Economics and Engineering for Preserving Digital Content©

H.M. Gladney
HMG Consulting

Abstract: The keynote address of ECDL 2007 suggests that progress towards practical long-term digital preservation is stalled. The current article responds by sketching how a modest software development team could implement and deploy a previously described conceptual solution, Trustworthy Digital Object (TDO) methodology, for the technical component of digital preservation. It emphasizes scholarly and cultural digital content, but could be extended to discuss bureaucratic records.

Curators cannot afford unique technology, but must exploit marketplace offerings. Macro economic facts suggest shifting most preservation work from repository institutions to their users. Much of the software needed is available. It has, however, not yet been selected, adapted, integrated, or deployed for digital preservation.

Our earlier articles describe conceptual solutions for all known challenges of preserving a single object, but do not deal with software development or collection scaling. We describe a practical strategy for detailed design and implementation of software to automate the clerical component of digital preservation. Tools for daily work can embed packaging for preservation without much burdening their users. Document handling is complicated by human sensitivity to communication nuances. Our engineering section therefore suggests how project managers can master the many pertinent details.

Introduction

Long-term preservation of digitally represented information (abbreviated LDP below) has been discussed for over a decade by librarians, archivists, and information scientists, an informal group sometimes called the digital preservation community (DPC).[†] From the perspective of engineers accustomed to the pace of I/T research and development, progress reported in the DPC literature has been surprisingly slow.[1]

> [No] concise and well-developed strategy that represents the views of a broad community has yet emerged. Since 1989 at least twelve have been published. Ross[2]

International conferences and workshops occur in rapid succession. However, these discuss mostly LDP urgency and social processes for collaboration, without introducing novel concepts or addressing engineering specifics.[3] A recent notice is typical:

[♥] An early draft version was made available on the ERPA Eprints server in December 2007.

[†] All cited Web pages were viewed in November 2007 or more recently.

[1] The digital preservation community (DPC) alluded to this article is an informal community whose membership can be understood from the citations, considered recursively, of periodicals such as DPC/PADI's *What's New in Digital Preservation* and *Digital Document Quarterly*.

[2] S. Ross, *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, 11th European Conference on Digital Libraries (ECDL), Budapest, September 2007.

This reflects a European perspective. For a critical review of the U.S. National Digital Information Infrastructure Program (NDIIPP), see H.M. Gladney, *Digital Preservation in a National Context: Questions and Views of an NDIIPP Outsider*, D-Lib Magazine 13(1/2), January 2007.

[3] Readers can judge the opinions expressed by inspecting conference proceedings, such as that available at http://www.kb.nl/hrd/congressen/toolstrends/programme-en.html, and plans such as that of M. Bellinger et al., *OCLC's digital preservation program for the next generation library*, Advances in Librarianship 27, 25-48, 2004.

> Leading figures from the international science community will meet today to try and save the digital records of the world's scientific knowledge from being lost. Policy-makers from the European Commission and national governments will convene with world-renowned research organisations and digital preservation experts at a strategic conference in Brussels. They will discuss creation of an Alliance and European infrastructure for preserving and providing permanent access to digital scientific information currently stored in formats which are fast becoming obsolete and growing exponentially in volume. Jackson[4]

Seamus Ross's cited article reviews preservation state of the art, partitions his view of current challenges, and proposes that "as a community we need to re-think how we are approaching research … [and] need to engage … researchers in this process, and especially those with a strong computing science and engineering background." The current article responds to this invitation, sketching a software engineering approach to the technical component of LDP.

Synopsis of the Current Article

The current article describes practical steps towards realization of *Trustworthy Digital Object* (TDO) methodology, building forward from this published LDP architecture.[5] It emphasizes design of each individual digital object to have structure and properties which, considered together with linked TDOs, resist technological obsolescence and undetected improper change. It builds on earlier papers[6] by sketching proposed action by a software development team. The article is dense, depending on extensive literature. Readers who find its background summary difficult to absorb quickly are referred to its many citations

TDO methodology is designed to handle all the most difficult and general cases—tempting targets for malevolent modification, objects represented with unusual file formats, and computer programs whose behavior might be compromised by tiny changes. Its descriptions leave optimization for less sensitive cases to other authors.[7] The author's prior work shows that it can be added to digital content management offerings without disrupting their current services.

Compared to most LDP proposals, the TDO approach is unorthodox. For an unorthodox proposal to be credible and also useful to engineers, it must explain:

1) Why continued effort along conventional directions is unlikely to succeed;
2) Principles which are likely to enable a successful solution; and
3) Strategy to guide software development teams.

The first issue is dealt with in our ***Literature Analysis*** section. Ross represents DPC views comprehensively enough to use as a template, starting with his "key research challenges."[2] ***Economic Analysis*** reviews macro economics, suggesting why some LDP tactics will be impractical.

The second issue, conceptual design based on sound epistemology, is handled in ***Trustworthy Digital Objects*** (TDO, Figure 1) and in ***Logical Analysis***. Since other publications address such topics, these sections summarize with pointers to more detail.

[4] C. Jackson, DPC, broadcast e-mail on 15th November 2007. See http://www.alliancepermanentaccess.eu/.

H. Hockx-Yu, *Progress towards Addressing Digital Preservation Challenges*, Ariadne 53, Oct. 2007, is more about objectives and process than about progress.

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

[6] H.M. Gladney, *Trustworthy 100-Year Digital Objects: Evidence After Every Witness is Dead,* ACM Trans. Office Information Systems 22(3), 406-436, July 2004.

H.M. Gladney and R.A. Lorie, *Trustworthy 100-Year Digital Objects: Durable Encoding for When It's Too Late to Ask*, ACM Trans. Office Information Systems 23(3), 299-324, July 2005.

[7] Other authors propose methods for widely used file formats. See, for instance, Betsy A. Fanning, *Preserving the Data Explosion: Using PDF,* DPC Technology Watch Report, April 2008. Such treatments need to be re-examined scrupulously because they might mishandle the distinction between essential and accidental information, inadvertently misrepresenting authors' intentions.

Engineering Analysis suggests tactics for handling unavoidable complexity and integration based on mostly available software.

Scope Limitations of This Article

Digital preservation is too broad a topic for comprehensive treatment in an article intended to add to the state of the art, in contrast to reviewing it.[8] Many articles emphasize that effective LDP requires much more than a technical solution. Preservation challenges include issues of content selection, service delivery by archives,[9] social acceptance, technology uptake by end users, and professional education.[10] The current article is, however, limited to technical measures, leaving other aspects to other authors.[11] This choice is not intended to imply that technical issues are the only or even the most important digital preservation issues. However, those who have argued, or might argue, that solving the technical challenges should wait for progress on other fronts should consider the possibility that skilful technical solutions can reduce other problems significantly, sometimes even bypassing them by inducing changes in how the world works.

A professional archivist's criticism of a draft version of this article included, "While saving the bits may be easy; saving the meaning is anything but. Most of what is really hard the author ignores." This is correct, and deliberate, because conveying meaning is theoretically impossible, as taught by epistemologists. This fact is famously conveyed the final paragraph of Wittgenstein's *Tractatus Logico-Philosophicus,* "What we cannot speak about we must pass over in silence".[12]

Within the already limited scope suggested by the above paragraphs, the current article assumes without further discussion the availability of near-term digital content management (digital library) support. Such support would be required even for a world in which information accessibility and trustworthiness did not degrade with time. We take long-term digital preservation (LDP) to mean "mitigation of deleterious effects of decay, obsolescence, and human error or misbehavior that might impair the value of digital document copies long after originals were created." This choice partitions "archiving" into digital library services for the near term and compensation for degradation. The current article leaves to others questions of digital library infrastructure to please content management clients for the current decade.

Specific technologies mentioned are examples, not recommendations among alternatives. The article argues neither that TDO architecture is the only solution nor that it would be optimal for millions of users and billions of objects. Such topics have been dealt with elsewhere.[13]

The information to be protected is assumed to be collections of scholarly and cultural objects—text, pictures, audio/video recordings, computer programs, engineering designs and simulations,

[8] For a review, see Priscilla Caplan, *The Preservation of Digital Materials,* Library Technology Reports 44(2), February/March 2008.

[9] R.J. Cox et al., *Machines in the archives: Technology and the coming transformation of archival reference*, First Monday 12(11), November 2007, http://journals.uic.edu/fm/article/view/2029/1894.

[10] Bruce W. Dearstyne, *The archival enterprise : modern archival principies, practices, and management techniques*, 1993.

Bruce W. Dearstyne, *Riding the Lightning: Strategies for Electronic Records and Archives,* pp.139-159 in Bruce W. Dearstyne, *Effective approaches for managing electronic records and archives*, 2002.

Charles M. Dollar, *Authentic Electronic Records: Strategies for Long-Term Access* (Chicago: Cohasset Associates, 1999), 117-128.

[11] The relationship between technology and professional archivists' methodology is carefully discussed by Richard E. Barry, *Technology and Transformation of the Workplace,* pp.1-21 in Bruce W. Dearstyne, *Effective approaches for managing electronic records and archives*, 2002.

[12] Ludwig Wittgenstein, Tractatus Logico-Philosophicus (originally Logisch-Philosophische Abhandlung), 1922, §7. "Wovon man nicht sprechen kann, darüber muß man schweigen."

[13] H.M. Gladney, *Preserving Digital Information*, Springer Verlag, 2007.

…, and similarly structured information found in business and government offices. The current article does not treat preserving business transaction records.[14]

Terms of Reference

Key terms, including “archive”, “archiving”, and “digital preservation” seem to be used and understood differently by different people. A criticism of a draft version of the current paper, in which the critic seems to have ignored the narrow scope announced above, suggests that the consequent difficulties are even greater than I had previously supposed.

Perhaps this problem will be mitigated by re-emphasis of the near-term/long-term distinction and the following announcement. Within this paper at least, I will use “digital preservation” as a relatively broad term that alludes to the entire literature in which the term occurs.[15] I will use “LDP” for the narrower definition: “mitigation of deleterious effects of decay, obsolescence, and human error or misbehavior that might impair the value of digital document copies long after originals were created”. With this distinction, I see nearly all current digital content management technology as limited to services provided today or needed within the next few years.

The subject of this paper is exclusively LDP technology intended ensure that information stored in the near term continues to be fully accessible and useful in the long term. Authors sometimes use “100 years from now” as a synonym for “in the long term” because that is far enough into the future so that digital technology will have changes that we cannot confidently predict and so that no-one who might remember today’s details will be available for consultation.

Technical Objectives

> Digital preservation … is about maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its ‘interrelatedness’, and about securing information about the context of its creation and use. Ross[2]

Our objective is a future with deployed machinery for preserving any digital content whatsoever in ways that please its eventual users. Towards such a happy situation, the current article tries to sketch soundly based design for tools and infrastructure. For that, we choose an architecture capable of handling every data format[16] and sufficient for information that has high risk for fraudulent or accidental distortion.

This choice might be suboptimal. For some kinds of content and for some social circumstances, the general design might be unnecessarily complex and less convenient than alternatives. We handle such cases by architecture that permits simpler mechanisms to replace the general mechanism—by enabling optimizations as a later stage within the general design framework.

What requirements should an LDP solution address? What might one of our descendants want of information stored today? He would be satisfied if, for any document preserved in a world-wide repository network, he could:

1. Retrieve a copy of the bit-string that represents the content if he is authorized to do so;
2. Read or otherwise use the content as its producers intended, without adverse effects caused by mistakes and inappropriate changes made by third parties;
3. Decide whether the information received is sufficiently trustworthy for his application;

[14] This limitation is merely to avoid a too-lengthy exposition, which would be needed to describe how to accommodate the environmental and statistical differences between bureaucratic records and cultural information. These are described in Gladney, loc. cit. footnote 13, §9.4.

[15] Compare the definitions on page 7 of Priscilla Caplan, *The Preservation of Digital Materials*, Library Technology Reports 44(2), February/March 2008.

[16] An institutional archive must accommodate every kind of information generated by its constituency, and be readily extensible to future information representations. In particular, technical research generates computer programs among which some surely should be preserved.

4. Exploit embedded references (links) to reliably identify and retrieve contextual information and to validate the trustworthiness of contextual links, doing so recursively to as much depth as he feels he needs; and
5. Exercise all this functionality without hindrance by technical complexity that can be hidden.

In addition to professional authors, editors, and businessmen, some citizens will want to preserve information without asking anybody's permission to do so. They will want convenient tools and infrastructure to:

6. Package any content to be LDP-ready, doing so in some way that ensures that their descendants can use this content as specified immediately above;
7. Submit such readied content to repositories that promise to save it, possibly in return for a fee for archiving service.[17]

What technology will repository institutions want? In addition to perfect world digital library technology, they will want support for:

8. Continuing to use deployed content management software without disruption originating in extensions for LDP;
9. Sharing content and metadata without adjustments requiring human judgment;
10. Sharing LDP effort with their clients to avoid burdens beyond their own resources; and
11. Ensuring that preserved information survives the demise of a large subset of all repositories.

Information worth considering for LDP belongs to at least two classes. Bureaucratic records are created and managed by governmental and business enterprises to describe transactions, with each record constrained to conform to structural and usage patterns shared by many other records. Scholarly, artistic, and other intellectual records are mostly created by individuals without much a priori constraint of style or content and only later collected for retention and dissemination. How bureaucratic records are and should be managed is sufficiently different from how cultural information is and should be handled to merit their own LDP discussion.[18] The current article therefore should be read as treating only cultural and scholarly information.

Literature Analysis

Librarians and archivists have diligently explored how traditional repository methodology might be adapted for LDP. This topic interests information producers and consumers at most indirectly. Instead, what most stakeholders most care about is whether deposited content will be reliably delivered when requested, whether recipients will be able to use records as their creators intended, and whether they can be confident about content authenticity and integrity.

DPC literature seems repetitive, describing how various groups are exploring similar ideas, with little attention to know-how originating outside a small community.[19] It pays little attention to software developed for business use and identifies no technical challenges for which in-principle solutions are unknown. Its most prominent focus has become organization and management of archives, sometimes called "Trusted Digital Repositories" and more recently "Trustworthy

[17] 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 and pedigree evidence!

[18] See §9.4 of H.M. Gladney, *Preserving Digital Information*, Springer Verlag, 2007; ISBN 978-3-540-37886-0.

[19] DPC articles rarely cite from ACM or IEEE periodicals, or from the trade literature that is the best source for content management tools. The latter literatures seem unaware that anybody cares about digital preservation.

Repositories".[20] However, nobody has published a persuasive argument that repositories can solve what this literature calls "the challenge."[21]

Different communities have different notions of worthwhile research. If a computer scientist can describe how to satisfy a service requirement, he would say it is not a proper research topic. In contrast, the U.S. NDIIPP plan[2] reflects a common view that a research topic exists for any information management need unsupported by available software. In IBM Research corridors in the 1980s, the boundary between research and practical engineering was called "SMOP"—"a simple (or small) matter of programming." This did not necessarily mean that the task being discussed was either uncomplicated or inexpensive. Instead it meant that computer scientists knew answers to its difficult questions, allowing most of the work to be passed to software developers. Patent law wording is apt; one cannot obtain protection for an artifact or process design "obvious to someone versed in the state of the art."

Challenges Summarized

Ross[2] articulates a research agenda that "responds to the lack of progress … in the delivery of preservation solutions, methods and techniques over the past twenty years." Restated in abbreviated form, the challenges he identifies as needing attention are:

1. Restoration: when digital objects have broken, how can we ensure and verify the syntactic and semantic correctness of restored versions?
2. Conservation of intact digital objects: how can we ensure copy integrity and authenticity?
3. Collection and repository management: what archival methods and software design will satisfy collection users' quality expectations?
4. Risk management: how can preservation practitioners quantify content risks and benefits to choose how to spend limited resources?
5. Preserving digital object interpretability: how can archivists save any digital object to be intelligible or fully functional (if it is a program) in centuries to come?
6. Collection cohesion and interoperability: how can archives integrate collections contextually for sharing across administratively independent repositories?
7. Preservation automation: given the immense number of digital objects, what automation opportunities exist and how can these be realized practically?
8. Preserving digital object context: what are the legal and social expectations for contextual information and how can these be realized reliably and economically?
9. Storage technologies and methodology: how can prior experience be exploited to implement and deploy a practical network of archival repositories?

The skills needed vary by challenge. Is the challenge technological or a matter of social acceptance, such as that required for data interchange? Is it that no-one knows a solution, or

[20] Center for Research Libraries, *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, 2007*, http://www.crl.edu/PDF/trac.pdf.

[21] H.R. Tibbo in a 15th Oct. 2007 posting to the *MOIMS-Repository Audit and Certification* blog (moims-rac@mailman.ccsds.org), writes, "What is the purpose of [the TRAC] standard? … Even the highest level of certification 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 in some way.

N. Beagrie, *E-Infrastructure for Research: Final Report from the OSI Preservation and Curation Working Group*, Jan. 2007. §3, its "theoretically ideal situation (10 year time horizon)", includes:

> "Long-term threats to preservation and curation of digital information arising from organisational and administrative disruption, funding instability, or lack of clarity surrounding handover of curatorial responsibility will have been addressed. This will have been achieved through development of a network of repositories and services, replication and collaboration between them, longer-term funding frameworks, and definition of different types of repository, roles, and responsibilities over the lifecycle of research information.
>
> "We will have a complex network of trusted digital repositories and policies in place across sectors and disciplines."

merely that the solution is not familiar to DPC members? Is it that basic principles are unknown, or merely that known principles have not been reduced to practice? If the latter, is it that existing software has not been packaged to be convenient for archivists? To what extent is the challenge one of community education?

For each challenge, consider an engineer's opinion as a conjecture for critical consideration.

1. Restoration: restoration work is not urgent, except for the rare cases when prompt action can involve the original author. Readers interested in at-risk content can foot R&D expenses when they decide they want access.
2. Conservation of intact digital objects: how to do this is known, but faces SMOP (see above in *Literature Analysis*). Open questions include whether cheaper methods exist, whether and when a file-type-specific method is preferable to the generic solution, and how to persuade stakeholders to use a solution.
3. Collection and repository management: over 100 digital library packages exist. Each repository institution can choose one and seek tailoring for its preferences. LDP will require at most modest adaptation.
4. Risk management: crude risk estimates will be good enough. What is worth saving is highly subjective opinion inappropriate for research publication.
5. Preserving digital object interpretability: a generic method is known.[22]
6. Collection cohesion: context is subjective choice not amenable to automation—a topic for authors and subject experts. Being required for today's services, this need not be considered an LDP topic.
7. Preservation automation: methodology below treats only how to manage each individual collection. Handling tomorrow's scales is a task worthy of the best engineering skills.
8. Preserving digital object context: existing efforts to standardize technical and provenance metadata are on the right track.[23] How to reliably package unique context with any object is known, as is how to reliably link widely shared context.
9. Deploying storage technologies: how engineers can exploit available software offerings is suggested below.

These statements are styled for management. Software engineering demands a different formulation, worded to enable objective assessment of purported solution quality.

Archival Science

> Archival practice and science has responded well to the changing environment of information production and use. Its core principles of authenticity, trust, context, provenance, description and arrangement, and repository design and management evolved during this period. Ross[2]

Bankers, businessmen, and attorneys have long understood the importance of integrity, authenticity, and evidentiary audit trails—objectives that Ross identifies as "principles" of diplomatics[24] and archival science. Such objectives are not questioned, if they ever were.

[22] R.A. Lorie, *The UVC: a Method for Preserving Digital Documents. Proof of concept*, IBM/KB Long-Term Preservation Study Report Series #4, 2002. http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf and http://www.erpanet.org/assessments/show.php?id=1093856769&t=2.

[23] Northwestern Univ. Digital Library Committee, *Inventory of Metadata Standards and Practices,* at http://staffweb.library.northwestern.edu/dl/metadata/standardsinventory/.

[24] L.E. Boyle in J.M. Powell, *Medieval Studies: An Introduction*, Syracuse U.P., 1976, pp.69-101, sums up the key diplomatics questions as quis?, quid?, quomodo?, quibus auxiliis?, cur?, ubi? quando? (Who? What? How? What's related? Why? Where? When?)

Alerted by widespread Internet chicanery, a wary public is becoming sensitized to issues of content reliability.[25] The issues are practical mechanism and persuading its widespread usage.

The shift from paper to digital media and digital collection scales[26] make it doubtful that procedures designed for and by archives many decades ago will work well in the 21st century. Some patently collapse. An example is reader surveillance by uniformed custodians, as the British National Archives uses to protect 18th-century paper. Some will prove unaffordable. An example is having librarians create metadata for every interesting holding.

Other traditions do have good digital analogs. Nobody seems to have systematically sought all that can be adapted. For instance, "a well-documented chain of custody"[2] within an institution would be insufficient, because information-tampering during network delivery to and from repositories would be easy.[27] However such documentation can be embedded in any TDO.[28]

Economic Analysis

> [T]here has been relatively little discussion of how we can ensure that digital preservation activities survive beyond the current availability of soft-money funding; or the transition from a project's first-generation management to the second; or even how they might be supplied with sufficient resources to get underway at all.
> Lavoie[29]

> [P]reservation of digital materials [is] a labour-intensive artisan or craft activity. … there is widespread agreement that the handicraft approach will not scale to support the longevity of digital content in diverse and large digital libraries.
> Ross[2]

The Information Revolution is relatively new, beginning about 50 years ago and still evolving rapidly. In contrast, today's infrastructure for managing content on paper is about two centuries old. Widespread library access began about 130 years ago with Carnegie funding. Digital library software packages first appeared about 20 years ago, and librarians began to pay attention to LDP even more recently.

Our grandparents had little energy, resources, or inclination for deep reading, theater attendance, or other cultural activities. University education was a rare privilege. Today we have time and opportunity to enjoy what our previous generations missed. And our children take digital content, anytime and anywhere, for granted.[30]

A century ago, recorded data held in governmental institutions and in the private sector were several orders of magnitude smaller than today. Written information to manage individuals' personal health and welfare hardly existed. Audio-visual recording was little more than a laboratory curiosity. Scientific records were notebook scribbles. Creating and disseminating written works was slow and labor-intensive. Such factors are qualitatively changed today.

The research community has grown 100-fold since 1930, when physics conferences typically had about 50 participants. Today, American Physical Society conferences have about 5000 participants. Large increase in faculties, coupled with the publish-or-perish syndrome, has created a periodical subscription crisis for university libraries. Our reading is burdened by many scholarly articles that convey little new. Similar difficulties are evident in the popular press and

[25] L. Graham and P.T. Metaxas, *"Of Course It's True; I Saw It on the Internet" Critical Thinking in the Internet Era*, Comm. ACM 46(5), 70-75, May 2003.

[26] J.F. Gantz et al., *The Expanding Universe: A Forecast of Worldwide Information Growth Through 2010*, IDC White Paper, 2007, http://www.emc.com/about/destination/digital_universe/.

[27] Tools for "Internet exploits" are widely publicized. For instance, see *SANS Top-20 Internet Security Attack Targets*, 2006, https://www2.sans.org/top20/.

[28] Gladney, loc. cit. footnote 13, §11.13.

[29] Brian F. Lavoie, *The Fifth Blackbird,* D-Lib Magazine 14(3/4), March/April 2008. http://www.dlib.org/dlib/march08/lavoie/03lavoie.html

[30] Schwartz, Evan I., *The Internet Time Lage: Anticipating the Long-Term Consequences of the Information Revolution*, A Report of the Tenth Annual Aspen Institute Roundtable on Information Technology, 2002.

business trade press, with several dozen subscription invitations appearing in some home mailboxes every month. Many magazines are free, but a drain on productive time for anyone who looks at them.

Deploying information tools regarded with confidence by a dependent public and an interested professional community is paced by consensus processes that cannot be much hurried. It is hardly surprising that the infrastructure for digital content management is immature compared to that for content on paper.

Stakeholder Communities

> Each individual seeks to build [his] own … archives about [his] own family and heritage. … we value highly those linguistic scraps of personal documentation which have come down to us from our ancestors—a grandparent's diary, the name scribbled on the back of a photograph, the entries in parish registers and gravestone inscriptions—all of which provide evidence of our own pedigree. Crystal[31]
>
> Maybe we need to empower the individual, or, even, to understand that individuals will come to assume more and more responsibility for preserving our digital heritage—rather than records professionals' constant search for the magic solution for all systems in all institutional and individual applications. [I]nspiration comes from Leonardo Da Vinci's … personal recordkeeping: "He was an endless doodler … who tucked several notebooks of varying sizes into his waist belt to record his thoughts…" Shneiderman[32]

The professional community ready to invest in preserving digital content numbers between 500 and 5000 archivists and academics worldwide. A few governmental institutions have begun to invest in content management infrastructure, but almost no private sector enterprises are displaying interest, much less investing manpower.[33] Exceptions might be found in the pharmaceutical and entertainment industries because their key assets include electronically recorded information holdings. Nevertheless, reports of private-sector LDP investment are hard to find.[34]

The information technology community providing tools with which businesses, academics, and private individuals create and share digital content is much larger. Computer programmers probably number between 200,000 and 2,000,000. Many work on making it easier for "ordinary people" to create, edit, save, and share digital documents. They necessarily focus on what will appeal to their customers, especially by dramatically lowering digital content management cost.

How many people add to the information pool suitable for LDP? Perhaps between 10,000,000 and 100,000,000.[26] If technology to record personal activities[35] appeals to the public, this number will increase ten-fold.

Our statistics are admittedly crude. Any might be incorrect by a factor of three. However such uncertainty hardly affects what the numbers imply—that no plausible increase in repository institution resources will eliminate the apparent challenges. If research libraries and archives want to contribute to digital progress, they will need to make radical methodological changes.

For instance, the DPC is too small to keep up with software engineers' activities and ordinary citizens' software uptake.[36] The pace of information creation greatly exceeds repository

[31] In Cox, loc. cit. From D. Crystal, *Language death*. Cambridge U.P., 2000.

[32] In Cox, loc. cit. From B. Shneiderman, *Leonardo's laptop: Human needs and the new computing technologies*, MIT Press, 2002.

[33] Business priorities are elsewhere: avoiding disclosure of customers' private information and document retention for audit as required by Sarbanes-Oxley legislation. See http://www.soxlaw.com/.

E. Thornton, *Perform or Perish,* BusinessWeek, 38-45, 5th Nov. 2007, typifies financial press articles.

[34] An exception is the BBC PrestoSpace project (http://prestospace-sam.ssl.co.uk/) R. Wright, *Digital preservation of audio, video and film*, VINE 34(2), 71, 2004.

[35] G. Bell and J. Gemmell, *A Digital Life,* Scientific American 296(3), 58-65, March 2007.

[36] L. Feigenbaum et al., *The Semantic Web in Action,* Scientific American 297(6), 90-97, 2007, describes a recent addition.

institutions' ability to select for LDP. This suggests that no attempt professional collectors make to archive a significant fraction of newly created digital content can succeed. Archivists should consider strategies for selection only after sufficient time has elapsed so that well-informed communities can choose what is most valuable.[37]

Will delaying preservation for 20 years or longer from information creation risk losing much that is valuable? Of course it will, but the risks of the alternative are greater. It is unlikely that society will manage to save everything. Even prompt LDP selection will lose much that is valuable. It will confound what is saved with content later considered uninteresting. Finding good material might become more difficult than it already is.

Neither the private sector nor ordinary citizens have shown enough LDP interest to act in what some curators might believe to be in the citizens' own interests. Software improvements for day-to-day work apparently have higher priority, as does satisfying regulatory requirements.[33]

Cultural repository institutions have never created the software they depend on, but always depended mostly on what was developed by private sector enterprises. Today commercial software is complemented by excellent open-source software that is free to acquire, but not necessarily inexpensive to adapt and maintain. However, the DPC community does not seem to have adopted an I/T customer role. This might have contributed to NDIIPP's difficulty in engaging I/T providers effectively.[2] There is little evidence that the DPC has systematically considered available sources to devise a software acquisition strategy.

Implications

LDP costs will seldom be recoverable from intended beneficiaries, if only because these might not yet be alive. The only practical tactics seem to be support from a parent institution, philanthropic support, taxing current collection users by membership subscriptions and/or information delivery fees, and making the costs acceptable to information producers. This article offers no comments on the first three alternatives. To make the final alternative workable, we need to find ways of minimizing the effort needed, of distributing the load widely among stakeholders, and of embedding the needed work into existing document processing.

Librarians and archivists cannot themselves accomplish LDP they call for. Instead, they must seek partnerships, particularly with software engineers, effective beyond any that they have already established.

The possibility of delaying archival ingest until several decades after families have created memorabilia suggests that LDP tools should be packaged to be convenient for any citizen.

[37] For instance, the Library of Congress received Leonard Bernstein's correspondence only after he died in 1990.

Trustworthy Digital Objects

> We need to be able to reason about preservation risks in the same way as, say, an engineer might do in the construction industry, … [While our] toolkit[38] enables organisations to reason about risk at the repository level, we need similar tools to reason about risk at the object levels as well. Ross[2]

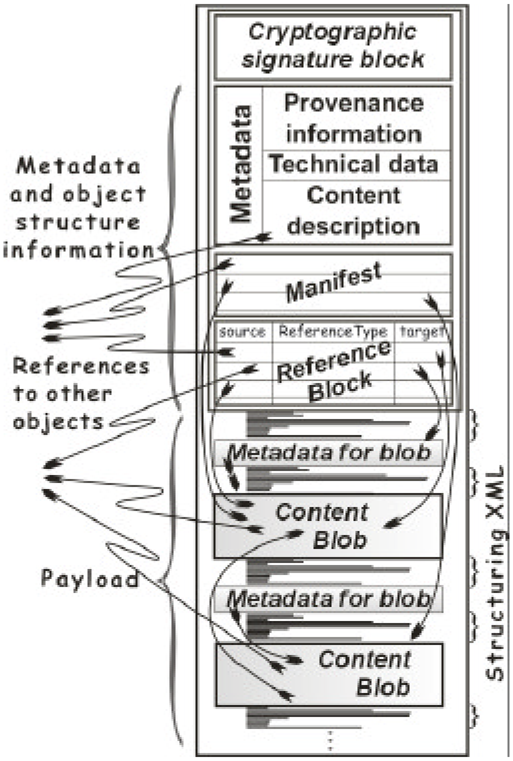

Figure 1: Trustworthy Digital Object (TDO)

An alternative for repository-centric approaches is methodology based on a LDP-ready "trustworthy digital object" (TDO).[39] Its scheme is compatible with information interchange conventions that are being developed by other authors and will be compatible with metadata standards that have not yet been finalized.

Conceptual schema suffice for analyzing whether this approach is correct and complete. More detail would obscure critical structural patterns.

Figure 1 is a convenient starting point for talking about TDO methodology. TDOs embed their own provenance metadata. Each has an RDF-compatible[40] relationship block for both internal cross-references and external references. A TDO can include any number of content blobs.[41] Each blob is more or less a conventional data file; there is no constraint on its data type.

The reference block represents a mathematical relation. Each relationship is a triple: [source, referenceType, target]. Its source and target names link to objects, into objects, or to extents within objects. Its referenceType can be an object link, but is more likely to be a relationship name.

Any intellectual work, together with context needed to understand it, is in fact a collection. A TDO can represent any information structure. It should be constructed to include whatever links its author deems important for correctly interpreting its content. The resulting structure is a potentially unbounded semantic web. Its users must decide which linked information subset interests them.

Each TDO contains technical and provenance metadata, tightly bound to the content they describe. This whole bundle is secured by its creator's cryptographic signature. Each embedded reference includes the signature of the object it links. Networked TDO repositories provide scaling and evidentiary infrastructure.[42]

This scheme supports information sharing between repositories, as well as with and among individual information producers and consumers. This can be seen by comparing the TDO schema with that for the *Object Reuse and Exchange* (ORE) initiative.[43] ORE specifies link

[38] A. McHugh et al., *Digital Repository Audit Method Based on Risk Assessment,* 2007, http://www.repositoryaudit.eu.

[39] Detail is available. Gladney, loc. cit. footnote 13.

[40] RDF Working *Group, Resource Description Framework*, 2004, http://www.w3.org/RDF/.

[41] A digital object is sometimes called a "blob" (binary large object) to emphasize that its internal structure is irrelevant for the conversation of the moment.

[42] G. Caronni, *Walking the Web of Trust*, Proc. 9th Workshop on Enabling Technologies. 2000.

[43] Open Archives Initiative, *Compound Information Objects: the OAI-ORE Perspective,* May 2007, http://www.openarchives.org/ore/documents/CompoundObjects-200705.html.

semantics within a digital object and between digital objects—semantics sufficient for specification of software enabling object sharing among repositories and end users. It extends TDO by specific relationships: “hasLineage”, “hasEntity”, “hasIdentifier”, “hasProviderInfo”, “hasDatastream”, and so on.[44] This extension informs supporting software about suitable actions, such as graph-following for retrieval and compound object construction.

ORE does not preclude other relationships. Nor does it include cryptographic signing and sealing. It would permit such extensions.

XDFU defines a competing interchange structure.[45]

Each Figure 1 content blob is represented either with ISO encoding or with a generic method using a simple virtual machine that Lorie called a Universal Virtual Computer (UVC).[22] The latter method, based on the Church-Turing thesis,[46] exploits four facts: (1) that any feasible computation can be accomplished with a Turing machine; (2) that any bit-string whatsoever can be created by some program’s execution; (3) that complete specification and an implementation of a suitable Turing machine are surprisingly concise; and (4) that their correctness can be tested today, rather than only when it would be too late to detect errors and repair shortfalls.

The effect is that any information can be represented for later reliably faithful recovery, independently of its original file type. Applications of Turing machines are familiar to computer scientists. Other interested readers are referred to careful expositions.[47]

Logical Analysis

> The task of philosophy is not to provide understanding of what is—that is the exclusive province of science. Rather, its task is the removal of misunderstandings. Rescher[48]

LDP design encounters unusual risks. Subtle design errors and weaknesses might not be discovered until information consumers access content saved many years earlier. This risk can be reduced by fundamental analysis that is usually only implicit in good engineering design. Given that our objective is knowledge communication, the appropriate sources are epistemology and philosophy of language.[49]

Design must start with conceptual models that progress from broad structure to as much detail as engineers want. Figure 1 partitions data, depicting structural aspects of archival objects. Figure 2 partitions information flow, suggesting paths and steps of object transmission among participants. Figure 3 partitions repository mechanisms, suggesting software structure.

Philosophers distinguish between objective matters and subjective choices, between essential and accidental information, and among conversation participants’ roles and contexts. They also teach explicit attention to the meaning of key words.

For instance, consider design to preserve authenticity and evidence of authenticity. A recent presentation asserts problematically, “Authenticity is difficult to define precisely and may be different for different kinds of objects in different business process contexts, leading to different

[44] See Figure 3 of Van de Sompel and Lagoze (loc. cit.)

[45] XFDU objects are structured like TDOs; CCDS, *XML Formatted Data Unit (XFDU) Structure and Construction Rules,* Sept. 2004, http://sindbad.gsfc.nasa.gov/xfdu/pdfdocs/iprwbv2a.pdf.

[46] B.J. Copeland, *The Church-Turing Thesis,* Stanford Encyclopedia of Philosophy, 2002, http://plato.stanford.edu/entries/church-turing/.

[47] Gladney, loc. cit. footnote 13, chapter 12.

[48] N. Rescher, *The Rise and Fall of Analytic Philosophy* in *Minding Matter and Other Essays in Philosophical Inquiry*, Rowman & Littlefield, 2001.

[49] Seminal works include Wittgenstein’s *Tractatus Logico-Philosophicus,* Cassirer’s *The Problem of Knowledge vol.4,* Carnap’s *The Logical Structure of the World,* Quine’s *Word and Object,* Polanyi’s *Personal Knowledge,* and Ryle’s *The Concept of Mind.* These authors were so successful that their teachings are embedded, without attribution, in current science education and often regarded as mere common sense.

preservation criteria."[50] The problem is not that the assertion is incorrect. Even a careful explanation might provoke criticism. Its problem is that its author uses it as an excuse not to try.

Somewhat better, a task force concluded, "[w]hen we work with digital objects we want to know that they are what they purport to be and that they are complete and have not been altered or corrupted."[51] This does not, however, provide much help to a software engineer.

Contrast a definition that captures more of what people mean when they describe signals, manuscripts, other material artifacts, or even natural entities, as "authentic":[52]

> Given a derivation statement R, "V is a copy of Y (V=C(Y))",
> a provenance statement S, "X said or created Y as part of event Z", and
> a copy function, "$C(y) = T_n(\dots (T_2(T_1(y)))),$"
> we say that V is a *derivative* of Y if V is related to Y according to R.
> We say that "by X as part of event Z" is a *true provenance* of V if R and S are true.
> We say that V is *sufficiently faithful* to Y if C conforms to social conventions for the genre and for the circumstances at hand.
> We say that V is an *authentic copy* of Y if it is a *sufficiently faithful derivative* with *true provenance.*

Each T_k represents a transformation that is part of a Figure 2 transmission step. To preserve authenticity, the metadata accompanying the input in each transmission step would be extended by including a T_k description. (This is not needed for steps creating identical copies.) These metadata might identify who is responsible for each T_k choice and other circumstances important to consumers' interpretations and judgments of authenticity.

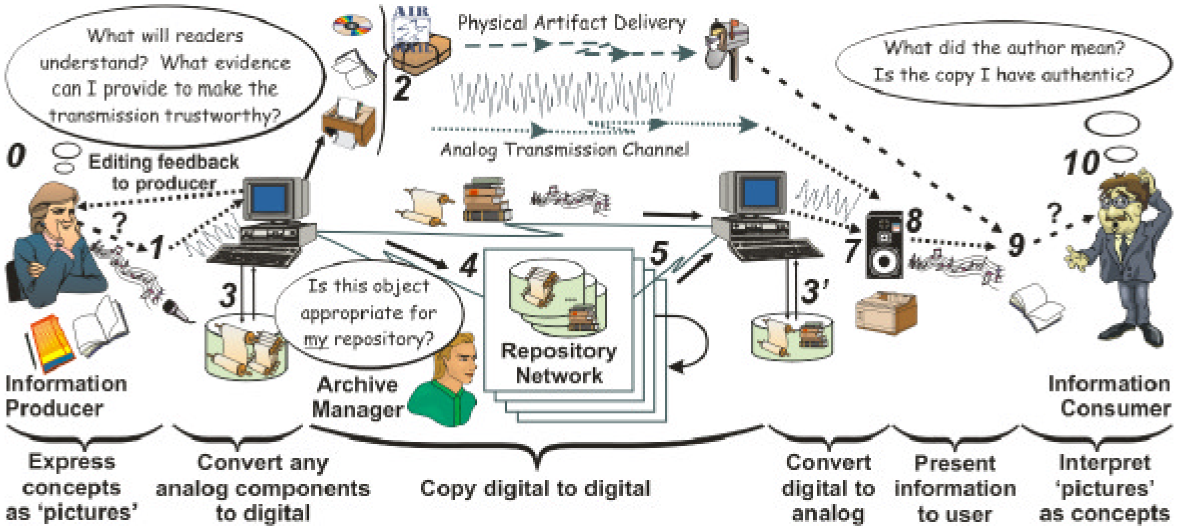
What will readers understand? What evidence can I provide to make the transmission trustworthy?
0
Editing feedback to producer
?
1
2
AIR MAIL
Physical Artifact Delivery
Analog Transmission Channel
What did the author mean? Is the copy I have authentic?
10
?
3
4
Is this object appropriate for my repository?
5
Repository Network
7
8
9
3'
Information Producer
Archive Manager
Information Consumer
Express concepts as 'pictures'
Convert any analog components to digital
Copy digital to digital
Convert digital to analog
Present information to user
Interpret 'pictures' as concepts

Figure 2: Information flow

Every document representation confounds essential information with accidental information. It might also lose essential information. For instance, in a recording of recited poetry, the speaker's voice pitch is likely to be regarded as accidental information. Social convention identifies the Figure 2 information provider as the primary authority for the distinction between what is essential and what is accidental. The secondary authority is the Figure 2 information consumer, whose notion of essential is governed by his objectives. For instance, an old book's

[50] R. Verdegem, *Back to the future: Dioscuri, the modular emulator for digital preservation*, 2007, http://www.kb.nl/hrd/congressen/toolstrends/presentations/Verdegem.pdf.

[51] InterPARES Authenticity Task Force, *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, 2004, http://www.interpares.org/book/index.cfm.

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

marginal notes might be considered essential information by a paleographer, and accidental by almost everybody else.

The most demanding scholar will want each original author's choice of every essential detail. The **0? 1** and **9? 10** transmission steps of Figure 2 are annotated with "?" marks to remind viewers that they necessarily involve subjectivity. Every other transmission step can be objectively described after it occurs and creates no difficulty if its output is identical to its input. Otherwise it injects subjective opinions of whoever chose the transformation function.

Confidence that eventual readers will understand authors' intentions is hampered by accidental information. Human dialog permits some reduction of this uncertainty, but there is no similar opportunity for archival documents. Curators will want not to exacerbate such difficulty. Combating file format obsolescence by repeated transformative migration[53] is a bad idea not only because archivists might make mistakes. Even if they accomplish precisely what they intend, they risk degrading the balance between essential and accidental information.

Discussions of the PDF format and its PDF/A variant[54] are about this problem. It affects Lorie's UVC method.[22] It is particularly bad in Rothenberg's computer hardware emulation,[55] because the architecture of the computer on which a document was created will be mostly irrelevant to what its author intends.

Some accidental information can be identified as such by comparing the document of interest to a set of documents represented with the same schema. If the set members address distinct topics, their similar attributes are likely to be accidental aspects.

Engineering Analysis

Human beings are sensitive to communication nuance. They expect this sensitivity to be reflected in information automation, perhaps not early in the history of any tool, but certainly in some idealistic future. They also expect much more of digital objects than they do of works on paper—more reliable authenticity, rapid linking to context, rapid creation of derivative works, and so on, all delivered so as to minimize human effort. Nobody should be surprised to hear that design for document handling is complicated.

What is the most important engineering ability? If the task is to build a bridge, aesthetic intuition would be desirable. More important, however, is reliable design—so that the bridge never falls. The best engineers know what could go wrong and make sure that it does not happen.

Imagine for a moment that a 50-person software development team had been authorized to respond to a promising and urgent market for LDP technology offerings. Suppose further that that other parties' software could be reused with affordable licensing fees. How might the team manager assign tasks to subordinates, knowing that he himself will be appraised in about a year for demonstrable progress towards customer service?

Design Process

Until archiving is partitioned into almost independent components, tasks that can be defined by objectively described rules are hopelessly confounded with subjective tasks that require human creativity, judgment, and taste. Many engineers favor "divide and conquer", designing no

[53] P. Mellor et al., *Migration on Request, a Practical Technique for Preservation,* Proc. 6th European Conf. on Research and Advanced Technology for Digital Libraries, 516-526, 2002.

[54] PDF/A intends to improve PDF durability. However, S. Abrams et al., *PDF/A, The Development of a Digital Preservation Standard,* SAA 69th Annual Meeting, August 2005 (http://www.aiim.org/documents/standards/PDFA69thSAA805.pdf) warns that "PDF/A *alone* does not guarantee preservation; PDF/A *alone* does not guarantee exact replication of source material; the intent of PDF/A is *not* to claim that PDF-based solutions are the best way to preserve electronic documents …"

[55] J. Rothenberg, *Ensuring the Longevity of Digital Documents*, Scientific American 272(1), 42-47, 1995.

mechanism or process until it is partitioned from the other mechanisms sufficiently with well-defined interfaces with these other components.

An effective partitioning identifies itself as such. The behavior of each partition is readily described entirely in terms of its interfaces. Such descriptions can be precise, complete, and compact. Any partition's internals can be changed without affecting other partitions. Widely used application programming interfaces and information interchange conventions become formalized as standards that are implicit contracts between independent engineers.

After engineers choose a set of partitions, they ask how to manage each, perhaps doing so in collaboration with potential participants in its implementation and use. They ask: what skills are needed to design it? To implement it? What kind of organization would be its natural "owner", some governmental department or private sector elements? Which design and implementation tasks are already addressed in available components?

Complete design for a software component is rarely affordable or feasible in its early versions, if ever. Nuanced human preferences are mostly unknown before users react to prototypes, and often imperfectly known even long after implementations are deployed. Practical designs must allow for non-disruptive changes and additions to every deployed version.

A paper describing software design—either high-level design or fine details—is likely to proceed logically. Typically it will start with business circumstances and objectives, moving successively to technical objectives, architectural principles and applicable standards, design layering and partitioning into practical chunks, choices of chunks to implement early, and several stages of increasing detail for selected chunks.

This sequence is logical rather than chronological. In practice, logically later stages suggest changes to earlier stages. Sometimes called “waterfall” methodology, iterative refinement may need to continue as long as the designed software is heavily used. What the current article provides is merely an idealized sketch that might never correspond to any implementation.

Producing LDP technology is unaffordable except by modest modifications of what exists for daily document handling. A software team manager would need to think about many software components and many identified requirements. To consider himself well informed, he would need to master hundreds of articles. Only an orderly approach would be workable.[56] A promising early step would be to create a comprehensive graph of content object classes, available technologies, and required human skills, tagged with references to articles promising solution components. Such a graph could link requirements to sources of software and know-how and identify missing technologies.[57]

Structuring a Solution

> The work achieved so far leads to post process the collections to ingest them in the trusted repositories (a large scale migration test). The main question for the future is how to do it in the workflow of collecting, pre-ingest and ingest at the Web scale. Lupovici[58]

Many facts—the number of digital objects, the number of authors, the speed of information creation and dissemination, the expectations of citizens, the cost trends of technology, relative skills of different communities, and so on—suggest shifting as much as possible of the responsibility from repository institutions to those who are served—information producers and information consumers.

[56] J. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole, 2000.

[57] Spreadsheets and mapping tools can speed the analysis of the hundreds of facts that merit attention. See, for instance, the no-charge CMAP offering at http://cmap.ihmc.us/. Such tools have entered the mainstream; see Brian Hindo, *Inside Innovation: Software that maps*, BusinessWeek 19-21, Nov. 26, 2007.

[58] C. Lupovici, *Archiving the Web: the mass preservation challenge*, Conference on Digital Preservation, Nov. 2007, http://www.kb.nl/hrd/congressen/toolstrends/presentations/Lupovici.pdf.

This will be feasible only if creating LDP-ready information is a very inexpensive addition to editing already required. LDP tools must be packaged within information producers' tools. Since document producers already want their output to be valued, it should be possible to persuade them to do additional work that does not take much time and is easy. As an incentive, prestigious repositories might limit what they accept for indexing and distribution to LDP-ready content.

Information might travel from its producer to its consumer by any of several paths (Figure 2). Participants will want the details of information copies to be independent of transmission channels. The simplest way of ensuring this is to arrange that the copy **3'** in a consumers' PC has precisely the same bit pattern as the copy **3** prepared by the producer.

In some unrealistically ideal world, LDP mechanisms would be automatic side effects of here-and-now information preparation, management, and exploitation. However some needed metadata choices and trustworthiness decisions are intrinsically subjective. Subjective decisions are often difficult matters of judgment and taste, particularly when they depend on community consensus. Conceptual models grounded in scientific philosophy can be used to separate what is objective, and therefore can be automated, from what is subjective, and therefore beyond automation.[59] Such analysis is fundamental for designing semi-automatic software to minimize human work needed.

Preparing and Examining Information

Preserving any information corpus requires metadata that is rarely included in today's Internet deliveries. Writers are typically the participants best informed about missing metadata facts. Only an information consumer can properly estimate damage to his affairs should information be incorrect. Such circumstances and scales already mentioned suggest shifting preparation for archiving, and also TDO inspection, from repository employees to repository clients to the extent possible.

Information producers will need to inspect older TDOs. Support for information consumers will therefore be a subset of that for information producers. Both roles are therefore treated here.

What will be needed is mostly PC software. A TDO editing program can be created by adapting text, picture, and structured data editors. It can be semi-automatic, choosing representations for data blobs and prompting document creators for data needed, such as provenance metadata, cryptographic signatures, and certified contextual links. Such software would also validate field entries, prompting for problematic data. Tools for extracting TDO portions and for traversing metadata and signature graphs can have graphic front ends that might look like Figure 1.

The authenticity definition of **§Logical Analysis**, used together with Figure 2, can guide engineers towards distinguishing questions that can be objectively answered from those with subjective elements. Such distinctions help them design semi-automatic procedures that prompt users for those answers that only human beings can provide, together with self-identification that conveys authority.

When we first put forward the TDO scheme about four years ago, its formal specification included no syntactic schema. This was partly because we were not yet confident that the conceptual structure was adequate, partly because metadata standards were still being debated, and partly because no community seemed ready to work toward syntactic standard consensus that is a *sine qua non* for information interchange. Today, all these circumstances seem changed. Many articles address metadata for LDP. Workgroups are struggling for standards consensus for document sharing. Many XML editors are available. Tools being

[59] Gladney, loc. cit. footnote 13, chapter 4.

developed for metadata extraction include the NLNZ Metadata Extraction Tool[60] and the PLANETS technical metadata extractor.[61] At least three contenders for digital object packaging have appeared.[62] Such tools could be embedded as document editor extensions.

Handling relatively simple Figure 1 content blobs might avoid the seeming complexity of UVC methodology,[22] depending instead on long-term durability of standard formats.[63] The scope of this approach is unclear, because some formats are surprisingly complex.[64] Whether a software developer uses ISO- or UVC-methodology, he can hide such complexity from "the poor user". The ISO method might seem simpler than the UVC method without in fact being so, an opinion influenced by how comfortable its holder is with computer science.

Archiving Services

Figure 3 corresponds to the much-used OAIS Functional Entities depiction,[65] but emphasizes software function partitioning and layering. Its largest box, "Archival Institution" depicts a complete repository institution, including exemplary archivists. Egger suggests that "the OAIS model has several shortcomings … as a basis for developing a software system. It is therefore necessary to develop additional specifications which fill the gap between the OAIS model and software development."[66] The current section sketches a response.

[60] National Library of New Zealand, *Metadata Extraction Tool,* 2003, *http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool*.

[61] M. Thaller, *Characterizing with a Goal in Mind: The XCL approach,* 2007, http://www.kb.nl/hrd/congressen/toolstrends/presentations/Thaller.pdf.

[62] XFDU, loc. cit., footnote 45.

XAM (eXtensible Access Method), http://www.snia.org/forums/xam/technology/specs/.

WARC (Web ARChive format), http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml.

[63] A standard encoding can be used only if it will be durably intelligible. This might be true for JPEG 2000 representation of images. See R. Buckley, *JPEG 2000 - a Practical Digital Preservation Standard?* DPC Technology Watch Report, Feb. 2008.

[64] For instance, the popular TIFF format has about 14 variants, each requiring somewhat different handling, and over 70 tag types. See http://hul.harvard.edu/jhove/using.html#tiff-hul.

[65] CCSDS 650.0-R-2, *Reference Model for an Open Archival Information System* (OAIS), 2001, Fig. 4-1.

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

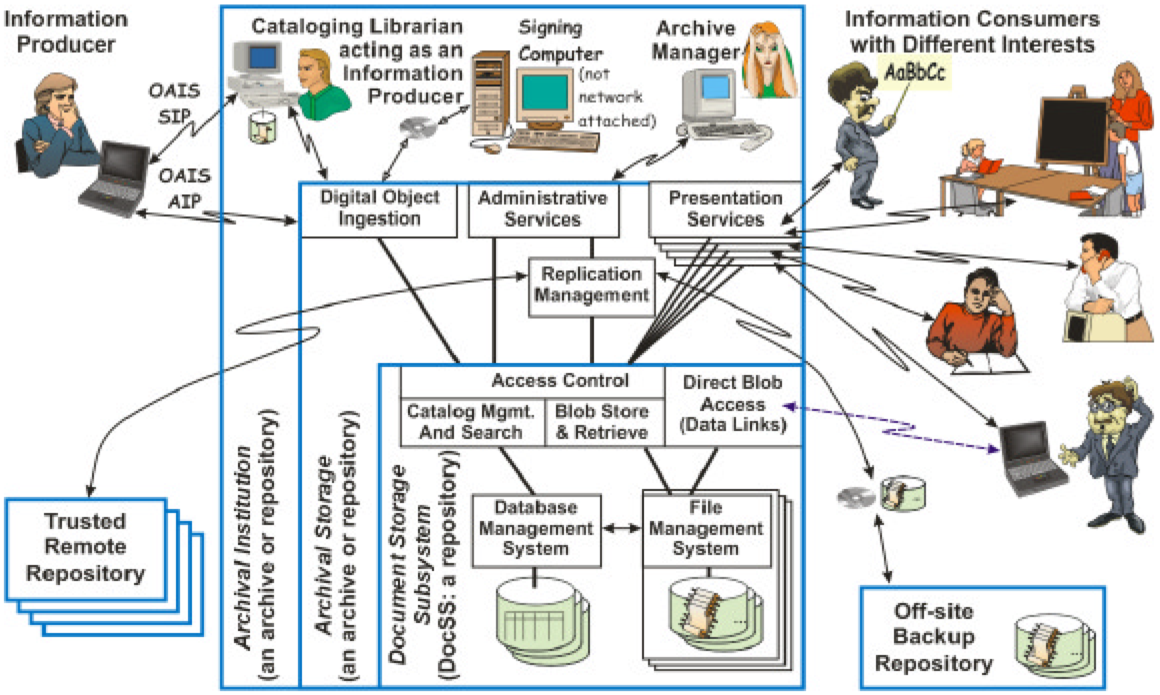
Information Producer
OAIS SIP
OAIS AIP
Cataloging Librarian acting as an Information Producer
Signing Computer (not network attached)
Archive Manager
Information Consumers with Different Interests
AaBbCc
Digital Object Ingestion
Administrative Services
Presentation Services
Replication Management
Access Control
Direct Blob Access (Data Links)
Catalog Mgmt. And Search
Blob Store & Retrieve
Database Management System
File Management System
Trusted Remote Repository
Archival Institution (an archive or repository)
Archival Storage (an archive or repository)
Document Storage Subsystem (DocSS: a repository)
Off-site Backup Repository

Figure 3: Nested repositories and human interfaces

The interface between an archival storage implementation and its users—archivists, information producers, and information consumers—can and should be with client/server relationships. The differences between service interfaces for archivists and for their clients will be more or less similar to the differences among interfaces for different client categories. For instance, the Figure 3 administrative services box could have been depicted as yet another presentation service box. User privilege differences will be reflected by access control database records.

The second largest Figure 3 box, "Archival Storage", depicts a hardware/software complex without suggesting much about how functionality might be partitioned among multiple computers. Most of its components become tailored for specific repositories and specific repository clients only by way of table entries chosen interactively by archive managers.

Data replication sites might be administratively independent services whose relationship to library content is similar to that of human users. Thus access control needs to intervene between stored data and both human users and replication services. Of course access privileges for a replication site are likely to be different from those for a human user, but this will be similar to how privileges differ among human users.

The boundary between the Figure 3 archival storage and its document storage subsystem distinguishes components that are likely to differ among institutional repositories from those whose functionality is the same in all repositories. This boundary corresponds to a software interface standardized in 2005—the *Content Repository API for Java* called "JSR 170" and its "JSR 283" refinement.[67] Repositories will choose among storage subsystem offerings to accommodate different hardware, different storage capacities, and different data traffic.

The archival storage layer includes only very little code implementing LDP functionality. An exception might be its ingestion component, with logic for deciding whether submitted

[67] See http://jcp.org/en/jsr/detail?id=170 and a *JSR 170 Overview* by R.T. Fielding. JSR 283 provides improved scaling, interoperability and access control. D. Nuescheler, the leader of these initiatives says, "JSR 283 [and JSR 170] … produce a content repository API that provides an implementation-independent way to access content bi-directionally." See http://jcp.org/en/jsr/detail?id=283.

documents conform to institutional criteria, possibly tagging stored objects to indicate different quality levels. A consequence is that most LDP software can be written for personal computers, a happy outcome!

Software supporting librarian participation in packaging information for preservation can be the same as that for information producers who are not repository employees.

Implementation of the low-level part of LDP, saving bit-string patterns, is more advanced than the portions already discussed. LOCKSS[68] enforces replication rules above database and file system interfaces. Its deployed version might need access control enhancements to permit different privileges for different remote repositories. Commercial storage management offerings include sophisticated support for safeguarding bit-strings, particularly in large "server farms". Scaling and automation for vast collections will hopefully be addressed by an industry workgroup.[69]

Discussion

For almost twenty years, the digital preservation community has sought practical methods of ensuring the longevity, integrity, authenticity, and interpretability of digitally recorded information—methods that will handle immense numbers of objects and deliveries. Its search space has been surprisingly narrow—methods for processes within archival institution perimeters, and methods of persuading clients that archival deliveries are trustworthy. Many years work without a solution should persuade participants to broaden their search.

DPC literature starts with a premise that it never questions. It assumes that LDP objectives can be achieved by modestly adapting how libraries and archives manage information on paper and other material media. Of course the purpose of digital repositories is much the same as that of libraries and archives established in the late 19th and the 20th century—storing and delivering written information that is authentic from its recipients' perspectives, even many years after the carrying documents and records were created. We agree with the apparent consensus that this should be what we seek under the rubric "digital preservation". However, the search for how best to accomplish it should include means without obvious analogs in repositories for paper.

Repository procedures alone cannot accomplish what's needed, partly because of the sheer number of digital documents and partly because digital objects are very easily changed. No plausible investment in repository institutions would change this fact. Workload needs to be shifted from digital collection managers to information producers and consumers—people and enterprises with economic incentives for creating and using preserved information. And the procedures for handling objects intended for preservation need protection against inappropriate modification, including sabotage while copies are in transit to or from repositories.

Inattention across the boundary between the DPC and essential engineering professions has clouded the distinction between what is not known and putative LDP software that has not been adapted, packaged, and deployed to please archivists and librarians—the distinction between fundamental invention and SMOP.

How to achieve LDP becomes obvious if one shifts attention from an archetypical archive to an archetypical saved object. As with diplomatic paper, the trick is to ensure that each document is fortified so that improper modification is readily discovered. This can be achieved by binding a document's parts firmly to one another and sealing the whole with a signature that is very difficult to forge.[42] Of course, how to accomplish this is different for digital documents than it is for paper documents.

[68] See http://lockss.stanford.edu.

[69] The Storage Networking Industry Association (SNIA) has recently established a "100-year archive initiative." See http://www.snia.org/forums/dmf/programs/ltacsi/100_year/.

With the shift from a repository-centric approach to the TDO approach, the scaling problem emphasized by Lupovici[58] vanishes without further effort!

The technically most difficult goal is reliably durable intelligibility of saved content while knowing little about future computers. That people expect better representation fidelity and more convenience than ever before amplifies the challenge, as does rapid change of natural and technical languages. 18th-century scholars spent lifetimes studying how to interpret early manuscripts. Tomorrow's students will expect their corresponding work to be rapid without requiring years of specialized study. The trick is mechanical assistance to render content in vernacular of the time. How to accomplish this is known.[22] Today's problem is less that the method is difficult than that its principles are unfamiliar to almost everyone interested.

The biggest social barriers to LDP progress seem to be market barriers. Private sector enterprises that might benefit from LDP are preoccupied with short-term concerns. Research libraries, archives, and other cultural institutions do not present themselves to technology providers in the guise of customers, but as supplicants for free technology and charitable support. Free market economies, of course, provide them little alternative. However, they have not much considered indirect opportunities: reducing the real expense of LDP by thorough automation, shifting workload from central services to their clients, burying LDP expense by integrating it into day-to-day content handling, and perhaps further tactics yet to be devised.

Nothing above is intended to suggest marginalization of repository institutions. However, to avoid marginalization[70] they will need to shift to roles for which they are better suited than competitors. Although the precise nature of such roles is still a subject for inquiry and social experimentation, plausible suggestions can be made. In addition to well-known digital library services, such as organizing collections and creating indices for information discovery, archives can make themselves essential for managing replicated copies, for administering quality criteria for content formats and metadata, and for creating and maintaining a trust web exploiting cryptographic signatures.

Next Steps

Anyone wanting to advance preservation practice is faced with an immense number of pertinent facts. These include, but are not limited to, literature comprising hundreds of DPC articles and thousands of computer science and engineering articles, hundreds of software offerings for document editing and management, more than a thousand file formats, varied intellectual property rules, and different expectations and skills in different stakeholder communities. The pertinent literature is voluminous, but laced with redundancy.

Remedy for such problems begins by broad analysis to identify the most promising projects and to help partition what's needed into tasks for almost independent teams. We need systematic examination of LDP literature. One effort would organize needs into formal engineering requirements. With each "line item" summarized in 2-5 lines, this document might have 100 pages. A second effort would create a pictorial map of candidate technology and software designs, labeled with pointers to detailed information. These two information resources would need to be iteratively refined as long as LDP development continues.

Software components exist for handling nearly every technical requirement. Adapting and integrating them will be business-as-usual software implementation. New tools continue to appear. Some potentially help with LDP. For instance, representing complex structures might be eased by a general format described in a prepublication paper.[71]

[70] Marginalization of professional archivists has been hinted at from within the profession. For instance, see Roy C. Turnbaugh, *What Is an Electronic Record,* pp.27 in Bruce W. Dearstyne, *Effective approaches for managing electronic records and archives*, 2002.

[71] H. Ishikawa, *Representation and Measure of Structural Information*, http://arxiv.org/abs/0711.4508.

Finally, we should begin building selected solution components for selected information genera. Of course, this cannot happen without customers and funding!

Conclusions

Everything under the sun has been said before. However, since nobody listened … – attributed to André Gide

Ross summarizes preservation state of the art with "after more than twenty years of research … the actual theories, methods and technologies that can either foster or ensure digital longevity remain startlingly limited."[2] Building on a published conceptual solution, we have described an engineering program that answers this challenge.

Our method addresses all extant and future data formats and can ensure testable integrity and authenticity, even for information that tempts fraudulent modification. It scales, starting by supporting information producers' LDP packaging, and finishing by allowing information consumers' tests that information received is trustworthy. Deployment need not disrupt today's repository services. It achieves all this by shifting attention from design for "Trusted Digital Repositories" to design of "Trustworthy Digital Objects."

A remaining challenge is to hide arcana from end users. Another is to persuade "buy in" by repository institutions, which could create a trust environment by managing hierarchical cryptographic signatures. This would be in addition to their providing information-finding services and preserving bit-string integrity and accessibility.

Our description provides sufficient guidance so that any senior software engineer could direct the next project steps.

We do not claim that TDO methodology is the only possibility in its class. It would be good to have an alternative for critical comparison. However none has been proposed. Nevertheless, we invite robust criticism of this and our other cited work.