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.

When Information Is Networked

Dan Cohen

abstract: Clifford Lynch is well-known for the depth of his knowledge in the domains of technology, libraries, and research. But perhaps his greatest contribution to the organization of knowledge and the pursuit of new methodologies has been his recognition of the transformative possibilities that emerge when these domains are connected through digital media. Once a collection or dataset becomes digital, it not only can, but likely will, be used in utterly new and unexpected ways through the power of networks. Corpora and data from loosely affiliated, decentralized institutions can be brought together in innovative combinations for diverse purposes. Digitized books may serve as surrogates for print books, but they may also be used as training materials for artificial intelligence. Computing cycles and remote instrumentation can be aggregated to create robust analytical tools that did not exist before. Imagining new combinations of networked information can help us forecast the future—and create it.

The Birth of Google Books

In the fall of 2005, Google had recently begun an audacious project that diverted the young company from its core business of copying and indexing web pages into a new venture of copying and indexing physical pages. That initiative would ultimately become Google Books. Millions of the books that slid into Google's scanners came from the shelves of major research libraries. Accessing an aggregated digitized library of this scale, from anywhere at any time, was a wildly transformative prospect for scholars and students, not to mention for a world of readers who lived far from these institutions.

But the project was also hugely fraught. It relied on an unusual and perhaps shaky corporate-academic collaboration. The effort was technically complex in bridging the analog and the digital. It cost Google, by then a public company, a fortune without a sure payoff. In addition, the project sparked heated legal arguments that swiftly led to a fiery lawsuit. Would this promising but problematic corpus of text emerge intact? If so, how useful would it be, and for what purposes?

portal: Libraries and the Academy, Vol. 25, No. 3 Supplement (2025), pp. 23–29.

In the midst of this disorienting development, one that was both exciting and unsettling for those of us in academia, I attended the Washington DC Area Forum on Technology and the Humanities in November 2005.[1] Like seemingly everyone else at the time, I had just started a blog, and my notes from the forum sit among my earliest posts.[2] Clifford Lynch spoke, along with the legal scholar Jonathan Band, on "Massive Digitization Programs and Their Long-Term Implications: Google Print, the Open Content Alliance, and Related Developments."

I witnessed Cliff's uncanny ability to communicate in plainspoken ways about complex matters and to distill these matters down to the most critical issues.

For the first time, I witnessed Cliff's uncanny ability to communicate in plainspoken ways about complex matters and to distill these matters down to the most critical issues. In his typically clear and concise fashion, Cliff surveyed the positives and negatives of the scanning operation that was then called Google Print. He situated it along a longer arc of the history and possible futures of mass digitization. The conversion of academic journals to bits had preceded Google's scanning of books, leading to new entities like JSTOR that were on their way to becoming essential pillars of scholarly research in the new century. A gigantic digital library of books was an obvious and worthy next step.

But Cliff also observed that books presented entirely new challenges in the passage from print to digital. The undertaking involved a mixture of legal issues related to the lengthy window of copyright, the difficulty of locating and aggregating the rights around millions of individual volumes, and the precious cultural value society places on books. These factors intensified the emotion around Google's project in a way that scanning old issues of *The American Historical Review* did not. From shelf to screen would not be an easy transition.

It would, however, be worthwhile for a vast array of reasons, Cliff concluded. Google's alliance with research libraries was a major advance toward the emergence of what the science fiction writer H. G. Wells called the "world brain"—a comprehensive store of human knowledge that was freely accessible to all. Moreover, if we could produce such a massive digital library, it would lead to entirely new avenues of thought and research. "Large scale open access corpora are now showing great value, using data mining applications: see the work of the intelligence community and the pharmaceutical industry," Cliff explained to the forum audience. The lecture hall was filled with scholars in humanistic fields, many of them unaware of what was happening, for example, at the Central Intelligence Agency or the pharmaceutical giant Pfizer. Cliff asked, "Will the humanities follow with these large digitization projects?"

The Novel Potential of Networked Digital Collections

This unexpected point stuck with me. In 2005, I was a young historian who was fascinated by digital media and technology. Like my departmental colleagues, I was unsure how scanned collections might change the practice of history beyond offering online access to primary resources and the ability to broadcast scholarly work across the globe

in this new medium. Cliff was saying something radical: networked digital collections transcend the benefits of mere remote access, the ability to read at a distance. They will inevitably lead to utterly new approaches and uses. Books would, of course, continue to serve as a rich narrative form of knowledge. But by becoming digital, books could mingle with other forms of information and data to enable research we had not previously considered or been able to do.

Books would, of course, continue to serve as a rich narrative form of knowledge. But by becoming digital, books could mingle with other forms of information and data to enable research we had not previously considered or been able to do.

For instance, Cliff noted in his forum talk, the value of quotidian humanities reference works, such as gazetteers and ontologies, would grow enormously in the future as aids to search the large corpora that were coming online. These highly structured works could be fruitfully joined with less structured text, unlocking previously hidden knowledge and enabling new forms of analysis. At the same time, free-form text corpora could generate or supplement new kinds of reference works. The key was to allow for the bidirectional movement and combination of knowledge across boundaries.

Cliff's core insight predicted that digitization and networking would lead to unexpected translational impacts and to new uses that arise from the interplay of collections and disciplines. The Lynchian perspective brought revolutionary clarity to the work of libraries, technologists, and scholars in the wake of the early web. If formerly isolated resources are digitized and connected, what capabilities might we gain? The many answers to this compelling question ignited two decades of progress in digital research and learning. Cliff was there to identify the embryonic possibilities before they fully emerged.

Cliff's core insight predicted that digitization and networking would lead to unexpected translational impacts and to new uses that arise from the interplay of collections and disciplines.

The Promise of Digital Scholarship

Soon after the 2005 forum on Google Books, Cliff published two essays that explored the unfolding landscape for digital scholarship: "Open Computation: Beyond Human Reader-Centric Views of Scholarly Literatures" (2006) and "The Shape of the Scientific Article in the Developing Cyberinfrastructure" (2007). Both identified what had begun to happen to research as corpora were redesigned for diverse uses beyond traditional narrative consumption and were placed on servers that could enable such uses.[3] His 2007 article explained,

> A number of interrelated technologies such as text mining and analysis are very active, vibrant and well-funded research areas, attracting extensive participation and investment from government and industry as well as academia. And, more recently, we are seeing experiments not only in computing on literatures to derive insights, but in the actual rehosting of literatures within new analysis, usage and curation environments: here a scholarly literature is actually imported into a new usage environment that adds value through computation and perhaps also through social interaction.[4]

A set of scholarly resources that was digitized (or, increasingly, born digital) could be synthesized, transplanted, and operated on by new computational techniques. The resources could also be curated and supplemented by a dedicated community of practitioners. The rapidly growing collection of open access scientific articles was swiftly heading in this promising direction: "The use of the corpus of scientific literature is already changing in other ways as well: not only do human beings read (and interact with) articles from the scientific literature one article at a time, but we are also seeing the deployment of software that computes upon the entire corpus of scientific literature."[5]

Considerable thought and work would be needed to create this potent new body of networked information. Legal issues around copyright, as well as the resistance of some scholars, would have to be addressed. The nascent open access movement—still relatively young 20 years ago—would need to expand significantly to cover a much greater percentage of scholarly output. Digital infrastructural issues related to servers, application programming interfaces, and the exact formatting of these resources would need to be determined as well. Platforms with multiple layers acting together on documents and data would have to be designed, built, and maintained by information technology specialists. Cliff said,

> Resolving this problem implies a somewhat "thicker" layer of software mediating between the machine-readable representation of articles in the cyberinfrastructure environment and the human reader. Today, articles are most typically delivered to readers in very unexpressive, semantically limited forms such as PDF [portable document format] or HTML [hypertext markup language], which are rendered by a PDF viewer or a web browser, respectively. As we build out the collaboratories and virtual workspaces to house the activities of our virtual organizations within the cyberinfrastructure, I hope that we will see a new generation of viewing and annotation tools presumably working on semantically rich XML [extensible markup language] document representations.[6]

In addition, this developing digital grid greatly increased the value of the "secondary" elements of scholarly production—the data, notes, links, bibliographies, charts, and other visualizations that accompany the narrative. These outputs could be just as useful as inputs to new computational modes of consumption as the primary text would be, perhaps even more so as the ability to combine information from different sources improved. Making it easier to decouple the different parts of a standard article or book, allowing them to be available for use and reuse independently, would soon pay dividends.

The payoff would extend beyond text and data mining and other quantitative methods, as new socio-technical processes materialized. This would be especially true for unique primary sources, as Cliff detailed in his remarks at the 2009 ARL-CNI (Association of Research Libraries and Coalition for Networked Information) Fall Forum,

"Special Collections at the Cusp of the Digital Age: A Credo."[7] Collections physically split by the happenstance of history could be virtually reunited. Decentralized remote audiences could then annotate those synthetic collections in a more thorough way than centralized local curators, perhaps even producing new critical scholarly editions. Materials that formerly interested only a small set of scholars in one discipline might hold value, through widespread access and networking, to experts in many other fields of study. Even more excitingly, in the case of some historical documents in poor condition, new imaging technologies with greater dynamic range might reveal their full contents for the first time, integrating them with existing scholarship.[8] "Put simply, special collections are a nexus where technology and content are meeting to advance scholarship in extraordinary new ways," Cliff concluded.

. . . through his writing, talks, and countless conversations, Cliff became our trusted guide to the transformative activities that lay ahead.

In the last three decades, through his writing, talks, and countless conversations, Cliff became our trusted guide to the transformative activities that lay ahead. While he was not blind to the many pitfalls and problems that would await, his underlying optimism served as a spur to all of us. Cliff thought we had much to look forward to as a research community, and he urged us to explore that future together, saying:

> The opportunities are truly stunning. They point towards entirely new ways to think about the scholarly literature (and the underlying evidence that supports scholarship) as an active, computationally enabled representation of knowledge that lives, grows and interacts with its contributors rather than as a passive archive or record.[9]

The future Cliff imagined earlier than most, and explained better than anyone else, came into being. New possibilities appeared through the combination of large-scale corpora, computational tools, and the distributed efforts of researchers. One can see these multifaceted opportunities in the wide range of presentations at the CNI meetings every year. What might we learn about the biology and migratory patterns of whales if we digitized the filing cabinets and photographs of marine biologists, modernized their old Microsoft Access databases, and merged them with contemporary images and data from drones?[10] How might we share complex scientific instruments among universities and colleges through open protocols and pooled data, enabling new forms of collaboration and democratizing access for researchers at smaller institutions?[11]

The Future of Networked Information

Machine learning and artificial intelligence (AI) are the most recent, and perhaps the most powerful, technologies to leverage the digital environment Cliff envisioned. Indeed, the fundamental breakthrough behind many recent advances in AI, the transformer, is, at its core, rather Lynchian. It shows how the latent connections between documents and data can produce, on the fly, surprising outputs and discoveries.[12] The utility of information is exponentially enhanced by being consumed as part of a highly processed network.

Cliff saw this possibility well before AI became a regular part of the toolkit of researchers. In his 2019 essay "Machine Learning, Archives and Special Collections: A High-Level View," Cliff cataloged how machine learning and AI could benefit cultural heritage organizations and the scholars and students who use the massive, diverse collections housed by those organizations:

> Some applications where machine learning have led to breakthroughs that are highly relevant to memory organizations include translation from one language to another; transcription from printed or handwritten text to computer representation (sometimes called optical character recognition); conversion of spoken words to text; classification of images by their content (for example, finding images containing dogs, or enumerating all the objects that the software can recognize within an image); and, as a specific and important special case of image identification, human facial recognition . . . The key strategy for the cultural memory sector will be to exploit these advantages, adapting and tuning the technologies around the margins for its own needs.[13]

The charge Clifford Lynch has given to us is not just to adopt a succession of new technologies for their own sake but also to imagine how we might shape those technologies to improve research, learning, and public understanding. The digitization and networking of resources might thus enable new insights. In the best of cases, this virtual effort will make an impact back in the physical world: new therapies to help patients, new ideas to improve communities, and new books to join the older ones on the shelves of libraries.

Dan Cohen is the vice provost for information collaboration, the dean of the library, and a professor of history at Northeastern University in Boston; he may be reached by email at d.cohen@northeastern.edu.

Notes

1. Roy Rosenzweig Center for History and New Media, "Google Print & Mass Digitization Projects: DC Tech & Humanities Forum to Be Held on 11/28/05," November 11, 2005, https://rrchnm.org/news/google-print-mass-digitization-projects-dc-tech-humanities-forum-to-be-held-on-112805/.
2. Dan Cohen, "Clifford Lynch and Jonathan Band on Google Book Search," blog post, November 28, 2005, https://dancohen.org/2005/11/28/clifford-lynch-and-jonathan-band-on-google-book-search/.
3. Clifford Lynch, "Open Computation: Beyond Human Reader-Centric Views of Scholarly Literatures," chap. 19 in *Open Access: Key Strategic, Technical and Economic Aspects*, Neil Jacobs, ed. (Oxford, UK: Chandos, 2006), 185–93; Clifford Lynch, "The Shape of the Scientific Article in the Developing Cyberinfrastructure," *CTWatch Quarterly* (August 2007).
4. Lynch, "Open Computation."
5. Lynch, "The Shape of the Scientific Article in the Developing Cyberinfrastructure."
6. Lynch, "The Shape of the Scientific Article in the Developing Cyberinfrastructure."
7. Clifford A. Lynch, "Special Collections at the Cusp of the Digital Age: A Credo," *Research Library Issues* 267 (December 2009).

8. See, for example, Reviel Netz and William Noel, *The Archimedes Palimpsest: How a Medieval Prayer Book Is Revealing the True Genius of Antiquity's Greatest Scientist* (Philadelphia: Da Capa, 2007); a digital version is available at http://www.archimedespalimpsest.org/.
9. Lynch, "Open Computation."
10. Harish Maringanti, "Unveiling Whale Wisdom: Digitizing the Patagonian Right Whale Dataset," project briefing, CNI (Coalition for Networked Information) Spring 2024 Membership Meeting, San Diego, March 26, 2024, https://www.cni.org/topics/digital-libraries/unveiling-whale-wisdom-digitizing-the-patagonian-right-whale-dataset.
11. Forough Ghahramani, Maureen Dougherty, and Barr von Oehsen, "The Ecosystem for Research Networking (ERN): Exploring Democratized Access to Research Instruments," project briefing, CNI Fall 2023 Membership Meeting, Washington, D.C., December 11, 2023, https://www.cni.org/events/membership-meetings/past-meetings/fall-2023/schedule-f23.
12. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017), https://arxiv.org/abs/1706.03762.
13. Clifford A. Lynch, "Machine Learning, Archives and Special Collections: A High-Level View," *Flash* 38 (2019).