

# Economics and Engineering for Preserving Digital Content

**H.M. Gladney**  
HMG Consulting

**Abstract:** Progress towards practical long-term preservation seems to be stalled. Preservationists cannot afford specially developed technology, but must exploit what is created for the marketplace. Economic and technical facts suggest that most preservation work should be shifted from repository institutions to information producers and consumers.

Prior publications describe solutions for all known conceptual challenges of preserving a single digital object, but do not deal with software development or scaling to large collections. Much of the document handling software needed is available. It has, however, not yet been selected, adapted, integrated, or deployed for digital preservation. The daily tools of both information producers and information consumers can be extended to embed preservation packaging without much burdening these users.

We describe a practical strategy for detailed design and implementation. Document handling is intrinsically complicated because of human sensitivity to communication nuances. Our engineering section therefore starts by discussing how project managers can master the many pertinent details.

To make discoveries one must not be too anxious about errors. One must be willing to state a theory clearly and crisply and say as the physicists do: I have worked on this for a long time, and this is what I have come up with; now tell me what, if anything, is wrong with it. And before one ever gets that far, one has usually found many of one's own attempts faulty and discarded them. What is most needful is by no means certainty but rather, to quote Nietzsche's happy phrase, "the courage for an attack on one's convictions."  
Kaufman<sup>1</sup>

## Introduction

Long-term preservation of digitally represented information (abbreviated LDP below) has been discussed for over a decade in articles by librarians, archivists, and information scientists, an informal group sometimes called the digital preservation community (DPC). This literature is mostly disjoint from that of computer science, software engineering, and information technology industry.<sup>2</sup> From the perspective of an engineer accustomed to the pace of I/T research and development, progress reported in the DPC literature is surprisingly slow.<sup>3,4</sup>

---

<sup>1</sup> W. Kaufman, *Discovering the Mind: Goethe, Kant, and Hegel*, 1980, p.8.

<sup>2</sup> This is suggested by inspecting literature citations. DPC articles rarely cite from ACM or IEEE periodicals, or from the trade literature that is the best source for emerging content management tools. And the latter literatures seem not to have noticed that anybody cares about long-term digital preservation.

<sup>3</sup> S. Ross, *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, 11<sup>th</sup> European Conference on Digital Libraries (ECDL), Budapest, September 2007.

<sup>4</sup> Ross (loc. cit.) 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.

International conferences and workshops occur in rapid succession. However, these seem to discuss little more than LDP urgency and social processes for collaboration, without introducing novel concepts or addressing engineering specifics.<sup>5</sup> A recent notice is typical:

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<sup>6</sup>

Contributors pay little attention to pertinent information technology provided to the public, including tools developed for business and private use. 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. Since progress towards comprehensive digital preservation seems to be stalled, it is time to consider fresh approaches, particularly those suggested by conventional software engineering practice. We sketch an engineering approach to the technical component of LDP.

### Summary of What Follows

[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.<sup>7</sup> Ross

For a technical proposal of an unorthodox "concise and well-developed strategy" to be credible and also useful to software engineers, it must provide three things:

- 1) An explanation of why continued effort along prior directions is unlikely to succeed;
- 2) A description of principles which are likely to enable a successful solution; and
- 3) Enough strategy description to demonstrate that the proposal is an actual solution and to guide a software development team as it begins its work.

The first issue is dealt with in our *Literature Analysis* and *Economic Analysis* sections. The former section draws heavily on Ross, which represents DPC know-how and views comprehensively enough to use as a template, starting with its "key research challenges."<sup>3</sup>

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<sup>4</sup> reflects a common view that a research topic exists for any information management need unsupported by available software. In 1980's IBM Research corridors, the boundary between research and practical engineering was called "SMOP"—"a simple matter of programming" or "a 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 fundamental questions, allowing most of the work to be passed on 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."

---

<sup>5</sup> Readers can judge our opinions 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.

<sup>6</sup> C. Jackson, DPC, e-mail on 15<sup>th</sup> November 2007. See <http://www.alliancepermanentaccess.eu/>. Similarly, H. Hockx-Yu, *Progress towards Addressing Digital Preservation Challenges*, *Ariadne* 53, Oct. 2007, reporting a recent LDP conference, is more about objectives and process than about progress.

<sup>7</sup> Listed in [footnote 66](#) of Ross, loc. cit.

Widely known economic facts suggest that some preservation tactics will be impractical for the flood of information expected.<sup>8</sup> Citizens' resources, interests and abilities have changed immensely since 1950. Although such economic circumstances do not themselves suggest how to proceed, they help identify what will not work.

The second issue, conceptual design based on sound epistemology, is addressed in our **Trustworthy Digital Objects**<sup>9</sup> (TDO, Figure 1) and **Logical Analysis** sections. Since we have addressed these topics, our treatment here is a summary with pointers to more detailed work.

The final issue is dealt with in our **Engineering Analysis** section. In the spirit of SMOP, this is not primarily about design. Instead, it describes a strategy for handling unavoidable complexity and issues of practical implementation. Ideally, separate teams would be able to provide solution components for later integration to please repository institutions, producers of digital content, and eventual users. Engineers know how to partition identified challenges into tractable components and manage their integration. The article identifies these and suggests where practitioners can find implementing software.

### Scope Limitations

Some DPC-identified LDP challenges are technical. Others are matters of content selection, service delivery by archives,<sup>10</sup> social acceptance, technology uptake by prospective users, and community education. The strategy described below is limited to technical aspects appropriate for software engineers. On the other hand, skillfully implemented technical solutions can be deployed to much reduce the other challenges.

The article takes "digital preservation" to mean "mitigation of the deleterious effects of decay, obsolescence, and human error or misbehavior that might impair the integrity or intelligibility of digital document copies long after originals were created." This choice partitions archiving into digital library services for a perfect world and compensation for unavoidable degradation. (Here, in a "perfect world" the utility of saved information does not decrease with time.) We leave to other authors questions of digital library infrastructure to please "perfect world" clients.

The article uses TDO methodology to help make the engineering description explicit. However, it argues neither that TDO architecture is the only solution in its class nor that it would be optimal for millions of users and billions of objects. Such topics and justifications have been published elsewhere.<sup>11</sup>

As is common in computer science, the TDO design objective is to handle all the most difficult and general cases—those that are tempting targets for fraudulent modification, objects represented with unusual file formats, and computer programs whose behavior might be drastically compromised by tiny changes. Optimization for easier and less sensitive cases is left for other articles and to other authors.<sup>12</sup>

Specific technologies are mentioned as examples, not as recommendations among alternatives. Such examples demonstrate the feasibility of adapting much that is already available. For optimal selection from completing software packages, one would need the comprehensive survey we call for in **§Engineering Analysis**.

---

<sup>8</sup> 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/](http://www.emc.com/about/destination/digital_universe/)

<sup>9</sup> H.M. Gladney, *Principles for Digital Preservation*, Comm. ACM 49(2), 111-116, February 2006.

<sup>10</sup> 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>, viewed 15-Nov-07.

<sup>11</sup> H.M. Gladney, *Preserving Digital Information*, Springer Verlag, 2007.

<sup>12</sup> For instance, other authors propose methods for widely used file formats. Such treatments need to be re-examined scrupulously because they might mishandle the distinction between essential and accidental information, inadvertently misrepresenting authors' intentions.

## Literature Analysis

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

DPC literature seems very repetitive, describing how various groups are exploring more or less the same ideas, with little attention to know-how originating outside their small community.<sup>13</sup> It identifies no technical challenges for which in-principle solutions are unknown. Its most prominent focus has become organization and management of archives, sometimes under the label “Trusted Digital Repositories” and more recently under “Trustworthy Repositories”.<sup>14</sup> However, nobody has published a persuasive argument that repositories can solve what this literature calls “the problem” or “the challenge.”<sup>15,16</sup>

An outside reader would find it difficult to determine which publication corpus the DPC targets for preservation. Is it academic and cultural literature typically similar to that held on paper in research libraries? Or is it all content that our descendants might find interesting, including practical information such as engineering details of the utilities under the streets of major cities, legal and financial records, personal medical records, or photographs of “dear old Aunt Jane”?<sup>17</sup>

### A Summary of Challenges

Ross<sup>3</sup> articulates a preservation 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 that he suggests need attention are:

1. Restoration: when digital objects have broken and are restored, how can we ensure and verify the syntactic and semantic correctness of restored versions?
2. Conservation for digital objects believed to be intact: what methods can we use to ensure the integrity of copies whose authenticity and provenance can be verified?
3. Collection and repository management: what archival methods and software design will satisfy the quality expectations of eventual collection users?
4. Risk management: what tools and methodology can help preservation practitioners quantify content risks and benefits to choose how to spend limited resources?

<sup>13</sup> Quality concerns are not unique to Information Science literature. See, for instance, David Lorge Parnas, [Stop the Numbers Game](#), Comm. ACM 50(11), 19-21, November 2007.

<sup>14</sup> Center for Research Libraries, *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, 2007*, <http://www.crl.edu/PDF/trac.pdf>, viewed 29-Nov-07.

<sup>15</sup> Such concerns are summarized by H.R. Tibbo in a 15<sup>th</sup> Oct. 2007 posting to the *MOIMS-Repository Audit and Certification* blog ([moims-rac@mailman.ccsds.org](mailto:moims-rac@mailman.ccsds.org)). It included, “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.

<sup>16</sup> 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.”

<sup>17</sup> DPC articles sometimes mention the collection at [Archivo General de Indias in Seville, Spain](#), evidencing the importance of commercial records even if the objective is limited to content for scholarly research.

5. Preserving interpretability and functionality of digital objects: 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. Automation in preservation: given the immense number of digital objects, what automation opportunities exist and how can these be practically realized?
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 to address each challenge depend on its nature. Is the challenge technological or a matter of social acceptance, such as that required for workability of data interchange standards? Is it that no-one at all knows a solution, or 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 librarians and archivists? To what extent is the challenge one of community education and training?

For each challenge, consider an engineer's opinion as a conjecture for validation or rejection.

1. Restoration: the cost of restoration can be borne by those who want the content in question. Work on restoration is not urgent. It can be mounted when its beneficiaries are prepared to foot R&D expenses.
2. Conservation for digital objects believed to be intact: how to do this is known, but faces SMOP. Open questions include whether cheaper overall methods exist, whether and when a file-type-specific method is preferable to the generic solution, and how to persuade stakeholders to evaluate the solution, considering whether or not to use it.
3. Collection and repository management: there exist over 100 digital library packages. Each repository institution can choose one that suits it well and seek tailoring to its idiosyncratic circumstances. LDP requires at most modest additions to available digital content management offerings.
4. Risk management: the current article suggests a first step to quantifying risks and to choosing investment strategies. Crude risk estimates will be good enough.
5. Preserving interpretability and functionality of digital objects: a generic method is known, but has neither been noticed by archivists nor realized in deployable software.<sup>18</sup>
6. Collection cohesion and interoperability: what context needs to be explicit for any content body is a mostly subjective decision not amenable to automation. Providing semantic context, starting with the meanings of key words and phrases, is a topic for authors and subject experts. Being required for today's services, this need not be considered an LDP topic except when long-term utility poses specifically identified additional requirements.
7. Automation in preservation: engineering methodology discussed below treats only how to manage each individual collection. Handling tomorrow's immense scales is a task worthy of the best skills and efforts towards automating every possible step.

<sup>18</sup> 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](http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf) and <http://www.erpanet.org/assessments/show.php?id=1093856769&t=2>, viewed 10-Oct-07, *UVC for Images*, [http://www.kb.nl/hrd/dd/dd\\_onderzoek/uvc\\_voor\\_images-en.html](http://www.kb.nl/hrd/dd/dd_onderzoek/uvc_voor_images-en.html), viewed 10-Oct-07.

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.

8. Preserving digital object context: existing efforts to standardize technical and provenance metadata are on the right track.<sup>19</sup> How to reliably package critical content with any object is known, as is how to reliably link widely used context. The critical challenge is to persuade information producers to use such tools.
9. Engineering and deploying storage technologies: how engineers can exploit available software offerings is discussed below.

These requirement and response statements are oriented towards management processes. Organizing software engineering and evaluating its products demands a different formulation. Each requirement needs to be worded to enable objective assessment of the extent and quality of purported solutions.

## **Traditional 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<sup>3</sup>

Bankers, businessmen, and attorneys have long understood the importance of integrity, authenticity, and evidentiary audit trails of provenance information—objectives that Ross identifies and treats as “principles” of diplomatics<sup>20</sup> and archival science. Alerted by widespread Internet chicanery, a wary public is rapidly becoming sensitized to issues of content reliability.<sup>21</sup> Such objectives are not questioned, if they ever were. The issues are practical mechanism and persuading its widespread usage.

The evolution Ross describes is primarily realization that the principles of diplomatics are important for many information genres in addition to royal treaties of the middle ages.

Ross asserts that an archive user can “determine whether a digital object is what it purports to be ... only if institutions have adequately and transparently documented the processes of digital entity ingest, management and delivery.” However, DPC literature does not teach how to do this for digital collections that are tempting targets for fraudulent modification. In fact, no-one confidently knows how to prevent misbehavior by invading software.<sup>22</sup>

The shift from paper to digital media and digital collection scales make it doubtful that procedures designed for and by archives many decades ago will work well in the 21<sup>st</sup> century. Some patently collapse. An example is reader surveillance by uniformed custodians, as the British National Archives uses to protect 18<sup>th</sup>-century paper. Some will prove unaffordable. An example is having of large numbers of librarians creating descriptive and technical metadata for every interesting holding. Some traditions will prove to have good digital analogs. An example is notary publics’ embossed seals combined with their handwritten dates and signatures.

Nobody seems to have systematically examined traditional methods designed for paper, seeking all that can be adapted well. For instance, “a well-documented chain of custody”<sup>3</sup> within an institution would be neither practical nor sufficient, if only because information-tampering during network delivery to and from repositories would be easy. In contrast, it is easy to add a modern equivalent of a royal signature and seal to any valuable document. The sealed

<sup>19</sup> Northwestern Univ. Digital Library Committee, *Inventory of Metadata Standards and Practices*, at <http://staffweb.library.northwestern.edu/dl/metadata/standardsinventory/>, viewed 24-Nov-07.

<sup>20</sup> Diplomats started with issues of trust for state correspondence. L.E. Boyle (deceased Librarian of Biblioteca Apostolica Vaticana), *Diplomatics*, in J.M. Powell, *Medieval Studies: An Introduction*, Syracuse U.P., 1976, pp.69-101, summed up the key questions as “quis?, quid?, quomodo?, quibus auxiliis?, cur?, ubi? quando?” (Who? What? How? What’s related? Why? Where? When?)

<sup>21</sup> 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.

<sup>22</sup> Tools for “Internet exploits” are widely publicized. For instance, see *SANS Top-20 Internet Security Attack Targets*, 2006, <https://www2.sans.org/top20/>, viewed 24-Nov-07.

package would contain the technical details and provenance of each modification in a standard format<sup>23</sup> and tightly bound to the content these metadata describe, all secured by the cryptographic signature of whoever made the change.<sup>24</sup>

One aspect of paper-based information storage has a ready digital equivalent—replication to protect against loss of the last copy. As suggested by Figure 3, it is easily automated.

## Economic Analysis

[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<sup>3</sup>

The Information Revolution is relatively new, beginning about 50 years ago and still evolving rapidly. In contrast, the current infrastructure for managing content on paper is about two centuries old. Widespread access to libraries began about 130 years ago with Carnegie funding. The first widely used digital library offering appeared only about 20 years ago,<sup>25</sup> and librarians and archivists began to pay attention to LDP even more recently.

A century ago, few people had the time, resources, education, or inclination to use cultural resources. During their working years, our parents had little energy or money for deep reading, concert attendance, or other cultural activities. Nor did many of them have college educations that teach how to benefit from and enjoy recorded information. That changed about 50 years ago, as evidenced by large increase in university enrollments and faculties. Our children take digital content for granted, and are likely to use it more and more, perhaps shifting the mixture of what the public might judge preservation-worthy.

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. Engineering and scientific records were notebook scribbles. Creating and disseminating written works was slow and labor-intensive. Such factors are massively changed today. A large fraction of newly generated information can be fetched only digitally.

The research community has increased by a factor of about 100 since 1930. For instance, physics conferences in the 1930's typically attracted about 50 participants. Today, American Physical Society conferences have about 5000 participants. And the large increase in college faculty sizes, coupled with the publish-or-perish syndrome, has created a periodical subscription crisis for the universities. Our reading is unduly burdened because only a small fraction of scholarly articles convey anything new. Similar difficulties are evident in the popular press and business trade press, with several dozen subscription invitations appearing in some home mailboxes every month. Many subscriptions are free, but a significant drain on productive time for anyone who looks at them.

The development and deployment of information tools regarded with confidence by a dependent public and an interested professional community are 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.

---

<sup>23</sup> This might be the *Metadata Encoding and Transmission Standard (METS)* described at <http://www.loc.gov/standards/mets/>, viewed 24-Nov-07.

<sup>24</sup> Gladney, loc. cit. footnote 11, §11.1.3.

<sup>25</sup> IBM Digital Library, now called IBM Digital Content Manager, was first marketed in 1993 after differently-named pilots had been used by customers for about 2 years. See H.M. Gladney, *A Storage Subsystem for Image and Records Management*, IBM Systems Journal 32(3), 512-540, (1993).

## DPC and Stakeholder Communities

The professional community ready to invest time and energy into preserving digital content numbers between 500 and 5000 individual librarians, 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.<sup>26</sup> Exceptions might be found in the pharmaceutical and entertainment industries because their key assets include electronically recorded information holdings. Nevertheless, reports of significant private-sector LDP investment are hard to find.<sup>27</sup>

The information technology community providing tools with which businesses, academics, and private individuals create and share digital content is much larger. Computer programmers and software engineers 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 and be workable for their customers. This includes lowering the apparent costs of participating in content management and exploitation. Examples of success are readily found. For instance, fail-safe storage for a home computer network today costs about \$0.40 per gigabyte, and reliable storage for private e-mail is available at no cost.

How many people add to the information pool that is candidate for digital preservation? The number seems to be between about 10,000,000 and 100,000,000.<sup>28</sup> If technology to record personal daily activities<sup>29</sup> appeals to the public, this number will increase ten-fold.

Each individual seeks to build [his] own ... archives about [his] own family and heritage. One linguist ... says that “as individuals, 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<sup>30</sup>

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<sup>31</sup>

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 wanted to contribute to significant progress, they would need to make radical methodological changes.

For instance, the DPC hardly has sufficient numbers to keep up with the changes being made by the software engineering community and ordinary citizens’ software uptake.<sup>32</sup> The pace of

<sup>26</sup> Business priorities for discretionary expenditures are elsewhere, such as toward avoiding disclosure of customers’ private information, near-term profitability, and document retention for audit as required by the Sarbanes-Oxley legislation. See <http://www.soxxlaw.com/>, viewed 24-Nov-07. E. Thornton, *Perform or Perish*, BusinessWeek, 38-45, 5<sup>th</sup> Nov. 2007, typifies financial press articles.

<sup>27</sup> A significant exception is BBC participation in the PrestoSpace project. See <http://prestospace-sam.ssl.co.uk/>, viewed 12-Nov-07, and also R. Wright, *Digital preservation of audio, video and film*, VINE 34(2), 71, 2004.

<sup>28</sup> The number of Web pages is between 500,000,000 and 5,000,000,000. The number of Internet users is more than 1,000,000,000. See <http://www.internetworldstats.com/stats.htm>, viewed 15-Nov-07, and pages it links.

<sup>29</sup> G. Bell and J. Gemmell, *A Digital Life*, Scientific American 296(3), 58-65, March 2007.

<sup>30</sup> In Cox, loc. cit. From D. Crystal, *Language death*. Cambridge U.P., 2000.

<sup>31</sup> In Cox, loc. cit. From B. Shneiderman, *Leonardo’s laptop: Human needs and the new computing technologies*, MIT Press, 2002.

<sup>32</sup> For instance, a recent addition underway is Semantic Web technology. See L. Feigenbaum et al., *The Semantic Web in Action*, Scientific American 297(6), 90-97, 2007. N. Spivak, *Minding the Planet: The Meaning and Future of the Semantic Web* (<http://lifeboat.com/ex/minding.the.planet>, viewed 29-Nov-07) attempts a detailed guide for non-technical people.

information creation greatly exceeds archival institutions' ability to select for preservation. This suggests that no attempt they make to archive a significant fraction of newly created digital content can succeed. Archivists need to consider strategies for selection only after sufficient time has elapsed so that well-informed communities can choose what is most valuable.<sup>33</sup>

Will delaying preservation for 20 years or longer from information creation risk not losing much that is valuable? Of course it will, but the risks of the alternative are greater. Since it is unlikely that society will manage to save everything, early preservation selection will lose much that is valuable and confound what is saved with content later considered uninteresting. The anticipated information glut might make finding good material more difficult than it already is.

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.<sup>34</sup> However, the DPC community does not seem to have adopted an I/T customer role, which might have contributed to NDIIPP's difficulty in engaging I/T providers effectively.<sup>3</sup> There is little evidence that the DPC has systematically considered commercial and open sources to devise a software acquisition strategy.

### ***Implications***

Preservation costs will seldom be recoverable from intended beneficiaries. The only practical tactics are to seek philanthropic support (mostly from governments) and to make the costs acceptable to information producers. No comment is needed on the philanthropic alternative. For the second alternative, we need to find ways of minimizing the effort needed, of distributing the load as widely as possible among stakeholders, and of embedding the needed work into existing document processing so that its incremental cost is more or less hidden, and no more than a small fraction of document preparation costs.

The DPC seems not to have the skills for creating tools or effectively managing software development to make preservation easy. Librarians and archivists cannot themselves accomplish preservation they call for. Instead, they must seek effective partnerships, particularly with software engineers, beyond any that they have actively encouraged.

Neither the private sector nor ordinary citizens have shown enough LDP interest to act in what some preservationists believe to be in their own interests. Functional and cosmetic software improvements for day-to-day work seem to have higher priority. So does satisfying regulatory requirements.<sup>26</sup> Available tools for creating preservation metadata have not become widely used. It is therefore not clear that preservation tools would be used if doing so costs users additional time to learn how and to edit documents. Someone would have to create and "sell" tools that preserve digital content as side effects of ordinary users' document editing and management.

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.

---

<sup>33</sup> For instance, the Library of Congress received Leonard Bernstein's personal correspondence in 1993 only after he died in 1990.

<sup>34</sup> [OpenLogic Exchange](http://www.openlogic.com/olex/) is a Web service to help enterprises find, download, use, manage and support enterprise-ready open source packages. See <http://www.openlogic.com/olex/>, viewed 25-Nov-07.

## 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<sup>35</sup> 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<sup>3</sup>

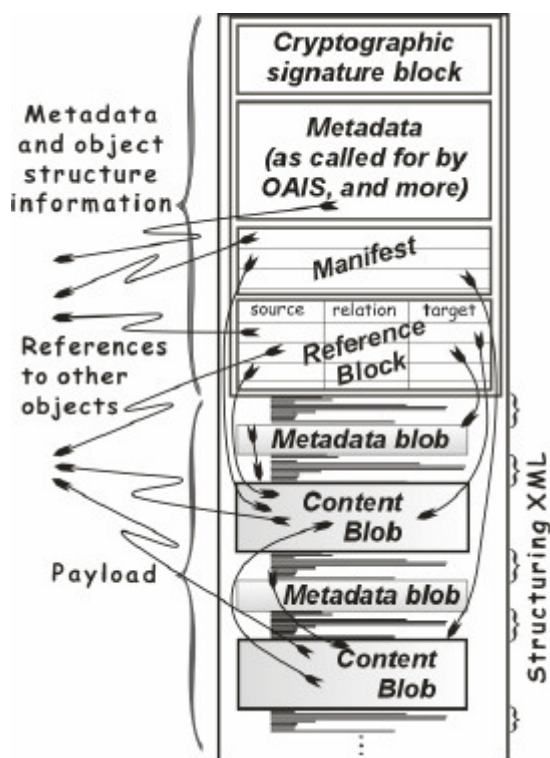


Figure 1: Trustworthy Digital Object (TDO) structure and content

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

An alternative for repository-centric approaches is methodology based on preservation-ready “trustworthy digital objects” (TDOs).<sup>36</sup> Its schema are 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.

A convenient starting point for talking about TDO methodology is its structure for interchangeable digital objects (Figure 1). Each TDO embeds its own metadata as verifiable evidence, and has an RDF-compatible<sup>37</sup> relationship block for both internal cross-references and external references. A TDO includes any number of content blobs.<sup>38</sup> 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 (as do the objects of a relational database). Each relationship is a triple: [source, referenceType, target]. Its source and target name links to objects, into objects, or to extents within objects. Its referenceType can itself be an object name, but is more likely to be the name of a relationship for which supporting software exists.

Any intellectual work, considered together with the contextual information it links, is in fact a collection. TDO structure can represent any kind of document, any collection, or any library catalog. A TDO should be constructed to include whatever links its author deems important for correct interpretation of its content. Any TDO structure is potentially unbounded. Its users must decide which linked information subset interests them.

Each TDO is cryptographically signed and sealed, as are its embedded references to other TDOs. The signature of a reference is a copy of the signature of the referenced (linked) object. Networked repositories storing TDOs provide scaling and evidentiary infrastructure.<sup>39</sup>

<sup>35</sup> A. McHugh et al., *Digital Repository Audit Method Based on Risk Assessment*, 2007, <http://www.repositoryaudit.eu>, viewed 22-Nov-07.

<sup>36</sup> Detail is available in a recent book: loc. cit. footnote 11.

<sup>37</sup> RDF Core Working Group, *Resource Description Framework*, 2004, <http://www.w3.org/RDF/>, viewed 24-Nov-07.

<sup>38</sup> 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.

<sup>39</sup> G. Caronni, *Walking the Web of Trust*, Proc. 9th Workshop on Enabling Technologies. 2000.

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.

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.<sup>40</sup> ORE specifies link semantics within a digital object and between digital objects—semantics sufficient to be a specification basis for software modules enabling information sharing among repositories and end users. Thus, it can be viewed as extending TDO by specific relationships, which it calls “hasLineage”, “hasEntity”, “hasIdentifier”, “hasProviderInfo”, “hasDatastream”, and so on.<sup>41</sup> This kind of extension is essential for informing supporting software about suitable actions, such as graph-following for retrieval, compound object construction, and so on.

ORE does not preclude other relationship types. Nor does it seem to include cryptographic signing and sealing for authenticity evidence. It would permit such extensions.

XDFU defines a competing interchange structure.<sup>42</sup>

Each Figure 1 content blob is represented either with an ISO encoding deemed durable or with a generic method using a simple virtual machine that Lorie called a Universal Virtual Computer (UVC).<sup>18</sup> This method, based on the Church-Turing thesis,<sup>43</sup> exploits three facts: (1) that any feasible computation can be accomplished with a Turing machine; (2) that any information whatsoever can be represented as Turing machine output; and (3) that a complete specification and an implementation of such a machine are surprisingly short and can be tested for correctness.

Design or implementation errors of UVC-based encoding can be detected today instead of being hidden hazards that might be discovered only after it is too late for reliable corrections. The effect is that any information can be represented for later 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.<sup>44</sup>

## 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<sup>45</sup>

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

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.

<sup>40</sup> H. Van de Sompel, C. Lagoze, et al., *An Interoperable Fabric for Scholarly Value Chains*, D-Lib Magazine 12(10), 2006; Open Archives Initiative, *Compound Information Objects: the OAI-ORE Perspective*, May 2007, <http://www.openarchives.org/ore/documents/CompoundObjects-200705.html>, viewed 15-Nov-07.

<sup>41</sup> See [Figure 3](#) of Van de Sompel and Lagoze (loc. cit.)

<sup>42</sup> XFDU objects have structure similar to TDOs. See CCDS, *XML Formatted Data Unit (XFDU) Structure and Construction Rules*, Sept. 2004, <http://sindbad.gsfc.nasa.gov/xfdupdfdocs/iprwbv2a.pdf>, viewed 25-Nov-07. Compatibility is likely to be incomplete, since the schemes were defined independently. Since the TDO definition did not include syntax, XFDU could be used for the metadata and structuring XML of Figure 1.

<sup>43</sup> B.J. Copeland, *The Church-Turing Thesis*, Stanford Encyclopedia of Philosophy, 2002, <http://plato.stanford.edu/entries/church-turing/>, viewed 1-Nov-07.

<sup>44</sup> Lorie, loc. cit. footnote 18. Also Gladney, loc. cit. footnote 11, chapter 12.

<sup>45</sup> N. Rescher, *The Rise and Fall of Analytic Philosophy in Minding Matter and Other Essays in Philosophical Inquiry*, Rowman & Littlefield, 2001.

Given that our objective is knowledge communication, the appropriate sources of fundamentals are epistemology and philosophy of language.<sup>46</sup>

Philosophers distinguish critically 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 technical 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 preservation criteria."<sup>47</sup> The problem with this assertion is not that it is incorrect. Even the most careful definition is likely to provoke criticism. Its problem is that its author seems to use 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."<sup>48</sup> 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":<sup>49</sup>

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.

<sup>46</sup> 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 "more honor'd in the breach than the observance." (Shakespeare, *Hamlet*, Act 1, Scene 4)

<sup>47</sup> R. Verdegem, *Back to the future: Dioscuri, the modular emulator for digital preservation*, 2007, <http://www.kb.nl/hrd/congressen/toolstrends/presentations/Verdegem.pdf>, viewed 20-Nov-07.

<sup>48</sup> 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>, viewed 22-Nov-07.

<sup>49</sup> H.M. Gladney and J.L. Bennett, *What Do We Mean by Authentic?* D-Lib Magazine 9(7), July 2003.

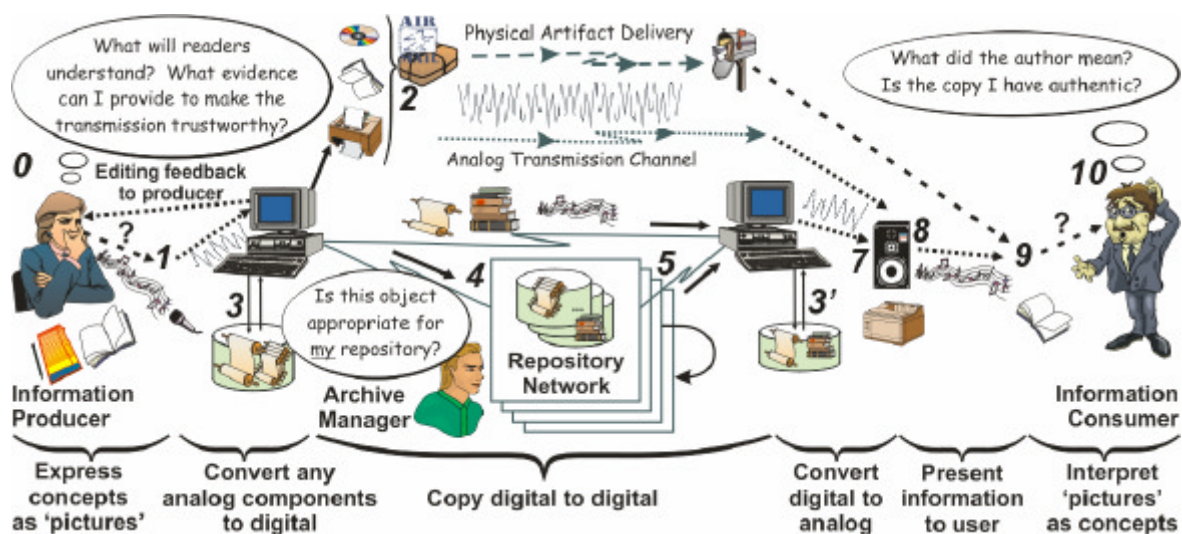


Figure 2: Information flow for current and preserved documents

Ross and his source<sup>50</sup> suggest that, "As every instantiation is a 'performance' representing a range of functions and behaviours, we need ways to assess the verisimilitude of each subsequent performance to the initial one and clear definitions of 'acceptable variance'." This might sound reasonable. However, if it is implemented by unilateral archivists' actions, it risks subtle information degradation.

Every document representation confounds essential information with accidental information. It might also lose essential information. Social convention identifies the Figure 2 information provider as the primary authority for the distinction between what is essential and what is accidental. (An editor or archivist who annotates an object with metadata would be the information creator for that metadata.) For instance, in a recording of recited poetry, the speaker's voice pitch is likely to be regarded as accidental information. The secondary authority is the Figure 2 information consumer, whose notion of essential is governed by his objectives. For instance, marginal notes in an old book 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 detail that this author considers essential to conveying what he intends. 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.

The accidental/essential circumstance contributes to uncertainty that an information consumer has correctly understood what an information producer meant. Human dialog permits some reduction of this uncertainty, but there is no similar opportunity for archival documents. Preservationists will want not to exacerbate such difficulty, which they can ensure only by avoiding transformations whenever possible.<sup>51</sup> Combating file format obsolescence by repeated transformative migration<sup>52</sup> is a bad idea not only because archivists might make mistakes. Even

<sup>50</sup> National Library of Australia, Guidelines for the Preservation of Digital Heritage, UNESCO, 2003, <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>, viewed 25-Nov-07.

<sup>51</sup> They might also show transformed versions to authors to collaborate on corrections. However, doing this will seldom be practical.

if they accomplish precisely what they intend, they are likely to degrade the balance between essential and accidental information.

Discussions of the PDF format and its variants for preservation are about specific cases of this problem.<sup>53</sup> It affects Lorie's UVC method.<sup>18</sup> It is particularly bad in Rothenberg's scheme of emulating computer hardware,<sup>54</sup> because most of the architecture of the computer on which a document was created will be mostly irrelevant to what the document is intended to convey. We should push representation decisions to be close to original authors and to editors who identify themselves to readers, to the extent that doing so is feasible.

To some extent, information that is accidental 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 very 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 new digital objects than they do of old 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 and essential expertise. 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 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 DP 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 would be appraised in about a year for demonstrable progress towards customer service?<sup>55</sup>

## 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 mechanism or process until it is partitioned from the other mechanisms sufficiently so that it can be treated with minimal, 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

---

<sup>52</sup> 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.

<sup>53</sup> PDF/A was developed 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>, viewed 30-Nov-07) 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; but once you have decided to use a PDF-based approach, PDF/A defines an archival profile of PDF that is *more* amenable to long-term preservation."

<sup>54</sup> J. Rothenberg, *Ensuring the Longevity of Digital Documents*, *Scientific American* 272(1), 42-47, 1995.

<sup>55</sup> In this, I am drawing on many years experience with technology transfer from the IBM Research Division to product development in other IBM units.

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 government 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 convenient 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 preservation technology is unaffordable except by modest modifications of what is already created 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.<sup>56</sup> 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.<sup>57</sup> It would help the manager partition the overall task so that independent teams could cooperate.

## 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<sup>3</sup>

We anticipate a future with deployed machinery for preserving any digital content whatsoever and pleasing eventual recipients. Towards such a happy situation, the current section seeks to sketch soundly based design for tools and infrastructure. For that, we choose a design capable of handling every data format<sup>58</sup> and sufficient for information that has high risk for fraudulent or accidental falsification.

<sup>56</sup> Applicable techniques are called "knowledge management." See J. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole, 2000.

<sup>57</sup> 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/>, viewed 25-Nov-07, and J. D. Novak & A. J. Cañas, *The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool*, *Information Visualization Journal* 5, 175-184, Jan. 2006..

That such tools have entered the mainstream is suggested by Brian Hindo, *Inside Innovation: Software that maps*, *BusinessWeek* 19-21, Nov. 26, 2007.

<sup>58</sup> A university generates almost every well-known information type, and also many novel types within its research programs. An institutional archive must accommodate every kind of information generated by its constituency, and must be readily extensible to information representations that become important in the future, even though

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 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 errors and inappropriate changes introduced by third parties;
3. Decide whether the information received is sufficiently trustworthy for his application;
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 being encumbered by avoidable technical complexity.

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 preservation-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.

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 preservation;
9. Sharing content and metadata without adjustments requiring human judgment; and
10. Sharing preservation effort with their clients to avoid burdens beyond their own resources.

## 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<sup>59</sup>

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.

This will be feasible only if creating preservation-ready information is an inexpensive addition to editing already required. Preservation tools must be packaged within information producers' tools. Since producers already want their output to be valued, it should be possible to persuade them to do additional work if this is inexpensive and easy. For instance, prestigious repositories might limit what they accept for indexing and distribute to preservation-ready content.

---

their specifics are unknown today. In particular, technical research generates computer programs among which some surely should be preserved.

<sup>59</sup> C. Lupovici, *Archiving the Web: the mass preservation challenge*, Tools and Trends: International Conference on Digital Preservation, Nov. 2007, <http://www.kb.nl/hrd/congressen/toolstrends/presentations/Lupovici.pdf>, viewed 26-Nov-07.

In some unrealistically ideal world, preservation 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, depending on human purposes and context evaluation that cannot be automated. 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. Such analysis is a fundamental basis for semi-automatic software to minimize human work needed.

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.

### Preparing Information for Preservation and Examining Preserved Information

Preserving any information corpus requires considerable metadata that is rarely included in today's WWW deliveries. Writers are typically the best informed participants about the missing facts. Only an information consumer can properly appraise potential damage to his affairs should information be inauthentic or incorrect. Such circumstances and scales already mentioned suggest shifting preparation for archiving from repository employees to information providers 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 editor 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 interactive graphic front ends that might look like Figure 1.

The authenticity definition of **§Logical Analysis**, used together with Figure 2, can guide an engineer towards distinguishing questions that can be objectively answered from those with subjective elements. Such distinctions can also help engineers design effective semi-automatic procedures that prompt information producers for those answers that only human beings can provide (together with self-identification that conveys authority). Similar thinking can assist design to help information consumers decide whether or not information provided is trustworthy.

When we first put forward the TDO scheme about four years ago, the formal specification deliberately 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 developed for metadata extraction include the NLNZ Metadata Extraction Tool<sup>60</sup> and the

---

<sup>60</sup> National Library of New Zealand, *Metadata Extraction Tool*, 2003, <http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool/>, viewed 27-Nov-07.

PLANETS technical metadata extractor.<sup>61</sup> At least three contenders for digital object packaging have appeared.<sup>62</sup> Such tools could be embedded as document editor extensions.

Each Figure 1 content blob is represented either in some ISO format or with a UVC-based method.<sup>18</sup> Handling relatively simple file types might avoid the seeming complexity of UVC methodology, depending instead on long-term durability of standard formats (ISO formats). The scope of this approach is unclear, because some formats are surprisingly complex, with many variants.<sup>63</sup> Whether a software developer uses ISO- or UVC-methodology, he must and 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 that will be strongly influenced by how comfortable its holder is with computer science.

The software engineering literature adds many document-handling suggestions every month. Some potentially help with preservation. For instance, representing complex structures might be eased by a general format described in a prepublication paper.<sup>64</sup>

## Digital Archiving Services

Figure 3 corresponds to the much-used OAIS Functional Entities depiction,<sup>65</sup> but is drawn to emphasize critical software 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.”<sup>66</sup> The current section sketches a response.

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. The privilege differences for access and alteration of repository contents 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 the functionality might be partitioned among multiple computers. Instead it suggests functional components and layering. 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 the specific rules for a replication site are likely to be

---

<sup>61</sup> M. Thaller, *Characterizing with a Goal in Mind: The XCL approach*, 2007, <http://www.kb.nl/hrd/congressen/toolstrends/presentations/Thaller.pdf>, viewed 27-Nov-07.

<sup>62</sup> XFDU, loc. cit., footnote 42.

XAM (eXtensible Access Method), <http://www.snia.org/forums/xam/technology/specs/>, viewed 30-Nov-07.

WARC (Web ARChive format), <http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>, viewed 30-Nov-07.

<sup>63</sup> 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>, viewed 24-Nov-07.

<sup>64</sup> H. Ishikawa, *Representation and Measure of Structural Information*, <http://arxiv.org/abs/0711.4508>, viewed 29-Nov-07. My attention was drawn to this example by an e-mail from Cornell U's arXiv.org listserv.

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

<sup>66</sup> 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>, viewed 3-Dec-07.

different from those for a human user, but this will be similar to how the rules differ for different 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 digital libraries. In fact, this boundary corresponds to a software interface that became the subject of a standardization effort in 2005—the *Content Repository API for Java* called “JSR 170” and its “JSR 283” refinement.<sup>67</sup> These standards make storage subsystems interchangeable. Repositories will choose among offerings to accommodate different hardware systems, different storage capacities, and different data traffic.

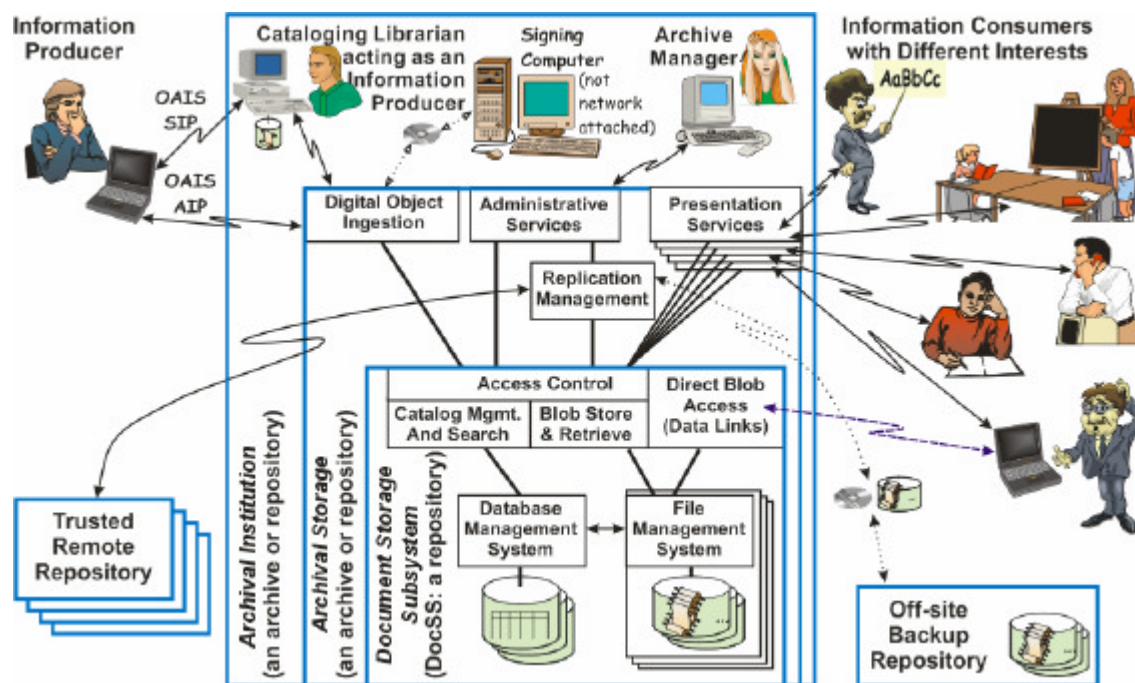


Figure 3: Nested repository structure and interfaces for human roles

The archival storage layer includes only very little code implementing preservation functionality. The only exception might be the ingestion component, which might include logic for deciding whether proposed ingestions conform to institutional criteria, possibly tagging stored objects to indicate different acceptable quality levels. A consequence is that most preservation software can be written for personal computers, a happy outcome!

Software supporting librarians that participate in packaging information for preservation can be more or less the same as that for information producers who are not repository employees.

The low-level part of LDP, saving bit-string patterns, is more advanced than the aspects already discussed. LOCKSS replication management<sup>68</sup> enforces rules above database and file system interfaces. Its deployed version probably needs access control enhancements to permit

<sup>67</sup> See <http://jcp.org/en/jsr/detail?id=170>. A Web search for “JSR 170” will yield over 300,000 links, including 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 adheres to the same general goal of JSR 170—to produce a content repository API that provides an implementation-independent way to access content bi-directionally on a granular level.” See <http://jcp.org/en/jsr/detail?id=283>, viewed 23-Nov-07.

<sup>68</sup> See <http://lockss.stanford.edu>, viewed 30-Nov-07.

different privileges for different remote repositories. At a still lower level, scaling and automation for vast collections will hopefully be addressed by an industry workgroup.<sup>69</sup>

## Discussion

For more than ten years, the digital preservation community has searched for practical methods of ensuring the longevity, integrity, authenticity, and interpretability of digitally recorded information. Its search space has been surprisingly narrow—methods to be used for human and machine processes within archival institution perimeters, and methods of persuading clients to trust that archival deliveries conform to well-known quality standards. Ten years work without a solution should persuade participants to broaden their search.

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. The challenges are exacerbated by the fact that essential skills are mostly located outside the DPC.

DPC literature starts with a premise that it never questions. It assumes that LDP objectives can be achieved by adapting established archival methodology with only modest changes. The model archive has been dedicated to preserving information stored on paper and other material media. However, an archive's purpose is not *per se*, but merely a means to another objective—to preserve and deliver documents, ensuring that deliveries are trustworthy.

No plausible investment in repository institutions would change the key determinants. 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.

It is somewhat ironic that how to achieve preservation is immediately 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 is done by binding a document's portions firmly to each other and sealing the whole with a signature that cannot be forged. Of course, how to do this is different for digital documents than it was for paper documents.

The technically most difficult aspect is ensuring durable intelligibility of the saved content without knowing much about future computers. That people expect better representation fidelity and more convenience than ever before exacerbates the problem, as does rapid change of natural and technical languages. An 18<sup>th</sup>-century scholar would have depended on his expertise developed by years of study for his painstaking interpretation of early manuscripts. Tomorrow's students will expect their corresponding work to be rapid, convenient, and possible without years of specialized study. The trick is mechanical assistance to render content in vernacular of the time. How to accomplish this is known. The current problem is less that the method is difficult than that its principles are unfamiliar to almost everyone interested.

The dominant 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 methods yet to be devised.

---

<sup>69</sup> 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/](http://www.snia.org/forums/dmf/programs/ltacsi/100_year/), viewed 30-Nov-07.

With the shift from a repository-centric approach to the TDO approach, the scaling problem emphasized by Lupovici<sup>59</sup> vanishes without further effort!

Nothing above is intended to suggest marginalization of repository institutions. However, they will need to shift to roles for which they are better suited than anyone else. Although the precise nature of such roles is still a subject for investigation 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 based on cryptographic signatures.

## Next Steps

Anyone wanting to advance the state of the art 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.

A remedy for such problems is to precede significant investment by a broad analysis of the landscape to identify the most promising projects and to help partition what's needed into tasks suitable for almost independent teams.

Someone should systematically examine a large fraction of the LDP literature to map what its authors teach. One effort would collect and organize distinct needs into a formal engineering requirements document. With each "line item" summarized in 2-5 lines, this document might have 100 pages or more. A second effort would create a pictorial map of candidate technology and software designs, labeled with pointers to detailed information. These information resources would need to be iteratively refined as long as LDP development continues.

Software components exist for handling nearly every technical requirement and are mostly inexpensive. Such tools are created by communities whose work has not been examined by the DPC. Adapting and integrating them would be business-as-usual software implementation.

## 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."<sup>3</sup> Building on a published conceptual solution, we have described an engineering program that would handle every explicitly identified technical requirement.

Our method is comprehensive, addressing all extant and future data formats and having sufficient security to ensure testable integrity and authenticity, even for information that tempts fraudulent modification. It scales by supporting information producers' preservation packaging, and 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."

The TDO methodology described is general, but does not pretend to be optimal. Instead, it allows optimization by specialized modules for popular data formats and by security shortcuts.

Most of the software needed for long-term digital preservation is available, but has not been tailored and packaged for digital preservation community convenience.

A significant remaining challenge is to hide arcana from end users. Another is to persuade "buy in" by archival institutions, which could create a trust environment by managing hierarchical

cryptographic digital signatures. This would be in addition to their preserving bit-string integrity and accessibility by digital object replication and providing information-finding services.

Our description is sufficient guidance so that any senior software engineer could direct the next preservation 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.

**Introduction**..... 1  
 Summary of What Follows ..... 2  
 Scope Limitations..... 3  
**Literature Analysis** ..... 4  
 A Summary of Challenges..... 4  
 Traditional Archival Science..... 6  
**Economic Analysis**..... 7  
 DPC and Stakeholder Communities..... 8  
 Implications..... 9  
**Trustworthy Digital Objects** ..... 10  
**Logical Analysis** ..... 11  
**Engineering Analysis** ..... 14  
 Design Process ..... 14  
 Technical Objectives..... 15  
 Structuring a Solution ..... 16  
     Preparing Information for Preservation and Examining Preserved Information ..... 17  
     Digital Archiving Services ..... 18  
**Discussion** ..... 20  
 Next Steps ..... 21  
**Conclusions** ..... 21

Figure 1: Trustworthy Digital Object (TDO) structure and content ..... 10  
 Figure 2: Information flow for current and preserved documents ..... 13  
 Figure 3: Nested repository structure and interfaces for human roles ..... 19