Critique of Architectures for Long-Term Digital Preservation

H.M. Gladney
HMG Consulting

Abstract: *Trusted Digital Repositories* (TDRs) and *Trustworthy Digital Objects* (TDOs) seem to be the only generic digital preservation methodologies proposed. Before any preservation method is recommended for wide use, it should be exposed to searching analysis.

Evolving technology and fading human memory threaten the long-term intelligibility of many kinds of documents. Furthermore, some records are susceptible to improper alterations that make them untrustworthy. We argue that the TDR approach has shortfalls as a method for long-term digital preservation of sensitive information. For specificity, we discuss a particular implementation.

TDO methodology addresses these needs, providing for making digital documents durably intelligible. It uses EDP standards for a few file formats and XML structures for text documents. For other information formats, intelligibility is assured by using a virtual computer. To protect sensitive information—content whose inappropriate alteration might mislead its readers, the integrity and authenticity of each TDO is made testable by embedded public-key cryptographic message digests and signatures. The authenticity of the keys is protected recursively in a social hierarchy grounded by publishing keys of well-known institutions.

A TDO is a specific kind of OAIS Archival Information Package convenient for sharing among repositories. The content and metadata of properly constructed TDOs are sufficient for creating the usual kinds of catalog records and search indices during repository ingestion.

Comparison of TDR and TDO methodologies suggests differentiating near-term preservation measures from what is needed for the long term. The proper focus for long-term preservation technology is signed packages that each combine a record collection with its metadata and that also bind context—*Trustworthy Digital Objects.*

If all that stuff was worth creating, surely some of it is worth saving!

Introduction

For much-discussed professional and social reasons,[1] we want to deploy tools for making any digital content whatsoever perpetually usable in ways that please its eventual users. We want these tools to be simple non-disruptive extensions of widely used digital content management (CM) services. The current article treats long-term digital preservation (LDP) as a compatible addition to day-to-day CM.

If a digital preservation method becomes popular, it will be applied to very large numbers of records—perhaps several hundred million new records every year, and to hundreds of record formats. These numbers are much larger than librarians and archivists will be able to handle properly with methods in use today. To the extent possible, every citizen should be enabled for creating, storing, and using durably reliable digital objects.

In view of the scale and inherent risks, before any digital preservation method is widely deployed, it should be subjected to rigorous critical examination based on the soundest and most dispassionate intellectual base available. Without this, subtle errors, avoidable ambiguities, and unintended limitations are unlikely to be discovered until long from now, when weaknesses and potential frauds might be much

[1] Francine Berman, *Got Data? A Guide to Data Preservation in the Information Age*, Comm. ACM 51(12), 50-56, 2008.

more expensive to correct than they are today. This examination should seek opportunities to reduce complexity that might mislead readers.

Technology for near-term preservation needs flexibility for software improvements. In contrast, technology for long-term preservation needs to be insensitive to changing technology and infrastructure. It therefore proves helpful to distinguish near-term preservation from long-term preservation.

What Is the Challenge?

> What is the meaning of *preservation*? Does the meaning change when it is applied to electronic rather than paper-based records? ... Will current strategies for preserving electronic records ensure longevity and authenticity? ... Have effective cost models been developed?[2]

The notion of a digital preservation theory[3,4] is recent, being mentioned earlier than 2007 only in comments about shortfalls. What do people expect of a theory to think it useful? To be most helpful for engineering, a theory would exhibit at least the following characteristics.

- It would be based on broad fundamental theory that is widely accepted as germane and successful.
- It would differentiate its topic from nearby topics, particularly topics that already have good theories.
- It would include a statement of objectives that does not limit the range of responsive system designs. Proposed solutions would be stated as external properties of systems or components.[5] Optionally, solution designs might be attached to demonstrate practicality.
- It would take into account the best prior work, exposing its reasons for not using earlier ideas.
- It would factor its sub-topics into portions that interact relatively weakly, so that autonomous engineering teams could contribute to practical solutions with only modest collaboration.
- If the theory refers to human participants, such references would be to human roles rather than to social or organizational positions.
- It would be written to be comprehensible by every interested reader.

What distinguishes a *theory of digital preservation* from a *software architecture*?

What Do We Mean by Long-Term Digital Preservation?

> Digital curation involves the management of digital objects over their entire lifecycle, ranging from pre-creation activities wherein systems are designed, file formats and other data creation standards, [and] capture of evolving contextual information for digital assets housed in archival repositories.[6]

Figure 1 depicts a model helpful for talking about modern communication objectives and mechanisms. For a comprehensive treatment, we must deal with the entire channel from each writer's knowledge **0** to each reader's judgments **10**, asking and answering at least the following questions.

- How can authors and editors ensure that eventual readers can interpret information saved today even though nobody can predict future computing technology?
- How can we avoid losing the last copy of any preserved object?
- How can we make authenticity evidence sufficiently reliable, especially for sensitive documents?
- How can we help writers provide metadata as a by-product of their efforts, thereby shifting cost from libraries and other institutions that cannot handle today's information flood?

---

2 Michèle Cloonan and Shelby Sanett, *Preservation Strategies for Electronic Records: Where We Are Now—Obliquity and Squint?* Am. Archivist 65(1), 70-106, 2002.

3 R. Moore, *Towards a Theory of Digital Preservation*, Intl. J. Digital Curation 3(1), 63-75, 2008.

4 Paul Watry, *Digital Preservation Theory and Application: Transcontinental Persistent Archives Testbed Activity,* Intl. J. Digital Curation 2(2), 41-68, 2007.

5 In a Dec. 2008 IEEE podcast, Grady Booch asserted that "as long as a system provides the right answers at the right time with maintainability, dependability, changeability, and so on, end users couldn't care less about what's behind the curtain making things work". See http://csdl.computer.org/rss/podcasts/audio/onarch.xml.

6 Christopher A. Lee and Helen R. Tibbo, *Digital Curation and Trusted Repositories: Steps Toward Success*, J. Digital Information 8(2), 2007.

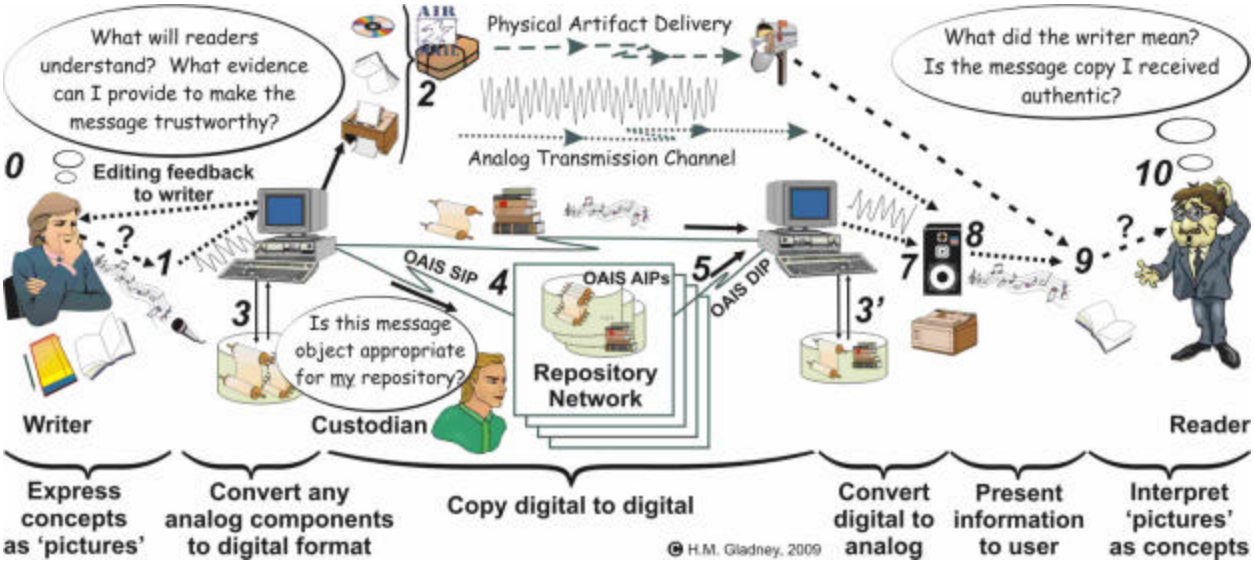

Figure 1: Model of communicating documentary information (messages) (PDI §1.4)

This model hides processing flow aspects important for conventionally published documents and for business records that might be preserved. These aspects are suggested for bureaucratic records by the Figure 2 Records Administrator, whose position might be replaced by a periodical or book editor for scholarly and artistic works.

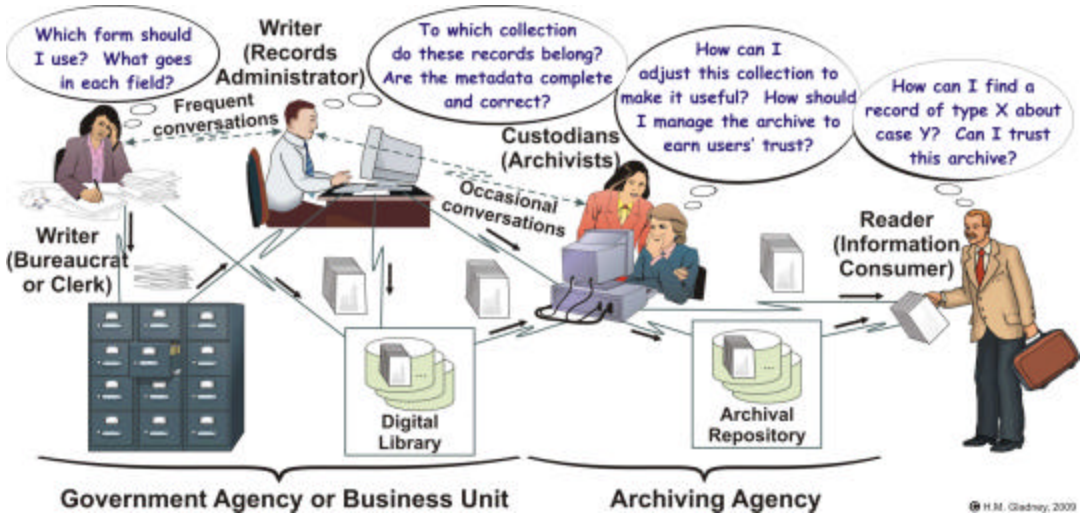

Figure 2: Model of bureaucratic records administration

Digital preservation is the totality of measures required and undertaken to mitigate information unreliability caused by machine failures, human misbehavior, technological obsolescence, and other ravages of time.

Long-term digital preservation (LDP) is a special case of communicating human thoughts and history. Precisely, LDP is the case in which it is impossible for a writer to converse with an eventual reader and impossible for the reader to clarify uncertainties by asking the writer or the writer's contemporaries.

Near-term digital preservation (NDP) has to do with the short period in which a collection's authors, editors, and custodians are available to answer questions about its interpretation, origins, and other details. It is practical to think of NDP as addressing the 5- to 10-year period in which any digital repository manager takes responsibility for the quality of his institution's services.

Scope Limitations

How to preserve what has been represented in digital form cannot be completely described or properly analyzed in an article of acceptable length without depending heavily on what is already written and also limiting the discussion to generally applicable methods. The treatment below is therefore limited to sketches that neither include technical details nor justify claims made, but instead cite literature providing the missing information. Readers not already familiar with this literature and wanting proper explanations will have to follow these citations. In particular, topics already handled in a book, *Preserving Digital Information*,[7] will be cited by identifying the book section in the style illustrated by "[PDI §9.4]".

Day-to-day information production and management include voluminous expressions of relationships and other descriptions that admit unspecified additions. Digital CM services refined since commercial digital library services emerged about 15 years ago are widely used. We treat only additions for preservation.[8] For similar reasons, we do not delve into risks to cryptographic measures and their mitigations.

What follows might obscure its main argument if it included potential optimizations that are often obvious or different communities' idiosyncratic preferences. We therefore limit discussion to aspects common to all kinds of information whatsoever and to methodology robust enough to handle at-risk information—documents for which misunderstandings by eventual readers are likely and also records whose inappropriate alteration might damage eventual readers, for instance by exposing them to fraud.

Our analysis insists on reliable communication between today's writers and eventual readers. It also insists on good evidence of information integrity and authenticity—evidence that any reader can test to decide whether information received is sufficiently trustworthy. Discussing optimizations to save people's time and machine cycles is deferred—especially optimizations for limited information classes.

Specifically, the scope of the current analysis is constrained to:

- Documents/records not yet shared—after we have completed this, we can backtrack.
- Basic methodology that will work for all content types—optimizations can be treated in other articles. We say little about handling differences for different classes of records and documents. [PDI §9.4]
- Riskiest cases—later we can consider shortcuts for less exposed instances.
- Technical aspects, leaving topics such as selection of what to save, how to plan services, and managing document collections to other authors.
- Not repeat well-known methods of day-to-day digital CM and archiving.

Synopsis

LDP can be treated as an extension to near-term CM services. This can be a way of packaging digital objects to contain or to refer reliably to context that their creators believe essential and a way of representing information[9,10] and critical metadata[11] so that our descendants will surely be able to interpret them correctly. What follows starts by sketching twenty years of scholarly literature.

Proper evaluation of our claims would be impossible without an impartial, fundamental theory—a basis sought by other authors.[12] We formulate and apply a preservation theory by analogy with mathematics,

---

7 H.M. Gladney, *Preserving Digital Information*, Springer Verlag, 2007, ISBN 978-3-540-37886-0.

8 H.M. Gladney, *A Storage Subsystem for Image and Records Management*, IBM Systems Journal *32(3)*, 512-540, 1993. Ideas pioneered in this work are today used in thousands of CM installations.

9 Steen S. Christensen, *Archival Data Format Requirements*, report from the Denmark Royal Library, 2004, available at http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf.

10 T. Phelps and P. Watry, *A No-Compromises Architecture for Digital Document Preservation*, Research and Advanced Technology for Digital Libraries 9th European Conference, ECDL2005, 266-277, 2005.

11 Brian Lavoie and Richard Gartner, *Preservation Metadata*, DPC Technology Watch Report 05-01, 2005, available at http://www.dpconline.org/docs/reports/dpctw05-01.pdf.

Judith Pearce, David Pearson, Megan Williams and Scott Yeadon, *The Australian METS Profile: A Journey about Metadata*, D-Lib Magazine 14(3/4), 2008.

Muriel Foulonneau and Jenn Riley, *Metadata for Digital Resources,* Chandos, 2008, ISBN 108334-302-8

12 Bruce W. Dearstyne, *The Archival Enterprise: Modern Archival Principles, Practices, and Management Techniques*, 1993, ISBN 0-8389-0602-8.

choosing end user requirements as the counterpart of axioms, philosophical theory of knowledge and language as the counterpart of rules of reasoning, and software design as the counterpart of theorems. This theory would be insufficiently crisp without attention to the meanings of key words and phrases. Even "digital preservation" has different meaning to different authors, obscuring the boundary between quotidian content management and preservation activities.

Three models are particularly helpful: a model of message transmission between writers and readers, depicted in Figure 1; a model of digital repository infrastructure, depicted in Figure 3; and a model of annotated digital content, suggested in Figure 4. A full version of each figure is included for completeness, even though the current article explains and draws on only part of what is depicted.

After extending human requirements statements with engineering considerations, our analysis considers the most prominent digital preservation model—the *Trusted Digital Repositories* (TDR) approach. To make the analysis explicit, it uses an example—work by the DICE (*Data-Intensive Computing Environments*) group in support of the U.S. NARA Electronic Records Archives project. It argues that TDR methodology insufficiently protects eventual users of many records.

It then sketches TDO methodology,[13] providing only enough description so that readers can understand the services possible and how these can be made robustly durable. The article argues that properly applied TDO technology would satisfy all identified preservation objectives.

A Conceptual Base for Digital Preservation

What can we preserve for future generations? We might hope that readers understand exactly what writers mean. The Figure 1 arrows with question marks suggest that communicating intended meaning unambiguously is impossible in principle. The other arrows depict communications that might include syntactic transformations that can be unambiguously communicated.

Today's Digital Preservation Research and Development

> [Many] digital libraries have been created, frequently by means of one-time grants and other soft funding. Sustaining these digital libraries … has become a critical concern for [their] stakeholders.[14]

Much has been written since digital preservation challenges were articulated in 1996.[15] More than a dozen annual conferences address digital archives. Study groups and task forces abound.[16]

Digital preservation work of the past two decades pays more attention to what might interest curatorial staff[17] than to what might please their clients—writers and readers. This might be related to the fact that so many authors are employees of repository institutions. These authors seem to have assumed without question the proposition that digital preservation can be accomplished by modest modification of repository procedures and management precepts.

> The academic research library has been predominant in collecting and preserving text-based scholarly literature, but it has not been the primary home for statistical data, cartographic materials, manuscript collections, prints and photographs, film, broadcast television and radio, folklore documentation, natural history specimens, and an overwhelming preponderance of primary source materials needed by scholars … Perhaps a preliminary answer to the question "What are the core functions of the research library with respect to collecting, preserving, and making accessible resources for scholarship?" might be that research libraries will be stewards of some sectors of the information universe, but they will not be the same sectors as before.[18]

---

13 H.M. Gladney, *Principles for Digital Preservation*, Comm. ACM 49(2), 111-116, February 2006.

14 Katherine Skinner and Martin Halbert, *Strategies for Sustaining Digital Libraries*, 2008, ISBN 0-977-29941-4, available on-line from http://metascholar.org/publications/StrategiesforSustainingDigitalLibraries.pdf.

15 John Garrett et al., *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information,* Commission on Preservation and Access, 1996.

16 Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustaining the Digital Investment: Issues and Challenges*2008, available from http://brtf.sdsc.edu/econ_sustainability.html.

17 Priscilla Caplan, *The Preservation of Digital Materials*, Library Technology Reports 44(2), 2008, available at https://publications.techsource.ala.org/products/archive.pl?article=2614.

18 Council on Library and Information Resources, *No Brief Candle: Reconciling Research Libraries for the 21st Century*, CLIR pub. 142, August, 2008, available via http://www.clir.org/pubs/abstract/pub142abst.html.

For business records and for family legacies, neither research libraries nor government archives are ready to become society's first choice of repositories, even though diverse objects are housed by similar institutions.[19] The Library of Congress, which holds many kinds of materials, is unlikely to be emulated by most research libraries, and might itself not extend to all kinds of materials that interest citizens.

> There is a renewed, beleaguered feeling that scholarship struggles valiantly against its own history and the new tides of change … Any new methodology that does not fit within the current framework … is met with anything from skepticism to exasperation, from giddiness to despondency. New approaches to how research is carried out must struggle against centuries of inertia.[20]

Professional archivists' literature includes calls for theory. Dollar calls for comprehensive study and reviews projects underway when he wrote.[21] Dearstyne summarizes the challenges, suggesting that, "Archival and records management principles and practices, developed over many years and embedded in the genes of professional associations, need to be modified or replaced outright to fit the electronic world. For insurance, archival concepts of provenance, original order, and appraisal-through-analysis need substantial change before they will map well to the new situation."[22]

Many people are delighted by personal opportunities created by the Information Revolution.[23] However, some scholars and librarians seem dismayed by changes that they see as threatening their institutions and professional status as intermediaries between citizens and information.

> [W]hat about preservation for the long-term? Well, [an Association of Research Libraries] report answered, the short-term strategies "are a bridge to the emerging solutions that are being developed to ensure long-term availability and access"—that is, a bridge to something hoped for rather than in place. The report admitted that "standards, guidelines, and best practices for producing and maintaining digital facsimiles for the long-term are in the development stage." But … the report insisted that, "best practices are in place to ensure that digital objects are being managed in such a way that keeps them safe now and allows us to implement long-term strategies as they emerge." The report further minimized the long-term preservation problem by flattering libraries for their historical ingenuity, as follows:
>
> > Ensuring high-quality image capture and providing for the long-term viability of digital objects is an admitted challenge, but the library profession has a long history of developing standards and best practices in order to support sustainable operations and facilitate inter-institutional collaboration. This tradition provides confidence that digital preservation challenges will be met.
>
> In short, we kept on producing digital resources because we had to while whistling in the dark about their long-term preservation.[24]

Notwithstanding their accepting digitization of deteriorating manuscripts and prints as an acceptable preservation method, research librarians have not achieved a confident sense of direction.

What Can Be Communicated?

The greatest difficulties of communicating reliably have little to do with technology. [PDI §3.3] Personal thoughts, particularly those we call *subjective* (having to do with opinions, tastes, judgments, purposes, or feelings) cannot be shared, except perhaps incompletely. However, after someone speaks or writes, his words are *objective* evidence of his thoughts.

Work between 1880 and 1960 provides insights essential to comprehending the boundary between what can be known and what can be told. Ludwig Wittgenstein,[25] Rudolf Carnap,[26] Ernst Cassirer,[27] Willard

---

19 For instance, El Archivo General de Indias de Sevillia houses commercial documents, such as ships' bills-of-lading, from the 16th-century Spanish exploitations of the Americas. In 1985, about 10% of this was scanned to become one of the earliest digital Libraries.

20 Bradley Daigle, *Digital Libraries and the Fate of Faculty Scholarship*, in Skinner, loc. cit. footnote 14.

21 Charles M. Dollar, *Authentic Electronic Records: Strategies for Long-Term Access,* 2000, ISBN 0-9700640-0-4. *Introduction* available at http://www.mybestdocs.com/dollar-c-preservation%20book.html#5

22 Bruce W. Dearstyne, *Strategic Adaptation to Changing Realities,* in *Effective Approaches for Managing Electronic Records and Archives*, 2002, ISBN 0-8108-4200-9**.**

23 Thomas L. Friedman, *The World Is Flat: A Brief History of the Twenty-First Century,* 3.0, 2007, ISBN 0-312-42507-4.

24 Deanna B. Marcum, *The Future of Preservation*, Symposium on the 3-D's of Preservation: Disasters, Display, Digitization, March 2006, available at http://www.loc.gov/library/reports/paris-speech-preservation.pdf.

25 Ludwig Wittgenstein, *Philosophical Investigations: The German Text, with a Revised English Translation*, Blackwell, 1953, ISBN 0-631-23127-7. See particularly the discussions of rules.

Van Orman Quine,[28] and Michael Polanyi[29] built on the work of Emmanuel Kant, Auguste Comte, Heinrich Hertz, Karl Weierstrass, Ernst Mach, Gottlob Frege, David Hilbert, Karl Kraus, and Bertrand Russell. These authors have been so successful in persuading Western scholars to accept and teach their views that their ideas and careful distinctions are often taken for granted as "mere" common sense, and too often then ignored. Wittgenstein sums up philosophical problems with:[30]

> Most of the propositions and questions to be found in philosophical works are not false but nonsensical. Consequently, we cannot give any answer to questions of this kind, but can only point out that they are nonsensical. Most of the propositions and questions of philosophers arise from our failure to understand the logic of our language.
>
> And it is not surprising that the deepest problems are in fact not problems at all.
>
> All philosophy is a 'critique of language'.

An archived record of a 1950 law is valuable because the expression it carries was inscribed to become the particular copy delivered to the archive in 1950 and because this process was managed to ensure the integrity of the preserved copy. The 1950 physical carrier (paper) is valuable today only if the procedures and records of the archive make it unlikely that it has been tampered with. The value of the paper is created by its evidentiary role. The paper and the archive exist in order to provide evidence for claims about a conceptual abstraction—a pattern.

To be eligible for copyright or for archiving, information must be fixed. The essential core of copyright protection is that, if a work has once been represented in tangible form, copyright protects the originator's beneficial rights for using the symbolic pattern represented.

> Fire swept through the converted grain silo that Naomi Marra has called home … Feared lost among the charred ruins is the last extant copy of her lyric ode, *Ruthless Boaz*. … [D]evotees hope that, following her many public declamations of the work, most or all of it may remain preserved in her memory. … Query: Is *Ruthless Boaz* still subject to statutory copyright protection?[31]

Nimmer's fable identifies what can be protected, and therefore much of what is worth preserving—patterns inherent in extant and potential replicas of each document. The essential patterns of a message are those needed to make it be Levy's "talking thing".[32]

The following assertions have been explained.

(1) There is a boundary between what can be mechanized and what must forever remain a human judgment or value decision. [PDI Chapter 3]

(2) Every Figure 1 transmission step potentially transforms the message. [PDI §1.4] Its **0➔1** step involves semantics (representation of meaning with symbols) and its **9➔10** step includes interpretation (understanding what meaning the signal's producers intended).[33] We can say nothing objectively certain about the relationship of a conceptual input **0** or conceptual output **10** to any other object in the transmission channel. In contrast, after a communication has been completed, we can describe precisely the transformation that occurred in each step between **1** and **9**.

(3) There is a distinction between knowledge and the knowledge subset we call "information"—between what one can know and what one can tell. ("To inform" is "to tell". The Figure 1 writer selects part of what he can tell from what he knows as part of effecting the **0➔1** transmission.)

---

[26] Rudolf Carnap, *The Logical Structure of the World,* Open Court, 2003, ISBN 0-8126-9523-2, originally *Der Logischer Aufbau der Welt,* 1928.

[27] Ernst Cassirer, *The Problem of Knowledge: Philosophy, Science, and History Since Hegel*, Yale U.P., 1969, ISBN 0-300-01098-2.

[28] Willard Van Orman Quine, *Word and Object: Studies in Communication*, MIT Press, 1964, ISBN 0-262-67001-1.

[29] M. Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy*, U. Chicago Press, 1958, ISBN 0-226-67288-3, distinguishes between "knowing about" and "knowing how to".

[30] Ludwig Wittgenstein, *Tractatus Logico-Philosophicus, English translation of the original German*, Routledge, 1921, quoting its proposition 4.003.

[31] David Nimmer, *Adams and Bits: of Jewish Kings and Copyrights*, 71 S. Cal. L. Rev. 219, 1998.

[32] David M. Levy, *Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation*, Proc. Third ACM Conf. on Digital Libraries, 152–161, 1998.

[33] An excellent introduction to semantics is available from J. Alberto Coffa, *The Semantic Tradition from Kant to Carnap to the Vienna Station*, 1991, ISBN 0-521-44707.

(4) Computers manipulate symbols that are surrogates for what they mean—symbolic representations of things and circumstances other than themselves. [PDI §3.5] A computer model is good if its pattern follows the pattern of what it stands for. [PDI §3.1] A *meaning* is a *relationship* between a symbol and some performance, concept, fact, or real world object or event. [PDI §4.3]

(5) A computer program is a representation of a rule or collection of rules.[25]

(6) The entities to be preserved are abstractions—patterns inherent in material and digital artifacts.[34] A documentary artifact is made valuable by its evidentiary role for the integrity and authenticity of the pattern it conveys. [PDI §4.6]

(7) When repeating subjective assertions, we should be explicit about whose opinion counts. [PDI §3.4]

(8) What we mean by "authenticity" is that metadata accompanying saved information properly and truthfully identifies who created that information and the attendant historical circumstances.[35]

(9) Every message contains accidental information in addition to what its writer deemed essential.[36] What is *essential* always pertains to somebody's particular purpose. The ambiguity inherent in a message can be reduced by repeating the message in different language, i.e., with a different assemblage of symbols. [PDI §4.1]

(10) For intelligibility, every message requires context, and is fully useful only to somebody who either understands critical context that is tacit[37] or can find whatever additional context is essential.

(11) There is no objective distinction between a "single information object" and a collection. In any kind of information package, references to other packages bind contextual information. [PDI §6.3]

(12) Every message participant wants autonomy.[38] However, sharing information depends on common language that includes EDP standards. Tension between autonomy and sharing cannot be avoided, resulting in subjective decisions about which different participants might disagree.

A reader might want to decide whether a preserved record is sufficiently trustworthy for his use, especially for information that is a tempting target for fraudulent modification. Tiny changes can create havoc and might evade discovery. A fictional example illustrates the risk:

> Queen of the Fairies: The law is clear—every fairy must die who marries a mortal!
>
> Lord Chancellor: Allow me, as an old Equity draftsman, to make a suggestion. The subtleties of the legal mind are equal to the emergency. The thing is really quite simple—the insertion of a single word will do it. Let it stand that every fairy shall die who doesn't marry a mortal, and there you are, out of your difficulty at once! [39]

Perhaps the most troublesome preservation challenge is the uncertain connection between message and meaning suggested by the Figure 1 arrows with question marks. It lurks behind every discussion of communication. It justifies what might be an emerging consensus that:

> [P]reservation in the digital age must be considered at the time of creation. Preservation cannot be an activity relegated to … libraries and archives, but rather must be seen as intrinsic to the act of creation.[40]

Who has the authority to decide what message→meaning mapping is correct? Social rules make the answer dependent on the type of communication and other circumstances. For intellectually innovative works by artists and authors, it is the originator or, when he is not available, diligent scholars who are viewed as "authoritative commentators" by scholarly communities and a skeptical public. For legal documents such as dispositive contracts and testaments, the specific wording has unconditional priority

---

[34] Compare Heather MacNeil, *Trusting Records: Legal, Historical, and Diplomatic Perspectives*, Kluwer Academic, 2000, ISBN 0-7923-6599-2. MacNeil's "essential attributes of a record" is precisely what we call a pattern.

[35] H.M. Gladney and J.L. Bennett, *What Do We Mean by Authentic?* D-Lib Magazine 9(7), 2003.

[36] Teresa Robertson, *Essential vs. Accidental Properties,* Stanford Encyclopedia of Philosophy, 2008, available at http://plato.stanford.edu/entries/essential-accidental/.

[37] M. Polanyi, *The Tacit Dimension,* Dover, 1983, (originally published 1967), ISBN 0-844-65999-1.

[38] Sarah Buss, *Personal Autonomy,* Stanford Encyclopedia of Philosophy, 2002, available at http://plato.stanford.edu/entries/personal-autonomy/.

[39] W.S. Gilbert and A. Sullivan, *Iolanthe*, Act II, 1882.

[40] Library of Congress, *Plan for the National Digital Information Infrastructure and Preservation Program* , 2002, p. 52. See http://www.digitalpreservation.gov/library/resources/pubs/docs/ndiipp_plan.pdf.

over the supposed opinions of the authors, even if those authors are available for questioning. Critical ambiguities are resolved by courts of law.

A distinction not explicitly identified in *PDI*[7] needs to be emphasized—the distinction between a human role and the occupant of an organizational position—a job-holder. What we mean by 'role' is a set of responsibilities and/or activities (in handling records). For instance in Figure 2, 'Archivist' is intended to suggest an organizational position and 'Custodian' to suggest a records-management role. In the current article, the notion of a role is most prominent in how we use 'reader', 'writer', and 'custodian' and how we intend these words to be interpreted. In contrast, the role/job-holder distinction seems not to be made in TDR literature such as the cited articles[3,4]–at least it is not explicit or emphasized.

Another critical distinction, that between *knowledge* and *information* is complicated, subtle, and the subject of philosophic analyses that are too long to be repeated here, apart from a brief reminder about how an individual or an organization acquires knowledge. We call it "learning", the process in which masses of information and experience are combined in the mind(s) of participants, who are said to "become knowledgeable". For our current objectives, it is not necessary to understand how this happens, but merely to recognize and acknowledge that it does happen—that there is a critical difference between knowing something and merely knowing about something. We call the latter "being informed".

These distinctions figure prominently in the TDR discussion below. Figure 2 records administrator responsibilities include *knowing* how long each record collection or portions thereof should be preserved, how and by whom it may and should be used, and everything else important about its content and use. He can and should *inform* archival custodians of those facts pertinent to information preservation.

What's Different from Information on Paper?

> Digital preservation is seen as complex and largely uncontrollable. Preserving books and other cultural objects looks straightforward in comparison.[41]

Discussions of the challenge often compare storing information on paper. They sometimes start by noting how difficult it is to alter information on paper surreptitiously, and how durable good paper is in benign environments. Mentioned less often, further facts are equally telling: that we individually have extensive training for working with paper; that society has deployed immense infrastructure for managing paper and moving it from place to place; and that social and individual quality expectations have increased greatly since sound recording was invented a century ago.

That digital records and documents are readily edited, readily copied, and readily transmitted makes widespread information sharing rapid and inexpensive, except perhaps for confidential documents. Quality expectations for business, personal, and cultural records have increased steadily since consumer electronics appeared, enabling recording accuracy that human beings cannot distinguish from live performances—except perhaps by the live performance glitches or audience noise. Digital technology has been driven to high precision by rising expectations,[42] including that automation should anticipate any potential problem. However, none of this creates any new problem of principle for digital preservation.

21st–century changes from the 20th–century world dominated by paper include that:

- Many business transactions, including most money management and government interactions, are carried out over the Internet. Our health care, business efficiency, government services, and education depend more on recorded information than ever before.
- In North America and Western Europe, nearly everything written today starts in digital form. Much of this is distributed by its writers only in digital form, and loses much in reductions to paper.
- Many citizens are blessed with better education, more leisure time, and better access to cultural involvement than ever before. Many generate personal information in digital form. All expect finding information to be easy and rapid.
- The total information to be managed is immense, has been growing rapidly for two decades, and is likely to continue exponential growth for some time to come.

---

41 Colin Webb, *Barriers or Stepping Stones? Impediments to Digital Archiving and Preservation Programs*, 2003.

42 Sarah McBride, *The Way We'll Watch*, Wall Street Journal, Dec. 8, 2008, accessible at http://online.wsj.com/article/SB122833913230576869.html.

Preserving information on paper follows three models. Information of wide interest is replicated in books, periodicals, and newspapers that are sold and also shared in orderly collections—the library model. The most important business and governmental documents are organized into strictly controlled repositories—the archival model. Financial documents, memoires, and pictures treasured by individuals and families are squirreled away in banks and attics—the private records model.

Between the time when an author completes his work and the time it becomes a library holding, it is typically prepared for publication by editing that adds conventional metadata, for instance in the front matter of a book, the page headings of a newspaper, and the packaging of audio-visual content. No equivalent step is today conventional for most digital works.

Language Describing Information Sharing

The complexity of human communication is reflected in our language. We use different words for different information genres (articles, poems, sound, pictures, video, rules, programs, ...), for their originators (journalist, poet, singer, artist, legislators, programmers, ...), and their recipients (reader, declaimer, audience member, viewer, citizen, PC user, ...). It is both desirable and feasible to discuss communication reliability without hinting at such distinctions, avoiding distracting detail.

To avoid confusions in discussing digital CM and preservation, we take extraordinary care with the meaning of key words and phrases. We are also unusually careful with the distinction between objective facts and subjective matters of conjecture, opinion, or ethics. We can paraphrase Wittgenstein with:

> Most of the proposals found in digital preservation literature are not incorrect, but about a different topic—managing collections of digital documents and records. ... Most of the difficulties relating to digital preservation arise from our failure to understand the logic of our language.
>
> And it is not surprising that achieving reliable digital preservation might in fact *not* be difficult at all.

Key words mean different things in different disciplines, and are sometimes ambiguous even within a domain of discourse. And sometimes different labels are used for roughly the same topic. For instance, 'digital preservation' is a replacement for other widely used names.[43] What used to be called "digital library services" in 1993 was relabeled "content management services" by IBM and other commercial enterprises about five years later. A different community later started to call it "digital archiving".[44] Sound and cinematic engineers refer to the same topic as "media asset management". And research librarians and information scientists started to call it "digital preservation" without anybody seeming to notice that a well-established topic had received a more fashionable name. Such jargon shifts tend to obscure what is known and are closely related to wasteful inattention across disciplinary boundaries. Partly to avoid perpetuating such obfuscation, the current article distinguishes between long-term digital preservation (LDP) and near-term digital preservation (NDP). For other key jargon, see the glossary below.

In the figures, icons for human beings suggest roles rather than job responsibilities or social positions. For instance, an archive manager who usually acts as a records custodian is likely to assume the role of a reader often and to be a writer occasionally. The entity assuming a role might be a machine process acting as the surrogate for a human role. Since human beings act by way of machine processes, this distinction is often irrelevant.

Requirements

Most digital preservation stakeholders will neither know nor care much about technology or about how their information requests are satisfied.[5] The lists below express an opinion about what human creators, custodians, and eventual readers of saved digital information will want and expect.

The critical reader will assess whether these lists are sufficient, whether every requirement expressed is truly necessary, and whether hidden assumptions about solution design have been avoided.

---

[43] Oya Y. Rieger, *Preservation in the Age of Large-Scale Digitization*, CLIR Publication 141, 2008, available at http://www.clir.org/pubs/execsum/sum141.html.

[44] Geoffrey Yeo, *Concepts of Record (1): Evidence, Information, and Persistent Representations,* Am. Archivist 70(2), 315-343, 2007, vehemently discusses difficulties associated with misuse of topical labels.

What Individual Human Participants Will Want

What might an eventual reader of information stored today want or expect? He will want every process to be as automated as possible, freeing him from clerical tasks to focus on what only human beings can accomplish. He will be satisfied if, for whatever record interests him, the following requirements are met.

Requirement name	Requirement statement
Content accessibility	Any eventual reader should be able to retrieve a faithful copy of the message that represents wanted content if he is authorized to do so.
Content intelligibility	Any eventual reader should be able to read or otherwise use message content as its writers intended, without adverse effects caused by inappropriate changes made by third parties.
Content authenticity[35]	Any eventual reader should be able to decide whether a received message is sufficiently trustworthy for his application. The trustworthiness mechanism should not interfere with other security wanted by information owners and custodians. Every message should be secured end-to-end from the writer's release of information to any reader's execution of authenticity tests.
Context reliability	Any eventual reader should be able to exploit embedded references reliably to retrieve contextual information and to validate the trustworthiness of contextual links, doing so recursively to as much depth as he wants.
Method simplicity	Any eventual reader should be able to exercise all this functionality without hindrance by technical complexity that could be hidden.

Every community participates in creating and sharing information.[45] In addition to professional authors, editors, and businessmen, some citizens will want to preserve information.[46]

Berman alludes to "preservation of our most valuable digital information".[65] Who will decide what is most valuable? An essential part of the response is that democratic society will tolerate no institutional arbiters of what is to be preserved. A consequence is that ability to preserve must be made available to anybody who acquires whatever skills are needed. Technologists' role is limited to making the process as simple as possible. Institutional curators can, and should, continue to choose and protect assets that they believe especially valuable and interesting to large numbers of potential readers.

In a nutshell, nobody should need to obtain anybody else's permission or help to preserve information. Anybody should be able to prepare documents for his and our descendants. Information creators will want convenient tools and infrastructure that accomplish the following requirements.

Requirement name	Requirement statement
Content preparation	Any writer should be able to package content of any type to be LDP-ready, doing so in some way that ensures that future readers can use this content as well as current readers are able to.
Content archiving	Any writer should be able to submit such readied content to repositories that promise to save it reliably, possibly in return for a fee for archiving service. (People are willing, in anticipation of death, to pay for storing their body remains. Surely they can be persuaded to pay for storing their intellectual remains together with high quality provenance information.)
Content type	It should be possible to preserve any digital object type and format, including types not yet invented.
Client autonomy	Readers and writers should have as much autonomy as possible.

---

45 Friedman, loc. cit. footnote 23, pp.93-126, *Uploading: Harnessing the Power of Communities.*

46 Gordon Bell and Jim Gemmell, *A Digital Life*, Scientific American 296(3), 58-65, 2007.

Catherine C. Marshall, *Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field* and *Part 2: Implications for Services, Applications, and Institutions*, D-Lib Magazine 14(3/4), 2008.

What will records custodians and repository institution administrators want of technology? In addition to well-known digital library technology, they will want support for:

Requirement name	Requirement statement
Repository autonomy	Any repository institution should be able to continue to use its currently deployed CM software without disruption originating in extensions for LDP. It should also be able to update or replace this software in future years, doing so without disturbing already preserved information.
Content sharing	Any repository institution should be able to share its content and metadata with clients and with other repositories without needing to make content adjustments requiring human judgment and intervention.
Workload sharing and scaling	Any repository institution should be able to share preservation effort with its clients to avoid burdens beyond its own resources.
Content durability	The repository community should be able to ensure that preserved information survives the demise of a large subset of all repositories.

In thinking about the challenges, we assume that the convenience of repository clients is unconditionally more important than that of repository managers, custodians, and curators.

Systematic Requirements

The requirements tabulated above focus on what might interest human individuals, paying little attention to engineering aspects. Comprehensive preservation also needs to address all sources of unreliability, all matters of scale, avoiding single points of failure, and integration with extant CM infrastructure. Many of these requirements pertain to digital library technology, have been thoroughly discussed years ago, and are satisfied by the best CM offerings. [PDI Chapter 9]

A possible exception is scaling. The number of digital objects being created is so large that, even if relatively few are to be preserved, their number will be too large for traditional curatorial attention. This would be true even if academic repository institutions were to receive funding they are seeking.[47] In addition, the number of data types is large[48] and growing because bureaucrats, businessmen, scholars, and engineers are creating new data formats for new applications.[49]

Core aspects of preservation schema need to be common world-wide, supporting document interchange among repositories and their clients, together with ready transfer of metadata from one repository to another.[50] LDP software must extend CM services and be deployable without disrupting ongoing service.

A Model for Digital Infrastructure

The Figure 3 digital repository model abstracts structure found in most deployed CM packages.[51] Critical archiving aspects are immediately apparent. For instance, almost all human interactions are via personal computers. Thus, data preparation for preservation can be with available content and metadata editing programs, with no more than modest additions to make records durable for the long term.

---

47 Fund raising seems to be the principal objective of the *Blue Ribbon Task Force on Sustainable Digital Preservation and Access.* See footnote 16.

48 Library of Congress, *Sustainability of Digital Formats: Planning for Library of Congress Collections,* available at http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml.

49 E. Durr, K. van der Meer, W. Luxemburg, and R. Dekker, *Dataset Preservation for the Long Term: Results of the DareLux Project,* Intl. J. Digital Curation 1(3), 29-43, 2008.

50 Open Archives Initiative, *Compound Information Objects: the OAI-ORE Perspective*, 2007, http://www.openarchives.org/ore/documents/CompoundObjects-200705.html. Also Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper, *Fedora: An Architecture for Complex Objects and their Relationships,* J. Digital Information 8(2), 2007.

51 Uwe M. Borghoff, *Vergleich bestehender Archivierungssysteme (Comparison of Existing Archiving Systems)*, Nestor Kompetenznetzwerk Langzeitarchivierung, 2005, page 10.

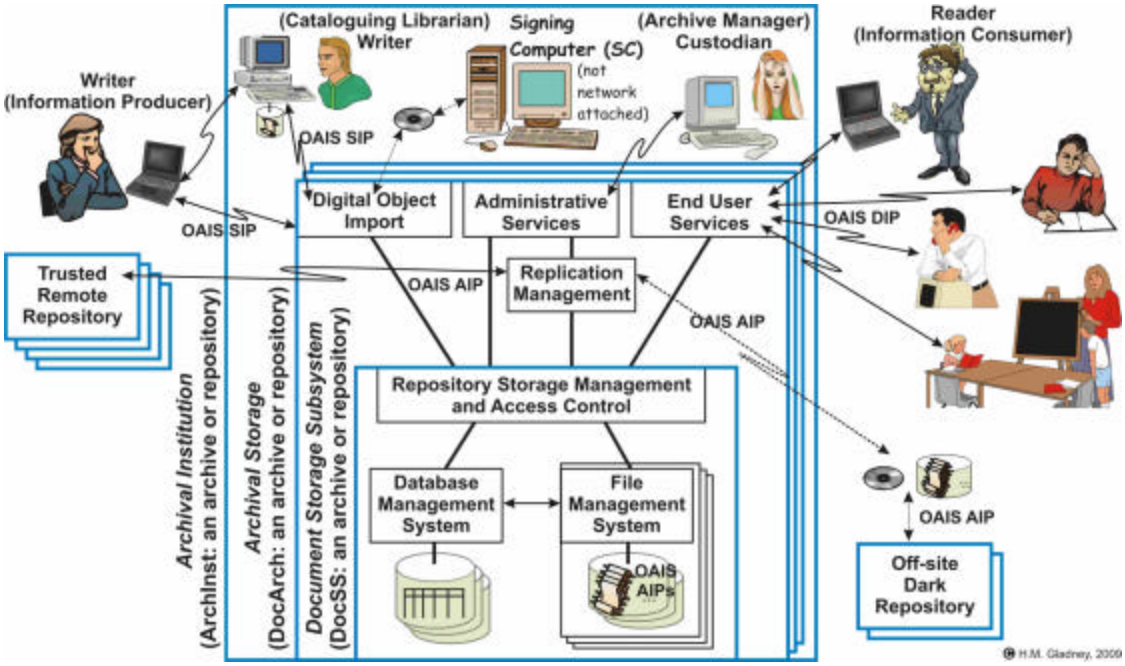

Figure 3: Model of repositories and related human roles[52]

A *sine qua non* of a digital repository is that, when a client requests access to information, the repository must be able to return a faithful replica of what was originally entrusted to it. This requirement is more stringent than the equivalent for libraries based on paper, partly because satisfying it is feasible and partly because altered messages have greater potential for harming their readers than do damaged books. A digital repository might provide many additional services. Some of these are so closely related to the stored content that it is hard to imagine a digital repository without them. Sometimes, however, it is advantageous to package library-related services in autonomous service offerings. Search services are an obvious example.

The figure shows nested repositories. (The word 'repository' is ambiguous.) An archival institution, **ArchInst**, addresses aspects pertinent to the current safety and ready availability of other people's records. Within **ArchInst**, one or more computing complexes, **DocArch**, provide data services to **ArchInst** employees and remote clients, usually acting as a server in client-server computing. A **DocArch** also presents interfaces to other repositories which might be organizationally autonomous and might also manage its own dark archives to forestall propagation of inappropriate content modifications.

Within **DocArch**, a nested **DocSS** integrates file subsystems and a database management subsystem to provide document storage subsystem services[8] that are becoming commodity software defined by a software interface standard.[53] Repositories will choose among storage subsystem offerings to accommodate different hardware, different storage capacities, and different data traffic. File and database subsystems might themselves be distributed among networked computers.

Readers might compare the Figure 3 model with the OAIS model,[54] which emphasizes different aspects. Specifically, OAIS emphasizes human administrative roles within an **ArchInst** and what it means to be a repository, defining vocabulary for discussing digital preservation.

Neither the archival storage layer, **DocArch**, nor the storage subsystem layer, **DocSS**, needs to include any code implementing LDP functionality. An exception is the **Digital Record Import** component, which should test whether submitted records conform to institutional criteria. As part of bit-string ingestion, this

---

52 Compare Figure 2 of Reagan W. Moore, *Building Preservation Environments with Data Grid Technology*, Am. Archivist 69(1), 139-158, 2006.

53 Java Community Process, *JSR 170: Content Repository for JavaTM technology API*, 2006, available at http://jcp.org/en/jsr/detail?id=170.

54 CCSDS, *Reference Model for an Open Archival Information System [OAIS]*, 2001, especially its Figure 4-1.

should register each submitted object's unique identifier, extract entity descriptors, and factor these data into its catalog, which should also be augmented to point at the safely stored bit-string replica.

We speculate that it will gradually become easy to assemble repository software from components provided by commercial and open-source suppliers to create functional replacements for offerings such as iRODS,[55] Fedora,[56] DSpace,[57] and Greenstone[58]—replacements that are easily tailored to institutional circumstances. When such component offerings conform to interface standards such as JSR 170[53] and XAM,[59] it will become easy to replace one component by another to exploit emerging technology, doing so without disrupting deployed repository services or stored information. In fact, some of the named projects have started to collaborate to share functionality.

It is unlikely that library catalog schema can be designed today to be fully satisfactory for periods of many decades. Relational database offerings have been designed to allow redesign without disrupting on-going service.[60] Improvements responding to shifting institutional interests, new data formats, and new ideas for information discovery are being explored in many projects.

Trusted Digital Repositories Methodology

> Collections are made more usable with tools that support gathering, organization, and transformation into new forms of scholarly output. …
>
> Inherently fragile digital objects are more likely to persist over time within a centralized and managed repository than in a distributed server environment in which … server and data management may vary.[61]

Most TDR articles address only near-term preservation. This might be because doing so is feasible with existing CM offerings, whereas adequate long-term software is not yet available. The issue at hand, however, is whether what they prescribe can achieve everything needed for long-term preservation.

> [A] fundamental issue associated with the authenticity of electronic records is the assurance that a trusted third party is responsible for [their] storage and for ensuring that they remain unaltered. The creators/users of electronic records [must] transfer them to the custody of a trusted third party where they … cannot be changed by anyone, including the individual or organizational component that initially created, maintained, and used them as operational records.[21]

This literature mostly pays attention to academic content, perhaps because of its authors' affiliations and perhaps because few businessmen, government officials, or citizens have expressed concern about vanishing digital creations. (An exception is the records of completed U.S. Presidencies.[62]) If challenged about limited preservation scope, a university librarian might reasonably respond, "Faculty work is at risk because so little of it receives curatorial care. We cannot be concerned with LDP when good work is being lost right now!" We find admonitions such as:

> Archival perspectives and planning need to be built into the creation and early management of all information that will enjoy long-term preservation. … Simply put, if creators of digital information do not take steps to preserve it early in its life, it will never reach any long-term preservation facility.[63]

---

55 San Diego Supercomputer Center information at https://www.irods.org/index.php/Documentation.

56 Cornell Digital Library Research Group information at http://www.fedora-commons.org/.

57 This software is used to promulgate MIT research articles. See http://dspace.mit.edu/.

58 University of Waikato software available at http://www.greenstone.org/.

59 See http://en.wikipedia.org/wiki/Xam and https://slx.sun.com/1179270715.

60 Don Chamberlin, *Using the New DB2: IBM's Object-Relational Database System*, Morgan Kaufman, 1996, ISBN1-55860-373-5.

James Hamilton, *Database Dialogue with Pat Selinger,* Comm. ACM 51(12), 32-35, 2008.

61 Katherine Skinner and Martin Halbert, *Strategies for Sustaining Digital Libraries*, Emory U.P., 2008, ISBN 0-977-29941-4, available on-line from http://metascholar.org/publications/StrategiesforSustainingDigitalLibraries.pdf. In this, see Leslie Johnson, *Principles … for Developing … Sustainable Repositories*.

62 Robert Pear and Scott Shane, *Bush Data Threatens to Overload Archives,* New York Times, 27th Dec. 2008, available at http://www.nytimes.com/2008/12/27/washington/27archives.html.

63 Anne J. Gilliland-Swetland, *The Archival Paradigm—the Genesis and Rationales of Archival Principles and Practices*, in *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*, CLIR Report 89, 2000, available at http://www.clir.org/pubs/reports/pub89/archival.html.

A high purpose for archives is to protect civil liberties by safeguarding records against chicanery by government officials and rogue governments' attempts to distort historical accounts. Neither repository employees nor repository owners (such as government agencies) can prudently be assumed to be fully trustworthy when such concerns arise decades after sensitive documents were purportedly created.

About Trust

TRAC formalizes the proposition that digital preservation can be based on trust for archival institutions.[64] Berman illustrates what many authors assume with: "One of the key components of [our] model is the formalization of the notion of trust between [collaborating repositories]".[65]

> The RLG/NARA trusted digital repository certification checklist defines a set of assessment criteria for preservation environments. The criteria can be mapped into data management policies that define how a digital preservation environment is operated. … By integrating a rule-based data management system …, we expect to demonstrate automated audits of the TDR checklist for a defined set of local policies. … [O]ne can also demonstrate the completeness and self-consistency of preservation environments. … all required preservation metadata are controlled by management policies.[66]

TRAC is asking readers who might depend on sensitive documents to trust assertions about "chain of custody". Articles asserting the importance of institutional relationships[67] might alert readers to the fragility of trust,[68] suggesting that TRAC has a weak foundation. Tibbo asks, "What is the purpose of [the TRAC] standard? … [C]ertification will not ensure digital longevity and authenticity, any more than best practices in analog repositories will ensure that no objects go missing or that none are defaced".[69]

A *trusted custodian* is a physical or juridical person entrusted with independently maintaining the records of electronic data interchange partners.[77] However, nowhere in TDR specifications is there a plausible explanation of who is expected to do the trusting, precisely which information qualities are to be made trustworthy, or which threats to authenticity and integrity will reliably be forestalled.[70] That concern is realistic is suggested by:

> Much data loss is due to human error; a very large number of attacks are carried out by insiders. Archives and libraries have often been targets in overt or covert wars. Consequently, there is every reason to expect that this will be the case with digital archives of key cultural materials. So the threats are real [for] the trans-institutional system but highly unpredictable for any given element in that system. Any model for sustainability and the associated costs must take such unpredictable considerations into account.[16]

---

64 RLG-NARA Digital Repository Certification Task Force, *Trustworthy Repositories Audit & Certification* (TRAC)*: Criteria and Checklist*, 2007, available from http://www.crl.edu/PDF/trac.pdf.

65 Fran Berman, Robert H. McDonald, Brian E. C. Schottlaender, and Ardys Kozbial, *The Need for Formalized Trust in Digital Repository Collaborative Infrastructure*, Proc. NSF/JISC Repository Workshop, 2007.

66 Reagan W. Moore and MacKenzie Smith, *Automated Validation of Trusted Digital Repository Assessment Criteria*, J. Digital Information 8(2), 2007, available at http://journals.tdl.org/jodi/article/view/198/181.

67 Glenn Dingwall, *Trusting Archivists: The Role of Archival Ethics Codes in Establishing Public Faith*, Am. Archivist 67(1), 11-30, 2004.

Ronald Jantz and Michael J. Giarlo, *Digital Preservation: Architecture and Technology for Trusted Digital Repositories,* D-Lib Magazine 11(6), 2005.

Seamus Ross and Andrew McHugh, *The Role of Evidence in Establishing Trust in Repositories*, D-Lib Magazine 12(7/8), 2006. http://www.dlib.org/dlib/july06/ross/07ross.html

M. Seadle and E. Greifeneder, *In archiving we trust: Results from a workshop at Humboldt University in Berlin*, First Monday 13(1), 2008.

Michael Day, *Toward Distributed Infrastructures for Digital Preservation: The Roles of Collaboration and Trust*, Intl. J. Digital Curation 1(3), 15-28, 2008.

68 Dusko Pavlovic, *Dynamics, robustness and fragility of trust*, 2008, available at http://arxiv.org/abs/0808.0732.

69 H.R. Tibbo in a 15th Oct. 2007 posting to the MOIMS-Repository Audit and Certification blog (moims-rac@mailman.ccsds.org).

70 RLG &OCLC, *Trusted Digital Repositories: Attributes and Responsibilities*, 2002, available at http://www.rlg.org/en/pdfs/repositories.pdf. Neither this report nor that cited in footnote 64 contains phrases such as "trusted by" or "trusted to" except for the very general passage, "They are trusted to store these valuable materials. They are trusted to provide access to them in order to … to foster the growth of knowledge. They are trusted to preserve these items to the best of their ability for future generations".

To make TRAC comprehensively effective, its proponents must provide plausible justification. How will they protect holdings from the kinds of attack suggested by Gow?[71] How can a repository manager ensure that no employee alters holdings, or that his repository has worked correctly for many decades before he assumed responsibility? How can he persuade a skeptic that every TRAC prescription has been faithfully and competently executed for a sensitive 100-year-old record? How can a client unversed in repository procedures assure himself that holdings he depends on are authentic and undamaged? Can clients afford the time required for thorough trust evaluations? Are they likely to do so?

The core problem is that TRAC protocols ask readers to trust archival custodians instead of providing what readers will really want—to be able to trust preserved records. TRAC rules might be adequate for institutions managing their own output, especially if these holdings are unlikely targets for inappropriate modification and unlikely to harm readers if they are modified. However, such trust cannot prudently be extrapolated to other circumstances without careful analysis that nobody has provided.

A TDR Example

For specificity about the TDR approach, we analyze DICE group work,[3] partly because its authors have published extensively and made available design detail[72] and partly because this work influences the giant U.S. NARA Electronic Records Archives project[73] and other repository projects.[66] It has readily identified similarities with independent projects.[74] Describing it, Watry asserts:

> Although the NARA prototype represents the current state of the art in digital preservation technologies, it also points to the necessity of a new generation of technologies, derived through research advances, which will fulfill its preservation goals, including:
>
> - *Authenticity,* the assertion that provenance descriptive metadata and integrity metadata remain inextricably linked to the electronic records … and [un]altered;
> - *Integrity*, the assertion that the electronic records have not been corrupted, …;
> - *Infrastructure independence*, the assertion that preservation [survives arbitrary infrastructure change].[4]

The DICE iRODS prototype fortifies preserved content to provide these qualities indirectly, not taking into account that each quality can be achieved more directly. Describing this work, Moore asserts that:

> [We] need to send into the future not only the information (records), but also a description of the environment that is being used to manage and read those records.[3]
>
> The challenge is designing a data management system that is able to support evolution of all of its constituent parts, from the storage system, to the database technology, to the authentication mechanisms, to the access mechanisms. Data grid technology … provides these capabilities …[52]

Description of iRODS[3] highlights a "preservation environment"—a co-managed digital repository network. It deals with maintaining data integrity in the face of unreliability of repository implementations, describing what can be done between Figure 1 input messages **4** and output messages **5**, but says little about composition of repository deliveries (OAIS DIPs), about intelligibility of preserved information when current workstation technology is obsolete, about trustworthiness in the face of felonious modification of archived records, or about scaling beyond what an institutional repository can handle.

> One can think of a preservation environment as the set of software that protects records from changes that occur in hardware systems, software systems, and even presentation mechanisms. Preservation environments insulate records from changes that occur in the external world … [with] mechanisms … to parse and present digital data, even after the original creation application has become obsolete. A preservation environment should be able to use modern technology to access and display "old" records.[66]

iRODS descriptions conflate the *writer* and *custodian* roles, ignoring the difference between *knowledge* and *information* about records being handled. This distinction becomes more critical as records age. Whereas today's *Archivist* can converse with today's *Records Administrator*, this will usually be

---

71 Brian Gow, Keith Epstein, and Chi-Chu Tschang, *The New E-spionage Threat*, Business Week, 33-41, April 21, 2008. Available at http://www.businessweek.com/magazine/content/08_16/b4080032218430.htm.

72 Private correspondence and debate during December 2008.

73 Kenneth Thibodeau, *If you build it, will it fly? Criteria for success in a digital repository*, J. Digital Info. 8(2), 2007.

74 Michael Factor et al., *Preservation DataStores: New storage paradigm for preservation environments*, IBM J. Res. & Dev. 52(4/5), 2008.

impossible for his successors. Durable repository services are necessarily limited to what can be accomplished solely on the basis of recorded information.

What Watry[4] and Moore[3] call "digital preservation theory" is what others usually call software architecture.

Trust and Authenticity

Although trust might not be a critical concern for scholarly and cultural works, it is for many legal documents, financial records, and medical patient information. In view of massive Internet chicanery, TDRs as specified today seem inadequate for sensitive information.

> The integrity of the evidential value of materials is ensured by demonstrating an unbroken chain of custody, precisely documenting the aggregation of archival materials as received from their creator …, and tracking all preservation activities associated with the materials.[63]

Chain of custody assertions might be good evidence for information on paper, but are inadequate for digital documents. Even for paper records, that a collection holds a document is, in itself, insufficient to be evidence in litigation. If challenged, it must be accompanied by testimony plausibly evidencing a standard of care that makes improper modification improbable. Whether the object is in paper or digital form, for an eventual reader to decide whether or not to trust information received from a repository network, he would need to know a great deal about that network and be skilled at evaluating that knowledge. It is unlikely that most readers will have the necessary expertise or patience.

The TDO approach shifts the locus of trust from the quality of repository management to the management of a relatively small set of public cryptographic keys. Given readily provided tools and instructions, any reader will be able to judge whether information has been distorted, doing so quickly without any immediate human help. Asymmetric cryptography is a better tool for authenticity management of publicly accessible digital records than any alternative described in archiving literature, with the possible exception of the internal procedures of a few superbly managed central government archives. Applied cryptography continues to be an active research topic. Although cryptography is not thought foolproof, it is much better understood than trust in institutions.

Bit-String Integrity and Scaling

Rosenthal asserts that petabyte-scale storage systems fail to store bits sufficiently reliably.

> The case for bit preservation not being solved … rests on the many orders of magnitude mismatch between the reliability requirements implied by society's expectations of the amount of data to be preserved and the length of time for which it should be preserved, and the observed performance of current storage hardware and software.[75]

This concern about the entire content of a petabyte store is misleading. Eventual readers will care about the integrity of a few preserved objects they retrieve, not about the integrity of entire repository contents. With this shift of attention, mitigating Rosenthal's concern seems straightforward. We can achieve Internet scales and economy by shifting software tool efforts from helping relatively few repository custodians to helping the much larger number of writers and readers, enabling their doing critical integrity testing on their own personal computers rather than on their trusting CM server machine reliability.

More generally, we need to consider exploiting the fact that modern equipment and software make bit-string errors relatively rare. For instance, we could reduce our integrity expectations from perfection to some modest failure rate, calling for bit-string errors in less than, for instance, 0.001% of objects retrieved. For the minority subset of cases for which this might be unacceptable, more rigorous integrity checking and error correction techniques could be provided.[76]

Infrastructure Independence

> A preservation environment is the middleware that shields records from the rapid evolution of technology.[3]

Digital objects can easily be represented without any dependency on repository infrastructure, thereby avoiding the need to describe middleware except by requiring that, when a stored message is requested

---

[75] David S. H. Rosenthal, *Bit Preservation: A Solved Problem?* Proc. iPRES2008, September 2008.

[76] Aiden A. Bruen and Mario A. Forcinito, *Cryptography, Information Theory and Error-correction; a Handbook for the 21st Century,* Wiley Interscience, 2005, ISBN 0-471-65317-9.

by sending its identifier to an appropriate repository, this repository nearly always returns the bit-string identified, together with its reliably-bound metadata.

> When a trusted custodian acquires digital records destined for continued preservation, … [he] can issue a data grid command to register the records of the creator from the original recordkeeping system into the data grid registry. … The records metadata can be extracted from the recordkeeping system if they are inextricably linked to the record …
>
> The trusted custodian may maintain the original records' names and organizational structure, … appraise the records and define the subset that will be archived and the standard encoding format that will be used. … and the trusted custodian may register the material … [F]or the preservation of Web sites, the records need to be extracted from the Web site, relinked, and organized before being registered … [A]uthenticity must be addressed. If a Web page is relinked so it will point to other Web pages within the preservation environment instead of the original site, the process used to do the relinking should be documented.[3]

In this iRODS process, an archivist first constructs relationships among records and metadata. Later, whenever a then-current preservation environment needs to be renewed, some archivist must migrate the representation of these relationships. I.e., an archivist #1 **first** brings content and metadata into his preservation environment, and **then** constructs expressions of relationships. Later, archivists (#2, #3, ..., who will often be different individuals) have the challenge of migrating these relationships to new preservation environments whenever the old environments approach obsolescence. The catalog and its procedures appear to include single points of failure to the entire preservation environment.

If archivist #1 were to express such relationships **before** bringing records into his repository, doing so in a way that makes no reference to potentially ephemeral aspects of repository design/implementation, the described efforts of archivists #2, #3, ..., would become unnecessary.

The iRODs mechanism will almost surely be adequate and cost-effective for e-mail collections[62] because e-mails include sufficient metadata. They also have simple formats likely to be intelligible for centuries, so that transformative migration will not be needed. Furthermore, as is the case for many other bureaucratic documents, an e-mail text has evidential priority over what might have been on its author's mind—a property not shared by typical works of scholarly or artistic innovation. For similar reasons, archivists and other agents who have little contact with original writers can safely preserve a great deal of WWW content.

For many kinds of bureaucratic record, the distinction between *writer* and *custodian* roles is unimportant. For many other kinds of information, archivists cannot preserve properly without writers' help.

Managing Information Quality

> [W]e express the management policies as sets of rules that control the execution of each micro-service. We evaluate the assessment criteria through queries on persistent state information that is generated by application of the rules. The rules are stored in a rule engine that is installed at each remote storage location, and the persistent state information is stored in a central metadata catalog.[3]

The DICE team has created about 100 tools to ensure that stored information conforms to preservation policies—"micro-services" that they store at each of their repository locations. Some of thise micro-services are distinctly for repository management, such as tools for managing client access to collections and discarding superannuated holdings. Others are clearly targeted at content while it is being used in a workstation. Examples are tools for creating message authentication codes and extracting metadata.

We need to consider the iRODS toolset for porting into TDO handlers.

Trustworthy Digital Object Methodology

TDO architecture is a scheme whereby a writer can bundle any collection, together with whatever metadata and contextual information he thinks readers might value. (Figure 1)

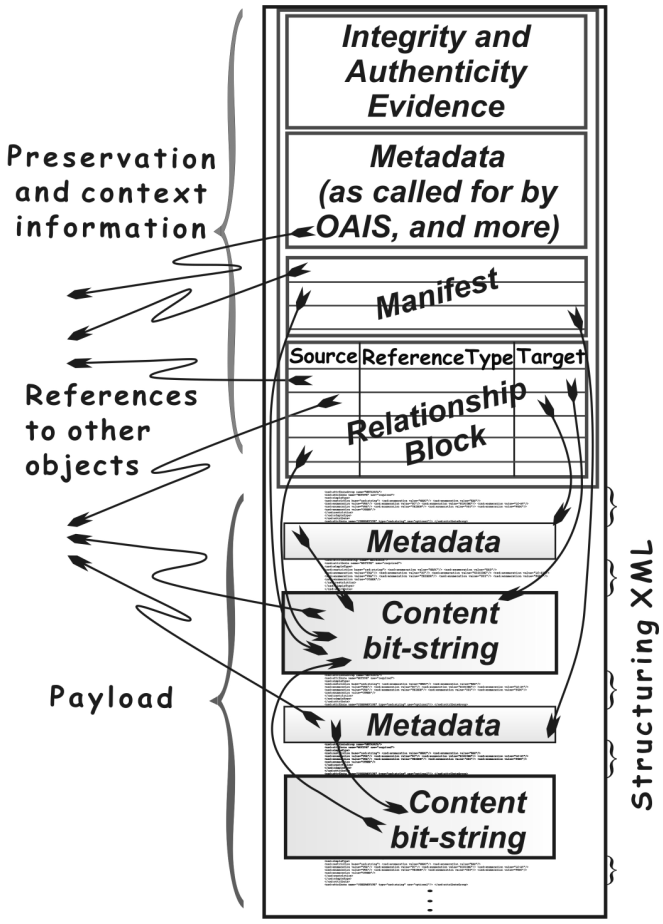

Figure 4**:** Trustworthy Digital Object (TDO) structure

TDO content representing a collection or a single work is necessarily chosen by an author, editor, or curator. This writer should include metadata and signatures identifying himself. [PDI §11.2] From an OAIS perspective,[54] a TDO is a Submission Information Package (SIP, as suggested in Figure 1 and Figure 3) and also an Archival Information Package (AIP). Catalog records and other management data a repository might want can be generated as a side effect of ingesting TDOs that writers have provided. Archivists could then use repository tools to supplement this information, possibly acting as writers themselves, sharing their additions in new TDOs bound to the ingested information. Such additions would be new intellectual contributions.

Trustworthy Digital Object (TDO) Architecture

> The authenticity of *electronic records* must be verifiable from elements of the records (i.e., either on their face or linked to them) and contextual to the records (i.e., belonging to their documentary, administrative or technological context), while the authenticity of *electronic copies* of authentic electronic records is attested by the preserver ... In other words, *any electronic copy of an authentic electronic record is authentic if declared to be so by an officer entrusted with such function*, namely the official preserver.[77]

Each TDO consists of a payload, metadata that other authors have called for,[78] structure information, references to critical context, and trustworthiness evidence. [PDI §11.4] The bundling conforms to ISO

[77] Heather MacNeil, *Providing Grounds for Trust: Developing Conceptual Requirements for the Long-Term Preservation of Authentic Electronic Records*, Archivaria 50, 52-78, 2000. Also *Providing Grounds for Trust II: The Findings of the Authenticity Task Force of InterPARES,* Archivaria 54, 24-58, Fall 2002.

[78] Rebecca S. Guenther, *Battle of the Buzzwords: Flexibility vs. Interoperability when Implementing PREMIS in METS*, D-Lib Magazine 14(7/8), July 2008.

Unicode and UTF-8 standards,[79] uses only a small subset of standard XML,[80] and conforms to OAI-ORE conventions for sharing information among archives.[50] This will be described and preserved using the *Abstract Syntax Notation One* [ASN.1] standard.[81] In a Figure 4 TDO:

- The bit-string set that represents a work is XML-packaged with registered schema.
- The package includes metadata and one or more identifiers of the object itself [PDI §7.3]. Enough has been written about metadata[11] that little needs to be added here. TDOs must identify which metadata schemes they use because no consensus for a world-wide standard exists.
- Each file that represents part of the work is represented by one or more bit-strings that should be encoded in a computing-platform-independent representation [PDI Chapter 12]. We recommend that the writer includes a bit-string facsimile of his original file.
- Integrity is assured by message digests, and authenticity by writers' signatures. These are encoded with asymmetric key cryptography (a.k.a. public key cryptography).[82] Public keys are grounded by keys published by widely trusted institutions [PDI §11.2.1]. Public keys can be preserved in TDOs.

  The MacNeil quotation above suggests that curators should add their signatures.
- Links to contextual information are secured by included message digest codes of the linked entities.
- Any content blob can itself be a TDO, a feature useful for recording version histories. [PDI §11.1.3]

The *Relationship Block* represents a mathematical relation.[26] Each relationship is a triple: {Source, ReferenceType, Target}. Sources and Targets link to objects, into objects, or to extents within objects. [PDI §6.4] A ReferenceType can be an object link, but is more likely to be a relationship name.[50] Such relations can express any structure whatsoever, have been standardized as RDF,[83] and are easily loaded into databases for content catalogs and search indices, and easily built from such databases.

How TDOs Satisfy LPD Requirements

A writer can package his work with a TDO editor similar to existing text and graphics editors. To preserve such a TDO, he must submit it to repositories. TDO support in a repository consists of ingestion process plug-ins that extract information to create catalog and index entries.

Requirement name	How this requirement is satisfied
Content accessibility	Is provided by conventional CM services. Each repository should ensure that TDO bit-strings are preserved intact. Extensions such as LOCKSS[84] can be used to replicate TDOs among repositories for copy safety. Reliability can be enhanced by storing copies into autonomous repositories.

---

79 The *Unicode Standard* is described at http://www.unicode.org/standard/standard.html. For *UTF-8*, see http://www.utf-8.com/.

80 XML 1.0, XML Namespaces, XPath, and XPointer are the core needed. See IBM, *A Survey of XML Standards*, 2004, available at http://www-128.ibm.com/developerworks/xml/library/x-stand1.html.

81 B.S. Kaliski, Jr., *A Layman's Guide to a Subset of ASN.1, BER, and DER*, 1993, available at http://luca.ntop.org/Teaching/Appunti/asn1.html.

82 Donald Eastlake and Kitty Niles, *Secure XML: The New Syntax for Signatures and Encryption*, Addison Wesley, 2002, ISBN 0-201-75605-6.

Filip Boudrez, *Digital Signatures and Electronic Records*, Archival Science 7(2), 179-193, 2007.

83 *Resource Description Framework*, described at http://www.w3.org/RDF/.

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

A. Rodriguez, *Preserving the Last Copy: Building a Long-Term Digital Archive*, Computer Tech. Rev. 25(3), 17, 2004.

Requirement name	How this requirement is satisfied
Content intelligibility	For content bit-strings (Figure 4) for which EDP standards are insufficiently reliable, this can be provided by rendering programs written in virtual machine code [PDI §12.2] or perhaps with multivalent coding.[10] A pilot implementation has validated the UVC approach.[85] However, economy is still an issue, so that investigation for practicality is urgent.
Content authenticity[35]	Readers will be able to test content integrity by using the message digest included in each TDO's Integrity and Authenticity Evidence (Figure 4) and authenticity with signature and provenance metadata. [PDI §11.2.2]
Context reliability	Readers will be able to test that any TDO linked by an external reference (Figure 4) is the intended contextual object by comparing its message digest to the one the dependant TDO binds to the reference.
Method simplicity	The algorithms just alluded to can be hidden within TDO content extraction tools. However, critics of the current article will be unable to judge whether simplicity has been achieved until we provide a pilot TDO implementation.
Content preparation	This will be supported by a TDO editor similar to text and graphic editors.
Content archiving	Any CM offering with sufficient TDO data extraction plug-in code in its ingest module—sufficient to support finding and accessing TDOs—will be enabled.
Content type	For eventual intelligibility, every distinct content bit-string data type will require one or more rendering methods. [PDI §12.2] Most of these will need to be written by providers other than the author and his colleagues.
Client autonomy	In addition to TDO editors/viewers alluded to, end users will need access to conventional CM and search services. No other dependencies exist.
Repository autonomy	Except for ingestion plug-ins already mentioned, TDOs require no addition to repository software. No relationship among repositories is presupposed.
Content sharing	TDO schema conform to OAI-ORE schema.[50]
Workload sharing and scaling	TDO preparation, testing, and reading can be achieved by repository clients. This can be used to shift overwhelming burdens from repository employees.
Content durability	A TDO can be submitted to any number of autonomous repositories. Any repository can back up its holdings in other repositories.[84] Familiar search engines can be harnessed to locate TDO copies.

Capturing authors' intents is not essential, but rather a service to eventual readers. Traditional scholars invest high expertise and long hours to interpret important texts. For instance, consider U.S. Supreme Court efforts to ferret out constitutional authors' meanings, particularly by strict constructionists. Similar efforts are unaffordable for most content, and are made improbable by explosive information growth, so that the value of durable representation is much greater than it otherwise might be.

A TDO as described above does not protect against plagiarism. A malefactor can extract and repackage TDO content under his own signature. Plagiarism protection can, however, be achieved by a new TDO that encloses the at-risk TDO and that is signed by a public official, just as legal paperwork is commonly signed by a notary public.

Table 1 compares TDR to opinions about TDO.

[85] J. R. van der Hoeven, R. J. van Diessen, and K. van der Meer, *Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects*, J. Info. Sci. 31(3), 196-208, (2005).

J. van der Hoeven, B. Lohman, and R. Verdegem, *Emulation for Digital Preservation in Practice: The Results,* Intl. J. Dig. Pres. 2(2), 123-132, 2007.

Table 1: A comparison of TDR and TDO emphases

Text from *Towards a Theory of Digital Preservation*[3]	Comments from a TDO perspective
The representation information includes descriptions of the preservation management policies, the preservation processes, and the state information needed to verify the correct working behavior of the system. … rule-based data grids can verify that prior policies correctly enforced preservation properties, while [preserving] descriptions of the current preservation management policies.	How can an eventual reader use this to validate that preserved information that he depends on critically is in fact authentic? (Such a reader is unlikely to have convenient access to the help of an archive employee.)
Any claims about the current state of authenticity and integrity rely upon a complete description of prior actions. This is the second major challenge for preservation environments, the ability to characterize how preservation processes have been controlled.	If the bit-string representing an archived object is not altered after it has been prepared for preservation[86] and stored reliably, authenticity and integrity are durable properties.
When we examine the components that are required to build a theory of digital preservation, we note that representation information about the preservation environment itself is required.	Repository clients are interested only in the quality of information that they retrieve. To them, repository internal activities are irrelevant.
Records … can be migrated to the new storage system without the names of the records changing, without the access controls on the records changing, and without the links [with] metadata being broken. The data grid transparently manages all required administrative metadata … while new technology is [installed].	If records are stored as TDOs, all this is looked after by bit-string replication without further effort. LOCKSS functionality[84] is sufficient to propagate information between repositories.
The iRODS micro-services [and rules] can be thought of as defining the minimal set of preservation functions [and policies] that need to be carried forward in time … to enforce trustworthiness.	This addresses preserving repository mechanisms, in contrast to preserving records that are repositories' raison d'être.
The preservation environment needs persistent names for identifying the records, the archivists, and the storage repositories. Assertions about the management of the records can then be based on attributes associated with the persistent name spaces.	Such name spaces are conventionally called "data types". This TDR treatment looks like a revival of 1980's object-oriented programming that could be applied to TDOs.
To maintain the ability to interpret and display the records, the preservation environment must characterize its own evolution, and the impact that preservation environment evolution has on record management.	This assumes that server software, in contrast to workstation software, is needed to interpret and display content. Such service is neither needed nor easily contributes to end users' convenience.
The acid test of a preservation environment is whether it describes the entire preservation information context sufficiently … This requires migrating not only the records, but also the characterization of the preservation environment context.	A TDO, considered together with its context, is self-describing. Programs needed to manage and exploit TDOs can themselves be preserved as TDO content—contextual information.

Moore correctly and insightfully suggests that such differences are largely the consequence of different objectives: "Our primary goal is to provide real software that supports production systems managing petabytes of data and hundreds of millions of files. We organize the distributed data into shared collections, and then manage properties of the shared collection".[72] In contrast, the cited TDO work attempts to design methods that will protect even the most sensitive digital objects long into the future, but has deferred building practical software and design for optimization until we are confident that basic objectives and feasibilities are understood and have been sufficiently exposed to peer review. The differences are part of unavoidable tension between accommodating current social practices and designing what might be best for the foreseeable future.

Discussion

Even a content management subset is too complex and too important to be specified and implemented by any small group of investigators and engineers. For robust wide appeal, information schema and supporting software need to be structured to encourage contributions from anybody able and willing to help. The DICE publications attempt to provide a complete preservation environment.[4] We prefer an

[86] H.M. Gladney and R.A. Lorie, *Trustworthy Digital Objects: Durable Encoding for When It's Too Late to Ask*, ACM Trans. Office Information Systems 23(3), 299-324, 2005.

open approach, partitioning what is required into components for nearly autonomous participants—both software developers and also end users.

Content management literature exposes excellent opportunities for partitioning digital preservation for synergetic action: (1) distinguishing between near-term and long-term preservation; (2) focusing on human roles rather than on organizational or social affiliations; (3) enabling transparency for information qualities; and (4) recognizing that preserving digital repository software is merely a special case of preserving software, but is not required for preserving other content.

Near-term and long-term: TDR efforts target the question, "What can today's repository managers do to protect records entrusted to them?" TDO design addresses a different question: "How can digital records be preserved for a century and longer?" The distinction partitions development for autonomous teams.

Much of TDR infrastructure is digital library technology that has been thoroughly understood for about a decade. Digital content management depends on tools for search index and metadata extraction from documents, searching and collection catalog creation, security token management, format registries, data display and editing programs, and many other components. It can safely be assumed that such services will continue to exist and to be refined without special attention under a "digital preservation" slogan.

In fact, most CM innovation has come and will come from contributors who pay little attention to preservation. An optimal preservation strategy will avoid competition with this software development juggernaut while focusing on compatible measures. TDR technology certainly can benefit from extensions such as the micro-services defined by the DICE group. However, even with such additions, it will be insufficient for reliably assuring record authenticity and intelligibility for 50 years and longer.

In contrast, TDO methodology treats only aspects for ensuring that preserved information is durably intelligible and trustworthy, without disturbing mechanisms people might choose for other needs. For instance, it says nothing about most aspects of collection management, most aspects of librarianship, or most aspects of knowledge management. This restraint helps make it compatible with software for the avoided areas, as well as with the literature discussing the avoided topics.

Human roles: As emphasized by our figures, we recommend formulating digital preservation theory in terms of the roles of human participants. This encourages thinking about what somebody will need to know for each action, as illustrated by the distinctions between what typical records administrators and typical archivists know and do. It conditions solution architecture to exploit object-oriented programming, leading us to consider using the Ruby programming language[87] for TDO implementation. It prepares the way for teaching end users how to prepare and how to read preserved objects in language that they will understand.

Transparency: An unmatched TDO strength is that any reader will be able to judge trustworthiness of conforming objects. Such a reader need not worry that hidden archival processes might be flawed or might not have been faithfully and correctly executed over the decades since some record was created. Everything needed for trustworthiness evaluation is in the TDO or in objects to which it links recursively.

TDO preparation for preservation can be executed as close to information creation as is wanted to capture original authors' intentions accurately.

Preservation Challenges Yet To Be Addressed

This article's description of TDO theory has focused on the most challenging cases. It has said little about optimizations for easier cases. It has not explored compromises, such as accepting tiny bit-string error rates. Such topics need to be considered to make large-scale preservation affordable and convenient. However, our first order of business continues to be deciding whether we have a sound and complete theoretical foundation. When this is confirmed, other investigators might fill in what's missing.

No complete TDO implementation exists. Much of what will be needed can be harvested from available open source software. We believe everything TDO methodology calls for is not only practical, but will prove to be economical. Partly this belief is based on pilots of portions of what cited literature calls for[49,50,85] and more basic tools.[79,80,81,82] We are well into design for TDO editors, authenticity and integrity

[87] David Thomas and Andrew Hunt, *Programming Ruby: the Pragmatic Programmer's Guide,* Addison-Wesley, 2001, ISBN 0-201-71089-7.

testing tools, and TDO content viewers. For reasons suggested by Friedman,[88] we plan to assert copyright privileges using the Apache style of license.

The biggest outstanding technical challenge is practical and economical software for durably representing objects of each important file format. While the UVC-based method [PDI Chapter 12] is feasible for every kind of information, it has not been shown to be inexpensive. Multivalent representation has been proposed for text-like documents,[10] but not yet adequately tested for preservation; nor are its scope, limitations, and cost thoroughly understood.

Specialized circumstances have not yet been considered. For instance, nobody has described how to handle confidentiality and access control for long-term secure archives.[89]

Having tools for preservation will not be enough. Teaching people to use them correctly will probably be a challenge. Persuading their use in bureaucracies is sure to be a challenge.[90] It would be naïve to assume that having TDO software will make it easy to persuade bureaucratic change.

Challenges to Skeptics

> *The Preservation of the Integrity of Electronic Records* [project] goal was to identify and define conceptually the nature of an electronic record and the conditions necessary to ensure its integrity, meaning its reliability and authenticity, during its active and semi-active life. The research resulted in ... rules for developing and implementing a trustworthy electronic record-keeping system.[36]

Gracy echoes librarians' admonitions not to expect a "silver bullet" solution.[91] No engineer would talk about a "single solution", because the phrase has little sense for computing procedures. However, if what such commentators mean is a concise prescription for the technical portion of LDP, they have been mistaken. To anyone who believes that the solution sketched above leaves any difficult problem unanswered, we repeat an earlier challenge: identify specific shortfalls.

If what commentators mean by a solution is a software package for preserving every kind of data under all circumstances, they are correct. No such package exists, or is likely to be created, because different institutions have different needs and preferences and because new data formats are always likely.

If what commentators mean by a solution is a framework and toolkit from which archiving support can be assembled for any specific situation, we know how to satisfy them. Furthermore, such an assemblage can be extensible to accommodate new data types and new needs. What is needed now is what, in IBM Research, we used call "a simple matter of programming".

If the last is what conservators want, they need to say so, they need to vet our solution or a better one if they can find it, they need to find funding, and they need to work with software engineers to ensure that what is built is truly satisfactory. Archivists have expressed their readiness for such responsible action.

> [T]he participants issued a series of resolutions, calling for greater involvement by record keepers in information technology initiatives; ... collaborative approaches to records and information technology projects; ... and the continued development ... of standards for electronic records management.[92]

Conclusions

The essence of what is worth preserving is a pattern embodied in some artifact.[31] The artifactual instance that we tend to value most highly, often calling it "authoritative", is one created by the author, artist, or clerk himself—an artifact we call "the original". For safety, we can create copies of digital originals, and

---

88 Friedman, loc. cit. footnote 23, pp.100-105.

89 M. W. Storer, K. Greenan, and E. L. Miller, *Long-Term Threats to Secure Archives*, Proc. Second ACM Workshop on Storage Security and Survivability, 9–16, 2006.

90 Adrian Cunningham, *Digital Curation/Digital Archiving: A View from the National Archives of Australia,* Am. Archivist 71(2), 530-543, 2008.

91 Karen F. Gracy review of Ross Harvey's *Preserving Digital Materials,* Am. Archivist 69(2), Fall/Winter 2006.

92 Laura Millar, *Authenticity of Electronic Records*, Report for UNESCO and Intl. Council on Archives, 2004, available at http://www.ica.org/en/node/30209.

store these independently of one another. Any such copy is considered "true" if it captures the essential pattern elements; it can change accidental aspects without these changes being regarded as defects.[93]

Every work has valuable context, some of which is essential for interpretation or as provenance evidence. Such context can sometimes be reconstructed after long delay, as we see in the work of archeologists, paleographers and historians. However reconstruction is so expensive and requires such rare expertise that prudence suggests recording critical context early in a work's history and binding it into the artifactual carrier of the pattern. We intend TDO schema and rules to be a complete framework for accomplishing this in formally structured objects that are easily shared, interpreted, and durably stored.

Widespread long-term digital preservation will not occur until the cost of preserving is small compared to the cost of creating original objects. A good TDO implementation will help accomplish this. When we consider all the pertinent social factors, we see that it is not only economically prudent, but perhaps even a sort of moral obligation, to identify and root out every possible source of unreliability and confusion.

We invite open criticism of our arguments and claims. We challenge readers to solve the TDR shortfalls identified, to identify hidden assumptions threatening TDO completeness and logical foundations, and to invent new and better methodology.

Acknowledgements

This article is informed by correspondence with Reagan Moore and with Paul Watry. Partly this has been a debate exploring differences of opinion and of objectives—differences that help to motivate the distinction between near-term and long-term digital preservation. The article has also benefited from detailed discussions and criticisms of the draft text by John Bennett, Tom Gladney, Keith Hoilett, and John Swinden.

Glossary

CM	(acronym) *content management*, which is a superset of digital library services; alternatively, *content manager*. Which form is appropriate will be clear from the context.
content management	(noun phrase) 21st-century phrase for the complex of services required to preserve, protect, and make accessible information mostly created by other people than the custodians. Content management grew out of what in the 1990s was called *digital library* services.
context	(noun) the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood (Concise Oxford English Dictionary); in the case of a TDO, the bundled metadata together with all the information referenced in its Figure 4 *Relationship Block*, considered recursively; context is best thought of as being unbounded (whether or not it is truly infinite only rarely has practical importance).
DICE	(acronym) *Data-Intensive Computing Environments,* the name chosen for itself by an R&D group at the San Diego Supercomputer Center and at University of North Carolina at Chapel Hill.
digital curation	(noun phrase) management of digital objects over their entire lifecycle, ranging from pre-creation activities wherein systems are designed, and file formats and other data creation standards are established, through ongoing capture of evolving contextual information for digital assets housed in archival repositories;[5] sometimes used as a synonym for *digital archiving*.
digital library	(noun phrase) see *content management.*
information	(noun) the subset of *knowledge* that a human being can communicate to another human being by speaking, writing, or drawing.
knowledge	(noun) what a human being (or animal) can know about the world or universe, including that of which he might not be consciously aware (as Sigmund Freud taught) or be unable to communicate adequately in words alone (for instance, how to ride a bicycle).[29]
LDP	(acronym) *long-term digital preservation*, which is the complex of measures required for and/or undertaken to mitigate information unreliability caused by ravages of time, including human misfeasance, fading human memory, and technological obsolescence.

[93] For innovative works, it can be difficult to decide which aspects are essential and which accidental. [PDI §4.1] This decision is perhaps the most difficult part of preservation, and can be controversial. For records of business transactions, this challenge is much less acute.

message	(noun) *information* to be conveyed from some *writer* to some eventual *reader*; often a synonym for *digital object*.
near-term	(adj.) in discussions of archiving, describing measures undertaken to please information clients today and in the next five to ten years—a period short enough that service managers can ascertain client satisfaction and react with service improvements.
preservation environment	(noun phrase) a co-managed network of digital repositories used to archive information, in contrast to storing information for current operations.[66]
reader	(noun) a role depicted in Figure 1, being somebody who uses *information*. The word is used to suggest independence of the content, style, and purpose of the information under discussion. For instance, the information might be a computer program to be executed. The reader might not be a human being, but a machine process instead.
records administrator	(noun phrase) in a business or governmental bureaucracy, a human role, being somebody who knows the reasons and organizational rules for record keeping, and administers compliance procedures, perhaps informing colleagues about these matters.
role	(noun) a set of responsibilities and/or activities assigned to, resp. carried out by, some human agent; a person's or thing's function in a particular situation (Concise Oxford English Dictionary).
TDO	(acronym) *Trustworthy Digital Object*, a formal structure combining content, metadata, and cryptographic signature information; see Figure 4.
TDR	(acronym) *Trusted Digital Repositories*, a widely discussed approach to digital preservation.
UVC	(acronym) *Universal Virtual Computer* devised by Lorie.[57]
writer	(noun) a human role suggested by Figure 1, being someone that exploits *knowledge* to create or edit *information* for others. The word is used to suggest that, in the discussion at hand, details of information genre are unimportant.

