By H.M. Gladney

PRINCIPLES FOR DIGITAL PRESERVATION

Focusing on end users' needs rather than those of archiving institutions.

Most information is now "born digital" and much is disseminated only in digital form. However, little of this is provided in forms that ensure its perpetual intelligibility or that include evidence that it can be trusted for sensitive applications.

Many articles about digital preservation come from the cultural heritage community, which is somewhat unfortunate as the IT community is not involved. The NDIIPP (National Digital Information Infrastructure Preservation Plan) [6] expresses urgency for preserving authentic digital works. However, since the 1995 appearance of *Preserving Digital Information* [2], little progress has been made toward technology for reliable preservation of substantial collections [7, 11].

Most of the preservation literature draws its examples from scholars' and artists' interests. We anticipate that the needs expressed will expand to those of businesses wanting safeguards against diverse frauds, attorneys arguing cases based on the probative value of digital documents, and our own dependencies on personal medical records.

This article deals exclusively with challenges created by technological obsolescence and the demise of information providers. Preservation know-how was summarized by Thibodeau in 2002 by observing that proven meth-

ods for preserving and providing sustained access to electronic records were limited to the simplest forms of digital objects. Even in those areas, proven methods were incapable of being scaled for the expected growth of electronic records. Furthermore, archival science had not responded to the challenge of electronic records sufficiently to provide a sound intellectual foundation for articulating archival policies, strategies, and standards for electronic records [10]. Here, a design that addresses all technical issues reported in the preservation literature is described.[1]

What Would a Preservation Solution Provide?

What might someone a century from now want of information stored today? Figure 1 suggests users' perspectives and helps illuminate preservation reliability questions. In addition to what content management offerings[2] and published metadata schema[3] already provide, a complete solution would:

- Ensure that a copy of every preserved record survives as long as desired;
- Ensure that authorized consumers can find and use any preserved record as its producers intended, doing so without impact from errors introduced by third parties;
- Ensure that any consumer can decide whether information received is sufficiently trustworthy for his application; and
- Hide technical complexity from end users (both information producers and consumers).

Viable solutions will allow repositories and their clients to use deployed content management software without disruption.

Challenges Exposed by Prior Work

Information in physical books, on other paper media, and in other analog forms cannot be copied without error and always contains accidental information that digital representations can avoid. Perfect digital copying is possible, and contributes both to the challenge of preserving digital content and to its solution. Preservation can be viewed as a special case of information interchange—special because information consumers can no longer obtain information producers' responses about missing information or puzzling aspects.

Pervasive Focus on Repositories. Much preservation literature focuses on so-called "trusted digital repositories." Recent articles [9] amplify prior calls for criteria to be used in audits that might lead to public certification that an institution has correctly executed sound preservation practices. However, to execute partly human procedures faithfully over decades would be difficult and expensive. Repository-centric proposals betray problems that call the direction into question. Fundamentally, they depend on an unexpressed premise—that exposing an archive's procedures can persuade its clients that its content deliveries will be authentic. Such procedures have not yet been described, much less justified as achieving what their proponents apparently assume. In addition, audits of a digital archive—no matter how frequent—cannot prove that its contents have not been improperly altered by employees or hackers many years before a sensitive record is accessed. Another problem is that the new code needed for digital preservation is likely to be mostly workstation software, not server software, so the people focusing on repositories will find it difficult to design solutions.

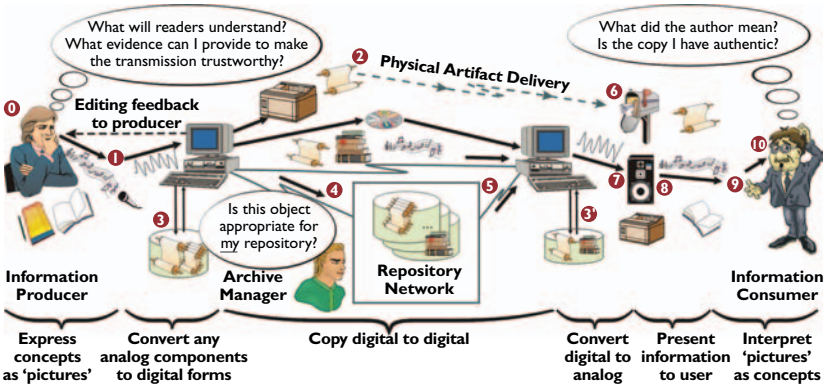

Figure 1. Documentary information interchange and repositories (the object numbering is taken from [5]).

The topical literature is replete with epistemological weaknesses. For instance, many of its references to trust are unmodified (unconstrained). Young children trust unconditionally; anyone else who does so is commonly considered childish. The mature formulation has the pattern, "X trusts Y to accomplish some

[1]Designs cited here have been published in *ACM Transactions on Information Systems*.
[2]Content management is not discussed in this article because archival needs can be satisfied by available software with at most modest and obvious extensions.
[3]These include general schema proposed for standardization, such as METS sponsored by the Library of Congress, and many topic- or discipline-specific extensions. An October 2005 Web search for material with "metadata schema" in their titles yielded over 300 hits.

action Z, or to refrain from some action or behavior W." If the authors of trusted digital repositories articles would adopt this pattern and consider the consequences of each Z and each W, they would materially advance their professed agendas.

As an objective, 'trusted' is misleading. Instead, one should focus on encapsulating information so that it is *trustworthy*.

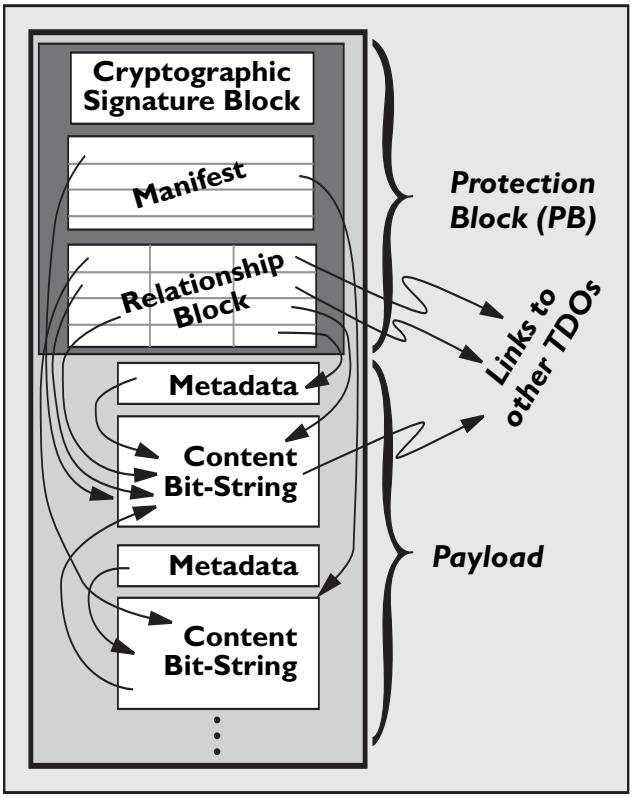

Figure 2. A trustworthy digital object (TDO).

What's 'the Original'? What's 'Authentic'?

In casual conversation, we often say that the copy of a recording is authentic if it closely resembles the original. But consider, for example, an orchestral performance, with sound reflected from walls entering imperfect microphones, signal changes in electronic recording circuits, and so on, until we finally hear a television rendering. Which of many different signal versions is the original?

Difficulties with 'original' and 'authentic' are conceptual. Nobody creates an artifact in an indivisible act. What people consider to be an original or a valuable derivative version is someone's subjective choice, or an objective choice guided by subjective social rules. We can, however, describe any version objectively with provenance metadata that expresses everything important about its creation history.

Conventional definitions, such as "authentic: of undisputed origin; genuine," do not help operationally. For signals, for material artifacts, and even for natural entities, the definition shown in the sidebar here captures what people mean when they say 'authentic'.

Each $\mathbf{T_k}$ represents a transformation that is part of a Figure 1 transmission step. To preserve authenticity, the metadata accompanying the input in each transmission step should be extended by including a $\mathbf{T_k}$ description. This metadata might identify the author of each $\mathbf{T_k}$ choice and other circumstances important to consumers' judgments of authenticity. Each eventual consumer will decide for himself whether the available evidence is sufficient for his particular purposes.

Preserving Dynamic Behavior. A prominent collaborative archivists' project suggests conceptual difficulty with preserving "dynamic objects" (representations of artistic and other performances) digitally [1]. We see no new or difficult technical problem; what differs for different object types is merely the ease of changing them.

A repeat R(t) of an earlier performance P(t) would be called authentic if it were a faithful copy except for a constant time shift from some t_{start}, that is, if $R(t)=P(t-t_{start})$. This seems simple enough and capable

Defining 'Authentic'

Given a derivation statement R, "V is a copy of Y (V=C(Y))," a provenance statement S, "X said or created Y as part of event Z," and a copy function, "$C(y) = T_n (\ldots (T_2(T_1(y))))$,"

we say that V is a *derivative* of Y if V is related to Y according to R.

We say that "by X as part of event Z" is a *true provenance* of V if R and S are true.

We say that V is *sufficiently faithful* to Y if C conforms to social conventions for the genre and for the circumstances at hand.

We say that V is an *authentic copy* of Y if it is a *sufficiently faithful derivative with true provenance*.

of describing any kind of performance. Its meaning is simpler for digital records than for analog recordings because digital records already reflect the sampling errors of recording performances that are continuous in time. The archivists expressing difficulty with dynamic digital objects do not express similar uncertainty about analog recordings of music.

Trustworthy Digital Object (TDO) Methodology

The TDO proposal focuses on methods for making the authenticity of preserved digital objects reliably testable and for assuring that eventual users will be able to render or otherwise use their contents. The objectives suggest solution components that can be nearly 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 [8].
- Schema for packaging a work together with metadata that includes provenance assertion and reliable linking of related works, ontologies, rendering software, and package pieces with one another.
- Standard bibliographic metadata and 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 computer environment.

To prepare the TDO that represents a work (see Figure 2), an editor converts each content bit-string into a durably intelligible representation and collects the results, together with standardized metadata, to become the TDO payload. In addition to its payload, each TDO has a protection block into which a human editor loads metadata and records relationships among its parts, and between it and other objects. The final construction step, executed at a human agent's command, is to seal all these pieces within a single bit-string with a *message authentication code*. 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. TDO metadata includes identifiers for the version and for the set of versions of the work and 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 and each cryptographic certificate is authenticated by a recursive certificate chain.

In the past, wax seals impressed with signet rings were affixed to documents as evidence of their authenticity. A contemporary digital counterpart is a message authentication code firmly bound to each important document. The structure and use of each TDO, emphasizing the metadata portions suggested by Figure 2, is described in [3]. The design includes the following features:

- Each TDO contains its own worldwide eternal and unique identifier and its own provenance metadata, and is cryptographically sealed to prevent undiscoverable changes;
- References to external objects are accompanied by their referents' message authentication codes;
- Certification keys are themselves certified. This recursion is grounded in the published and annually changed public keys of institutions that people trust to be honest witnesses. The stored results of this process chain constitute durable evidence of the TDO's publication date;
- Each person that edits a work being prepared for archival deposit nests or links the version he started with, thereby creating a reliable history;
- Each participant in the creation sequence usually is, or readily can become, acquainted with his predecessor and his successor. Thus the public keys that validate authorized version deliveries can readily be shared without depending on a Public Key Infrastructure (PKI) certificate authority. This arrangement avoids well-known PKI security risks.

Content represented with relatively simple and widely known data formats can be saved more or less "as is." For other data formats, [5] teaches how to encode any kind of content bit-string suggested by Figure 2 to be durably intelligible or useful. Its features include:

- That we enable each information producer to separate irrelevant information, such as operating system details, from information essential to his intentions, encoding only what's essential;
- Rewrite to the code of a Turing-complete virtual machine (extended to handle concurrency and real-time services)—an application of the Church-Turing thesis that any program or rule set producing a finite sequence can be implemented by a simple machine.
- And that such machines can themselves be

CONTENT REPRESENTED WITH RELATIVELY SIMPLE AND WIDELY KNOWN DATA FORMATS CAN BE SAVED MORE OR LESS "AS IS".

described completely and unambiguously.

A producer typically tries to encode information so that each consumer can read or otherwise use the content. In an ideal scenario as depicted in Figure 1, perfection would be characterized by the consumer understanding exactly what the producer intended to communicate. However, in addition to the consequences of human imperfections of authors and editors, the 0→1 and 9→10 steps suffer from unavoidable language limitations. (Jargon, expectations, world views, and ontologies are at best imperfectly shared. For example, I cannot tell you what I mean. I cannot know how you interpret what I say.)

Such difficulties originate in the theoretical limits of what machines can do. How we might mitigate them will be discussed in future articles. Philosophical arguments that TDO methodology accomplishes as much as any mechanical method can accomplish toward preserving digital information, and that it attempts no more are presented in [4]. A second work in progress examines what information producers can do to minimize eventual consumers' misinterpretations, given that communication invariably confounds intentional with accidental information.

Discussion

Premature digital preservation deployment would risk that flaws might not be discovered before large expenditures are made to create archival holdings of uncertain quality. Errors might distort meanings (for texts) or behaviors (for programs). The questions reach into epistemology—the philosophical theory of what can be objectively known and reliably communicated, in contrast to what must forever remain subjective questions of belief or taste. We are therefore reluctant to implement pilot installations until we have considered the applicable philosophy thoroughly and until experts have had the opportunity to criticize TDO design.

What's Missing from the U.S. Digital Preservation Plan? Engineers want questions that can be answered objectively. They expect plans to be clear enough so that every participant and every qualified observer can understand what work is committed and can judge whether progress is being achieved.

We expect a plan to articulate concisely each objective, the resources needed to meet it, commitments to specific actions, a schedule for each delivery, and a prescription for measuring outcomes and quality. If the plan is for a large project, we expect it to be expressed in sections that separate teams can address relatively independently. If the resources currently available are inadequate, we expect the plan to identify each shortfall. Finally, if a team has already worked on the topic, we expect its plan to list its prior achievements.

NDIIPP funding is commensurate with that for all foreign preservation work combined. Unfortunately, the technical portions of [6] contain little more than vague generalities and decade-old ideas. It identifies few technical specifics, no target dates, and few objective success measures. Engineers will find little to work with. Later publications do not repair its weaknesses. This is troubling for an initiative launched six years ago.

Competitive Evaluation. Firm assertions of TDO packaging advantages over alternatives would be premature before we have deployed a complete pilot. Ideally, we would compare our design to alternatives. However, nobody has designed one. Notwithstanding such uncertainties, we believe that, in addition to satisfying our starting objectives, TDO support infrastructure will exhibit the following desirable characteristics:

- Consumers will be able to evaluate TDO content authenticity without help from administrators.
- Metadata-to-object dissociation will occur at most

WHAT WILL MAKE IMPLEMENTATIONS EASY TO TAILOR IS THAT GOOD TOOLS EXIST FOR XML. WHAT WILL MAKE THEM SCALABLE IS THAT TDO STRUCTURE IS RECURSIVE AND USES LINKS EXTENSIVELY.

rarely, and will be discernible when it happens.
- Correct information delivery will be insensitive to Internet security risks. Objects might disappear, but if a TDO is delivered, its integrity can be validated.
- Identifier creation servers will not be needed. Specialized name-to-location resolvers might not be needed, because popular Web crawlers could readily include the function. Bit-string replication for robust TDO storage can include completely automatic management of Internet directories.
- Collection management can be simplified by exploiting TDO link reliability. If metadata is sufficiently standardized, users will be able to use automatic tools to create personal digital library catalogs that suit their special needs and preferences.
- TDO software can be brought into service without disrupting installed digital libraries. Preservation objects can be stored, cataloged, and served by any of several extant content manager offerings.

What will make implementations easy to tailor is that good tools exist for XML. What will make them scalable is that TDO structure is recursive and uses links extensively.

Conclusion

Most preservation literature emphasizes the perspectives of archiving institutions. This article and supporting TDO reports focus on end users' needs because these have precedence over repository needs. Principles for a TDO design have been articulated here to address every technical problem and requirement identified in the literature. The central elements are an encapsulation scheme for digital preservation objects and encoding using extended Turing-complete virtual machines. Correct TDO implementations will allow preservation of any type of digital information and will be as efficient as any competing solution.

Critical examination of this work by readers is encouraged and public discussion is called for because "getting it right" is too important for anything short of complete transparency. C

References

1. Duranti, L. The long-term preservation of the dynamic and interactive records of the arts, sciences and e-government. *Documents Numerique 8*, 1 (2004), 1–14.
2. Garrett, J. et al. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Commission on Preservation and Access and The Research Libraries Group, 1995.
3. Gladney, H.M. Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Trans. Info. Sys. 22*, 3 (July 2004), 406–436.
4. Gladney, H.M. Trustworthy 100-year digital objects: Syntax and semantics—tension between facts and values; eprints.erpanet.org/archive/00000051/.
5. Gladney, H.M. and Lorie, R. Trustworthy 100-year digital objects: Durable encoding for when it's too late to ask. *ACM Trans. Info. Sys. 23*, 3 (July 2005), 299–324.
6. Library of Congress. *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*, 2003; www.digitalpreservation.gov/repor/ndiipp_plan.pdf.
7. Marcum, D.B. Research questions for the digital era library. *Library Trends 51*, 4 (Spring 2003), 636–651.
8. Reich, V. and Rosenthal, D. LOCKSS: A permanent Web publishing and access system. *D-Lib Magazine 7*, 6 (June 2001).
9. Ross, S. and McHugh, A. Audit and certification of digital repositories; Dale, R. Making certification real: Developing methodology for evaluating repository trustworthiness. Both articles in *RLG Digi-News 9*, 5 (Oct. 2005).
10. Thibodeau, K. Knowledge and action for digital preservation: Progress in the U.S. Government. In *Proceedings of DLM-Forum 2002* (2002), 175–179.
11. Waters, D. Good archives make good scholars: Reflections on recent steps toward the archiving of digital information. In *Proceedings of the Council on Library and Information Resources* pub107, (2002).

H.M. Gladney (hgladney@pacbell.net) is the president of HMG Consulting in Saratoga, CA, and the publisher of *Digital Document Quarterly* (home.pacbell.net/hgladney/ddq.htm).