Editor's Note

In the summer of 2024, Clifford Lynch announced his retirement as executive director of the Coalition for Networked Information (CNI) after 28 years at its helm. CNI quietly launched a project to create this Festschrift to document and honor his legacy. Authors began contributing articles in early 2025, with a planned publication date of July 2025. Since the final membership meeting of Cliff's tenure was April 7–8 in Milwaukee, the plan was to surprise him, surrounded by colleagues and friends, with a presentation of the table of contents of this special issue. However, just two weeks prior to the meeting, Cliff's health worsened; he was told about the Festschrift and received project details and articles. Though unable to attend in person, he participated in the CNI membership meeting via Zoom and also virtually joined his retirement reception, which included readings of excerpts from each article in this volume. Sadly, on April 10, 2025, Clifford Lynch passed away. Festschrift contributors wrote their articles prior to his passing, and we have chosen not to alter their original language.

Lots of Cliff Keeps Stuff Safe

Victoria Reich and David S. H. Rosenthal

abstract: A long time ago in a Web far, far away, it is a period of civil war between two conceptions of how digital information could be preserved for posterity.[1] On one side is the mighty Empire, concerned with the theoretical threat of format obsolescence. On the other are the Rebels, devoted to the practical problem of collecting the bits and ensuring that they survive. Among the rebels are the Internet Archive and the LOCKSS Program. This is the story of how the rebels won, thanks in no small part to Cliff Lynch's sustained focus on the big picture.

Background

It all started just 30 years ago. In January 1995, the idea that the long-term survival of digital information was a significant problem was popularized by Jeff Rothenberg's *Scientific American* article "Ensuring the Longevity of Digital Documents."[2] Rothenberg's concept of a "digital document" was of things like Microsoft Word files on a compact disc—that is, individual objects encoded in a format private to a particular application. His concern was with *format obsolescence*: the idea that the rapid evolution of these applications would, over time, make it impossible to access content in an obsolete format.

Rothenberg was concerned with interpreting the bits; he essentially assumed that the bits would survive. Given the bits, he identified two possible techniques for accessing the content: (1) *format migration*, translating the content into a less obsolete format to be accessed by a different application, and (2) *emulation*, using a software implementation of the original computer's hardware to access the content using the same application. Emulation was a well-established technique, dating from the early days of IBM computers.

The Web

Five months later, an event signaled that Rothenberg's concerns had been overtaken by events. Stanford University pioneered the transition of academic publishing from paper to the World Wide Web when the HighWire Press team (of which Victoria was

portal: Libraries and the Academy, Vol. 25, No. 3 Supplement (2025), pp. 107–114.

a member) put the *Journal of Biological Chemistry* on the Web. By then it was clear that, going forward, the important information would be encoded in formats such as HTML and PDF. Because each format with which Rothenberg was concerned was defined by a single application, it could evolve quickly. But formats were open standards, implemented in multiple applications. In effect, they were network protocols.

The deployment of IPv6 (Internet Protocol version 6), introduced in December 1995, shows that network protocols are extraordinarily difficult to evolve because of the need for timely updates to many independent implementations. Format obsolescence implies backward incompatibility; this situation is close to impossible in network protocols because it would partition the network, splitting it into disconnected components. As David discussed in 2012's "Formats through Time," the first two decades of the World Wide Web showed that its formats essentially do not go obsolete.[3]

The rapid evolution of Rothenberg's "digital documents" had effectively stopped, because they were no longer being created and distributed in that way. Going forward, there would be a legacy of a static set of documents in these formats. Libraries and archives would need tools for managing those they acquired. Eventually emulation, the technique Rothenberg favored, would provide those tools.[4] But by then it turned out that, unless information was on the World Wide Web, almost no one cared about it.

The Integrity of Digital Information

Thus, the problem for digital preservation was the survival of the bits, not the retention of their format, aggravated by the vast scale of the content to be preserved. In May of the following year, Brewster Kahle established the Internet Archive to address the evanescence of web pages.[5] This impermanence comes in two forms: *link rot*, when links no longer resolve, and *content drift*, when they resolve to different content.

As Cliff did in many fields, he focused on the big picture. He understood the importance to digital preservation of simply collecting the content and ensuring its integrity.

This is where Cliff Lynch enters the story. As Cliff did in many fields, he focused on the big picture. He understood the importance to digital preservation of simply collecting the content and ensuring its integrity. Already in his 1994 article "The Integrity of Digital Information" he had written:

> A system of information distribution that preserves integrity should also provide the user with a reasonable expectation of correct attribution and source of works. Even if deliberate attempts at fraud, misdirection, or covert revision may sometimes slip through the routine processes of the system these problems can be adjudicated by a formal challenge and examination system … The expectation should be that violations of integrity cannot be trivially accomplished.

He had noted that even in the print world this expectation was fading:

> We assume that print is difficult to alter, that print authorship and source attribution are relatively trustworthy, and that printed works are normally mass-produced in identical copies. In fact, current technology trends undermine these assumptions. Printed publications are becoming increasingly tailored to very narrow audiences, and it has become easy to imitate the format of well-known and professionally presented publications.[6]

Cliff discussed how the survival of the bits could be confirmed using digital hashes, the potential for digital signatures to confirm authenticity, and why such signatures were not used in practice.

LOCKSS

In October 1998, we proposed to Michael Keller, Stanford's librarian, a decentralized system whereby libraries could cooperate to collect the academic journals to which they subscribed and preserve them against the three threats we saw: technological, economic, and legal. He gave us three instructions:

- Don't cost me any money.
- Don't get me into trouble.
- Do what you want.

Thus was born the LOCKSS (Lots of Copies Keep Stuff Safe) Program. The prototype was funded by two small grants, the first from Michael Lesk at the National Science Foundation (NSF) and then by Donald Waters at the Andrew W. Mellon Foundation. Both of them, like Cliff, understood the importance of assuring the survival of, and access to, the bits.[7] Development of the first production system was mostly funded by a significant grant from the NSF and by Sun Microsystems. We did not cost Keller any money, in fact, the reverse, considering Stanford's overhead on grants.

The LOCKSS system, like the Internet Archive, was a system for ensuring the survival of, and access to, the bits in their original format. This was a problem. Somehow, despite Rothenberg's advocacy of emulation, the conventional wisdom in the digital preservation community rapidly became that digital preservation should defend against format obsolescence by using format migration based upon collecting preservation metadata.

Actually, the sine qua non of digital preservation is ensuring that the bits survive. Neither Kahle nor we saw any return on investing in preservation metadata or format migration. We both saw scaling up to capture more than a tiny fraction of the at-risk content as the goal. Future events showed we were right. At the time, however, the digital preservation community viewed LOCKSS with great skepticism, as "not real digital preservation."

Paper Library Analogy

In a 1994 paper, Cliff had described how the paper world's equivalent of ensuring the survival of the bits might work. It could be summarized as "Lots of Copies Keep Stuff Safe":

In a 1994 paper, Cliff had described how the paper world's equivalent of ensuring the survival of the bits might work.

> When something is published in print, legitimate copies ... are widely distributed to various organizations, such as libraries, which maintain them as public record. These copies bear a publication date, and the publisher essentially authenticates the claims of authorship ... By examining this record, control of which is widely distributed ... it is possible, even years after publication, to determine who published a given work and when it was published. It is very hard to revise the published record, since this involves all of the copies and somehow altering or destroying them.[8]

Compare this with how we summarized libraries' role in our first major paper on LOCKSS, "Permanent Web Publishing." It recommended, "Acquire lots of copies. Scatter them around the world so that it is easy to find some of them and hard to find all of them. Lend or copy your copies when other librarians need them."

From a systems engineering viewpoint, we wrote:

> Libraries' circulating collections form a model fault-tolerant distributed system. It is highly replicated and exploits this to deliver a service that is far more reliable than any individual component. There is no single point of failure, no central control to be subverted. There is a low degree of policy coherence between the replicas, and thus low systemic risk. The desired behavior of the system as a whole emerges as the participants take actions in their own local interests and cooperate in ad-hoc, informal ways with other participants.

If librarians are to have confidence in an electronic system, it will help if the system works in a familiar way.[9]

Threats

Cliff's focus on the big picture meant he also understood that economic and legal threats were at least as significant as technological ones. For example, in 1996's "Integrity Issues in Electronic Publishing," he wrote:

> In the networked information environment, the act of publication is ill defined, as is the responsibility for retaining and providing long-term access to various "published" versions of a work. Because of the legal framework under which electronic information is typically distributed, matters are much worse than they are generally perceived to be. Even if the act of publication is defined and the responsibility for the retention of materials is clarified, the integrity of the record of published works is critically compromised by the legal constraints that typically accompany the dissemination of information in electronic formats.[10]

Cliff's focus on the big picture meant he also understood that economic and legal threats were at least as significant as technological ones.

He discussed some electronic journal pilots in a 1996 talk, writing that one key question was "how acceptable transactional pricing systems will be to end users or to producers, suppliers, and rights holders. Will such models cause streams of income and expenditures to become unworkably erratic?"[11]

Now there are two lawsuits from the copyright cartels aimed at destroying the Internet Archive, and it is easy to understand that the most critical threats to preserved content are legal. A quarter century ago, this was less obvious. But even then, facing the oligopoly academic publishers, it was obvious to us that LOCKSS had to be designed around the copyright law.[12]

Cliff continued to remind the library community of the economic and legal threats and of the broader issues impeding preservation of our digital heritage. Early examples include 1999's "Experiential Documents and the Technologies of Remembrance," where he wrote:

> The retention, reuse, management, and control of this new cornucopia of recorded experience and synthesized content in the digital environment will, I expect, become a matter of great controversy. This will include, but not be limited to, privacy, accountability and intellectual property rights in their broadest senses. And these materials will hopefully become an essential and growing part of our library and archival collections in the 21st century—particularly as we sort through these controversies.[13]

Cliff continued to remind the library community of the economic and legal threats and of the broader issues impeding preservation of our digital heritage.

In 1999's "On the Threshold of Discontinuity," he wrote:

> It is unclear how to finance archiving and preservation of these materials. Their volume is no longer driven by acquisitions budgets or by the scholarly publishing system, but by activities that may take place largely beyond the control of the library. And, of course, costs are open ended and unpredictable for digital preservation, unlike the costs associated with preserving modern printed materials (on acid-free paper).[14]

In 2001's "When Documents Deceive," Cliff explained, "Digital documents in a distributed environment may not behave consistently; because they are presented both to people who want to view them and software systems that want to index them by computer programs, they can be changed, perhaps radically, for each presentation. Each presentation can be tailored for a specific recipient."[15]

Cliff's 2003 article "The Coming Crisis in Preserving Our Digital Cultural Heritage" said, "Preservation of digital materials is a continuous, active process (requiring steady funding), rather than a practice of benignly neglecting artifacts stored in a hospitable environment, perhaps punctuated by interventions every few decades for repairs." The article added, "It is probably not an exaggeration to say that the most fundamental problem facing cultural heritage institutions is the ability to obtain digital materials together with sufficient legal rights to be able to preserve these materials and make them available to the public over the long term. Without explicit and affirmative permissions from the rights-holders, this is likely to be impossible." The article explained,

> What is threatening us today is not an abuse of centralized power, but rather a low-key, haphazard deterioration of the intellectual and cultural record that is driven primarily by economic motivations and the largely unintended and unforeseen consequences of new intellectual property laws that were enacted at the behest of powerful commercial interests and in the context of new and rapidly evolving technologies.[16]

The "Standard Model"

The LOCKSS team repeatedly made the case that preserving web content was a different problem from preserving Rothenberg's digital documents, and thus that applying the entire apparatus of "preservation metadata"—PREMIS (Preservation Metadata Implementation Strategies), FITS (Flexible Image Transport System), JHOVE (JSTOR/Harvard Object Validation Environment), and format normalization to web content—was an ineffective waste of scarce resources.[17] Despite this, the drumbeat of criticism that LOCKSS was not "real digital preservation" continued.

After six years, the LOCKSS team lost patience and devoted the necessary effort to implement a capability they were sure would never be used in practice. The team implemented, demonstrated, and in 2005 published transparent, on-demand format migration of web content preserved in the LOCKSS network.[18] This was possible because the specification of the HTTP (hypertext transfer protocol) that underlies the World Wide Web supports the format metadata needed to render web content. If it lacked such metadata, web browsers would not be possible.

Unsurprisingly, this demonstration failed to silence the proponents of the "standard model of digital preservation." So five more years later, in 2010, David published "Format Obsolescence: Assessing the Threat and the Defenses," a detailed exposition and critique of the standard model's components. Those were:

- Before obsolescence occurs, a digital format registry collects information about the target format, including a description of how content can be identified as being in the target format, and a specification of the target format from which a renderer can be created.
- Based on this information, format identification and verification tools are enhanced to allow them to extract format metadata from content in the target format, including the use of the format and the extent to which the content adheres to the format specification. This metadata is preserved with the content.
- The format registry regularly scans the computing environment to determine whether the formats it registers are obsolescent, and issues notifications.
- Upon receiving these notifications, preservation systems review their format metadata to determine whether they hold content in an obsolescent format.
- If they do, they commission an implementer to retrieve the relevant format specification from the format registry and use it to create a converter from the now-obsolescent target format to some less doomed format.
- The preservation systems then use this converter and their format metadata to convert the preserved content into the less doomed format.[19]

The critique observed that creating a format specification for a proprietary format and then implementing a renderer from it was almost impossible, that the existence of open-source renderers made doing so redundant, that most HTML on the World Wide Web failed validation (a consequence of Postel's Law, which recommends that when creating software or systems, the creators adhere strictly to specifications and standards in what they produce), that there were no examples of widely used formats going obsolete, and that Microsoft's small step in that direction in 2008 met with universal disdain and was abandoned.[20] It also noted that the standard model is based on format migration, a technique of which Rothenberg's article disapproves: format migration "suffers from a fatal flaw … Shifts of this kind make it difficult or impossible to translate old documents into new standard forms."[21]

Emerald Publishing awarded David's critique the 2011 Outstanding Paper Award for *Library Hi Tech*, but that honor failed to silence the standard model's proponents. Although we no longer follow the digital preservation literature closely, it is our impression that over the intervening 15 years, advocacy of the standard model has died down, thanks in no small part to Cliff's sustained focus on the big picture.

Victoria Reich and David S. H. Rosenthal cofounded the LOCKSS Program in 1998. Victoria stepped away from the executive director role in 2016 and retired in 2018. David retired from the chief scientist role in 2017.

Notes

1. Adapted from the opening crawl of "Star Wars: Episode IV—A New Hope," written and directed by George Lucas (1977; Los Angeles, 20th Century Studios), https://www.starwars.com/video/star-wars-episode-iv-a-new-hope-opening-crawl.
2. Jeff Rothenberg, "Ensuring the Longevity of Digital Documents," *Scientific American* 272, 1 (1995): 42–47, http://www.jstor.org/stable/24980135.
3. David Rosenthal, "Formats through Time," *DSHR's Blog*, October 9, 2012, https://blog.dshr.org/2012/10/formats-through-time.html.
4. David Rosenthal, "Follow-up to the Emulation Report," *DSHR's Blog*, November 10, 2015, https://blog.dshr.org/2015/11/follow-up-to-emulation-report.html.
5. David Rosenthal, "The Evanescent Web," *DSHR's Blog*, February 10, 2015, https://blog.dshr.org/2015/02/the-evanescent-web.html.
6. Clifford A. Lynch, "The Integrity of Digital Information: Mechanics and Definitional Issues," *Journal of the American Society for Information Science* 41, 10 (1994): 737–44, https://doi.org/10.1002/(SICI)1097-4571.
7. David Rosenthal, "A Tribute to Don Waters," *DSHR's Blog*, August 20, 2019, https://blog.dshr.org/2019/08/a-tribute-to-don-waters.html.
8. Lynch, "The Integrity of Digital Information."
9. David S. H. Rosenthal and Vicky Reich, "Permanent Web Publishing," in *FREENIX Track: 2000 USENIX Annual Technical Conference Proceedings* (San Diego, June 18–23, 2000), https://www.usenix.org/legacy/event/usenix2000/freenix/full_papers/rosenthal/rosenthal.pdf.
10. Clifford A. Lynch, "Integrity Issues in Electronic Publishing," chap. 8 in *Scholarly Publishing: The Electronic Frontier*, ed. Robin P. Peek and Gregory B. Newby (Cambridge, MA: MIT Press, 1996), 133–43, https://www.google.com/books/edition/Scholarly_Publishing/Uj-LIewe3TUC?hl=en&gbpv=0 OR https://mitpress.mit.edu/9780262661683/scholarly-publishing/).
11. Clifford A. Lynch and Carroll Davis, "Serials in the Networked Environment," *Serials Librarian* 28, 1–2 (1996): 115–20, https://doi.org/10.1300/J123v28n01_12.
12. Wikipedia, "Internet Archive."
13. Clifford Lynch, "Experiential Documents and the Technologies of Remembrance," in *i in the Sky: Visions of the Information Future*, ed. Alison Scammell (New York: Routledge, 2000), 140–46, https://doi.org/10.4324/9780203058725.
14. Clifford Lynch, "On the Threshold of Discontinuity: The New Genres of Scholarly Communication and the Role of the Research Library," in *Racing toward Tomorrow:*

Proceedings of the Ninth National Conference of the Association of College and Research Libraries, ed. Hugh A. Thompson (Chicago: Association of College and Research Libraries, 1999), 410–18, https://www.cni.org/wp-content/uploads/2014/07/clynch99.pdf.

15. Clifford A. Lynch, "When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web," *Journal of the American Society for Information Science and Technology* 52, 1 (2001): 12–17.
16. Clifford A. Lynch, "Chapter 18. The Coming Crisis in Preserving Our Digital Cultural Heritage," *Journal of Library Administration* 38, 3–4 (2003): 149–61, https://doi.org/10.1300/J111v38n03_04.
17. David Rosenthal, "PREMIS [Preservation Metadata Implementation Strategies] & LOCKSS," *DSHR's Blog*, March 28, 2014, https://blog.dshr.org/2014/03/premis-lockss.html.
18. David S. H. Rosenthal, Thomas Lipkis, Thomas S. Robertson, and Seth Morabito, "Transparent Format Migration of Preserved Web Content," *D-Lib Magazine* 11, 1 (January 2005), https://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html.
19. David Rosenthal, "Format Obsolescence: Assessing the Threat and the Defenses," *Library Hi Tech* 28 (2010): 195–210, https://www.abitare.org/papers/LibraryHighTech2010.pdf.
20. Mark Whitehorn, "'Draconian' Microsoft Promises to Make Office Work Again," *The Register*, January 5, 2008, https://www.theregister.com/2008/01/05/ms_office_sp3_woes/.
21. Rothenberg, "Ensuring the Longevity of Digital Documents," quoted in Rosenthal, "Format Obsolescence," 4.