting at http://e-records.chrisprom.com/?p=620, December 21, 2009.

[Schermer 2011] Schermer, Michael. *The Believing Brain: Why science is the only way out of the trap of belief-dependent realism*, Scientific American 305(1), 85, July 2011.

[Snow 1969] Snow, Charles Percy. *The Two Cultures: and a Second Look: an Expanded Version of the Two Cultures and the Scientific Revolution.* 1969. ISBN 052109576X

[Snyder 2011] Snyder, Laura J. *The Philosophical Breakfast Club.* Broadway Books, 2011, ISBN 978-0-7679-3048-2