IJDC | *Conference Paper*

From Building a First-Generation Digital Library Infrastructure to Reimagining Discovery

Stuart Snydman
Associate University Librarian and Managing Director for Library Technology Services, Harvard University

Martha Whitehead
Vice President for the Harvard Library and University Librarian, Harvard University

Abstract

Twenty-five years ago, Harvard University was in the early stages of a project to build a first-generation digital library infrastructure. The project was carefully named the Library Digital Initiative (LDI), signifying that 'digital' would be an integral and integrated aspect of 'library' and not a separate entity. The initiative aimed to develop knowledge and expertise relating to digital objects, as well as technical infrastructure to create, curate, access and preserve them, and to integrate the new digital collections with Harvard's extensive tangible collections.

Today, we still benefit from the foresight of this first-generation development and the subsequent ones it spawned, but we are also at a pivotal point of reflecting on lessons learned and opportunities to be seized as we rebuild and reimagine our digital infrastructure and services in a vastly expanded data ecosystem. Predicting what libraries will look like two decades ahead is always conjecture. What we do know, however, is that while the themes and challenges from the past two decades endure, the way we are tackling them is different. This paper examines what has changed since early library digital initiatives, and the imperatives we see for the future.

Submitted 3 February 2025 ~ *Accepted* 20 February 2025

Correspondence should be addressed to Stuart Snydman, Email: stuart_snydman@harvard.edu

This paper was presented at the International Digital Curation Conference IDCC25, 17-19 February 2025

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Libraries are at the foundation of academic and cultural institutions and research that crosses organizational and national boundaries. We play a crucial role in digital curation, preservation, and access to the world's knowledge and cultural memory, and in facilitating the creation of new knowledge and discovery.

At Harvard University, the Library Digital Initiative that launched in 1998 was a key element of the university's journey into the digital age and represented a transformative reimagining of how libraries integrated digital resources into the mission of our institutions. It foretold today's challenges of grappling with the vast expansion of knowledge in all formats, with a mandate to provide equitable and inclusive access to those resources.

The demands on libraries have changed dramatically over the past 25 years, in the context of rapid technological evolution, the increasing complexity and size of digital data, and the sophistication and expectations of users. At Harvard, our aim has broadened from curating, organizing and facilitating access to the collections inside our virtual and physical walls to envisioning a global network with equitable access to diverse research outputs and cultural resources held by communities around the world.

Now, as we look to the future, we must grapple with complex technical challenges and opportunities presented by a fragmented and diverse systems environment and the emergence of artificial intelligence and other transformative innovations, and help users navigate an information ecosystem where trust is easily compromised and authenticity is hard to determine. The journey from a first-generation digital library ecosystem to reimagining discovery requires a commitment to both innovation and enduring library values.

Aspirations: Advancing Open Knowledge

Harvard Library's strategic directions are outlined in a discussion paper titled *Advancing Open Knowledge* that explores the library's aim of expanding world knowledge and discusses broad strategies for advancing it in the years ahead. The paper notes that

> 'today's scholars and the general public are in the midst of a critical moment in our knowledge environment. Innovations in technology have raised expectations that any information we want, from any part of the world, is now available at our fingertips and always will be. In reality, the information globe is still dominated by the wealthiest nations, trustworthy information can be hard to find, and it can be gone tomorrow.' (Harvard Library, 2020)

At the foundation of *Advancing Open Knowledge* is a deep commitment to working collaboratively to champion access to diverse perspectives. It affirms that scholars at Harvard and everywhere will benefit from collaborative networks that support equitable access to a diversity of content, easy engagement with trustworthy information, and thoughtful preservation for the future—a global knowledge commons. As the paper states, 'Knowledge is both local and global: we aim to surface local creations and have them accessible in all parts of the world, to ensure a rich diversity of perspectives and cultures' (Harvard Library, 2020).

To support these aspirations, Harvard Library committed to the ongoing evolution of our digital infrastructure as a key enabler of our strategies. There is a strong legacy of

investments and expertise to build upon, but there are also challenges inherent to the way library digital infrastructure has developed over time.

Looking Back: The Library Digital Initiative (LDI)

The Library Digital Initiative (LDI) was launched in 1998 as a five-year program with $12 million allocated by Harvard University, marking an important milestone in the evolution of digital library infrastructure. During this time, many of Harvard's peer institutions were similarly investing in digitization initiatives, laying the groundwork for what would become a transformative period for academic libraries globally.

The late 1990s saw the nascent stages of digital library programs and departments, often emerging from early digitization projects and the foundational work of initiatives like SGML, XML, and TEI markup standards in the burgeoning field of digital humanities. These programs generated increasing volumes of digital data, prompting the need for solutions for its management, storage, access and use—critical issues that, at the time, lacked robust technological infrastructure.

While digital preservation was frequently discussed, it remained largely theoretical as the immediate priority was finding ways to store and organize data effectively. Harvard, alongside several other pioneering institutions, took bold steps to address these challenges at scale, implementing systems that set new benchmarks for the field.

The naming of Harvard's program as the Library Digital Initiative was intentional, signaling the deep integration of technology into the library's core mission. Dale Flecker, then-Associate Director, Planning and Systems, envisioned digital as 'just part of being a research library', a perspective that now resonates across modern academic institutions where technology is intrinsic to all aspects of library services. Flecker and his colleagues pursued an ambitious goal: to create a production system capable of supporting functions and requirements unmet by commercial technologies of the time.

Digital Repository Service (DRS)

The LDI included what Flecker described as a general repository: 'Its purpose is to provide a robust service to store, manage, protect, and serve heterogeneous digital objects. And to provide information and facilities for the preservation of those objects' (Flecker, 2000).

This vision reflects the foundational role digital repositories have come to play in modern research libraries and memory institutions. At the time, such systems were novel, with few academic research libraries operating at scale. Harvard's team rose to the challenge, building a system designed for longevity. The functional requirements for this Digital Repository Service (DRS) were completed in 1999, and since then the DRS has provided reliable preservation of Harvard's extensive digital collections, evolving alongside new file formats, user needs, and technological advancements.

The success of the DRS is not solely attributable to the system itself but to the ingenuity, dedication, and collaborative spirit of the staff who built and sustained it. Technologists and library professionals worked together to establish Harvard's DRS as a cornerstone of its digital preservation program. At the same time, the emergence of digital preservation as a specialized domain of professional expertise has driven significant advancements in preservation technologies and standards. Specialists from Harvard, in collaboration with peers from other institutions contributed significantly to innovations such as early web and email archiving systems, JHOVE (a format validation tool), PDF/A archival standards, the Audio Engineering Society's Audio Preservation Standard, and preservation metadata standards such as METS and PREMIS. These contributions underscore the vital role of digital preservation professionals in shaping the evolving landscape of library and archival practices. This period represented a pivotal moment of

growth and transformation for the library, driven by a spirit of innovation and a willingness to embrace risk.

Today, it is difficult to imagine a leading research library without access to a robust and durable digital preservation system. For Harvard, the DRS embodies the institution's commitment to preserving its invaluable assets, from its nearly 400 years of collections to the intellectual output of its faculty and students, including scholarly publications, electronic theses and dissertations, and research datasets. Harvard's contributions to digital preservation have been instrumental in defining the field's trajectory, ensuring the enduring accessibility and integrity of the cultural and scholarly record in an increasingly digital world. Looking ahead, the DRS continues to serve as a model for innovation and collaboration in digital preservation, with its next chapters poised to build on this remarkable legacy.

Open Collections Program (OCP)

A key program related to the LDI, and benefiting from external funding, was the Open Collections Program (OCP), a focused effort to provide open global digital access to historical resources in Harvard's libraries—digital content that the DRS was designed to serve and preserve. The goal was to support teaching, research, and public engagement by digitizing and organizing materials in thematic collections.

The OCP tackled the challenges of developing coherent digital collections from the independently operating physical repositories of Harvard's very decentralized library system and ensuring their long-term access. This required extensive collaboration across faculties and libraries, with multiple advisory committees including faculty, content, and technical groups to manage topic refinement, material selection, and workflow standardization. Thousands of books, pamphlets, and manuscripts were evaluated, and workflows established to ensure accurate metadata and cataloguing. The program adopted a hybrid approach of microfilming for preservation before digitizing using flatbed scanners or digital photography. Ambitions were scaled to the limitations of early technologies and available budgets. By 2003, over 480 books and multiple manuscripts had been digitized and uploaded to the DRS, and a prototype website was launched for testing the organization and accessibility of digitized materials. These modest beginnings, followed by continued investment in and evolution of Harvard's approach, paved the way towards the more than 6 million objects digitized from the library's collections and publicly available today.

The OCP, and similar digital collections programs emerging elsewhere, had broad implications for digital collections. It redefined how academic libraries approached digitization, focusing on creating accessible, integrated collections instead of isolated exhibits. By prioritizing usability for educators and researchers worldwide, the OCP bridged gaps in access to primary sources, enriching global scholarship. The OCP also established a close connection to Harvard's emerging digital preservation program, demonstrating the inextricable link between the dual imperatives of preservation and access. Perhaps most importantly for sustainability of future digitization initiatives, the program demonstrated the potential of an integrated library digital infrastructure.

The OCP exemplified the power of leveraging institutional strengths to create a scalable, sustainable, and impactful digital resource. By balancing ambitious goals with practical workflows, it provided a roadmap for future large-scale digitization efforts in the academic library sector.

Open Access and Research Data Curation

Though outside the direct scope of the LDI, closely related to it was a set of Harvard initiatives focused on providing open access to research outputs. Harvard Library had long been concerned about the unsustainable prices of subscription journals and the barriers

they presented to knowledge sharing and saw new opportunities in digital technologies. There was also a growing awareness of the need to curate and preserve the bourgeoning research data underpinning scholarly publications, and to make that data as openly accessible as possible and as secure as necessary.

In the realm of scholarly publishing, Harvard's approach to the fiscal unsustainability and academic constraints imposed by major journal publishers was exemplified by its 2004 cancellation of the Elsevier 'big deal.' To combat the challenges of traditional subscription models and advance the open access movement, Harvard pioneered an academic-led approach to open access. In 2008, it was the first US university to adopt an open access policy, and the first in the world to adopt an open access policy by faculty vote rather than administrative edict. This was the first university 'rights retention' open access policy, ensuring that authors and the institution held the nonexclusive rights needed to authorize open access. Today, about 70 university open access policies around the world (in North America, Europe, Africa, and Asia) are based on the Harvard model (Suber, 2020).

To support its open access policy, Harvard launched a repository called Digital Access to Scholarship at Harvard (DASH)[1] in 2009, using the open-source platform DSpace as its underlying technology. Today, DASH contains over 58,000 works of scholarship, including articles, conference proceedings, working papers, case studies, books and book chapters, theses, and dissertations.

The creation of DASH at Harvard, and the emergence of similar open access or 'institutional repositories' elsewhere, reflected a broader trend of large academic research libraries adopting purpose-built or specialized systems to support an increasingly diversified range of services. For Harvard, the development of an enterprise digital preservation repository like the DRS served distinct needs, while the unique workflows of open access necessitated a separate and specialized system. This differentiation extended to other library domains as well, with platforms dedicated to digital collections or specialized tools for managing archival collections and finding aids.

Similarly notable during this time is the emergence of specialized research data repositories. Dataverse[2], launched in 2007 as an initiative of Harvard's Institute for Quantitative Social Science (IQSS), was purpose-built to address the data-sharing challenges faced by researchers (King, 2007; Crosas, 2011). It evolved in alignment with the FAIR principles of data—that research data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson, 2016).

As a technology, Dataverse contributed to the trend of institutions developing open-source platforms to meet specialized needs. Designed for distributed and decentralized adoption, Dataverse could be implemented by individual institutions, consortia, or domains. Simultaneously, it embraced a standards-based approach to identifiers and machine-readable metadata, ensuring interoperability and portability of data. This approach demonstrated a commitment to global data-sharing principles, enabling broad discovery and interoperability of research data.

More broadly, the emergence of Dataverse—and the pivotal role libraries played in adopting and supporting it—reflected the growing recognition of research data as an essential scholarly output to be stewarded and discovered, complementing traditional publications.

Harvard's experience with both DASH and Dataverse exemplifies several trends that have defined the growth of library digital infrastructure over the past two decades. They highlight the library's expanded role in supporting the full lifecycle of academic research—from data collection to publication. They embraced technical solutions that enable interoperability across open networks. Additionally, they underscore another notable trend: the adoption of open-source software as a solution where the commercial

[1] Digital Access to Scholarship at Harvard (DASH): https://dash.harvard.edu
[2] Harvard Dataverse: https://dataverse.harvard.edu

marketplace has fallen short. For Harvard and our peers, these open access publication and research data platforms have reinforced the library's ongoing commitment to open scholarship and open science, while also reflecting the broader transformation of library technologies.

Current Challenges

Proliferation of Internal and External Digital Content

In writing about the LDI in 2000, Flecker observed,

> 'Perhaps the most striking feature of the LDI to date, and undoubtedly one of its major weaknesses, is that it has been predominantly focused on resources inside the University. The majority of any library's digital offerings will undoubtedly be held externally. Over time the key challenge in building institutional digital libraries will be the integration of the many heterogeneous external resources into coherent services for the population of the institution. To date, little attention (beyond the issues of portal organization and access management) has been paid to what it means to integrate internal and external resources.' (Flecker, 2000)

Twenty-five years later, that integration of internal and external resources remains elusive, and has various new dimensions. Today, we are grappling with the curation of digital content in several concentric circles:

Library-like objects: The LDI was a major step forward for its time, aiming to support not just the digital content of the traditional library but other 'library-like objects (research resources of lasting value)' at museums and other entities across the university. It was anticipated that many different specialized local repositories, content delivery applications, and discovery interfaces would be required over time, given the different types of digital objects being managed and used. Indeed, fast forward, and Harvard Library has developed exceptional systems for search and discovery like Harvard Geospatial Library[3] and HOLLIS for Archival Discovery[4] that specialize and benefit specific media and areas of study. They are each of their time, however, separated by years of differences in design approaches, shifted organizational priorities, and evolved technology standards. As a result, we host an array of discovery systems that, in the aggregate, betray their value by being both unintuitive and intimidating in their variety and complexity.

Institutional research inputs and outputs: The 'library-like objects' above can be the inputs or outputs of research, but today's university is also managing a plethora of other digital resources supporting the research lifecycle. A variety of different kinds of datasets are acquired by research teams for their own use, outside the library context of collections shared across the university, and often involving stringent regulations requiring specialized infrastructure. Increasingly, the library is acquiring or creating complex datasets that *are* shared university-wide, requiring a different set of services than the earlier serving of digital objects. And for research outputs, the library is engaging in research data management and scholarly communication services that involve library principles of curation, description and preservation, but in the context

[3] Harvard Geospatial Library: https://hgl.harvard.edu
[4] HOLLIS for Archival Discovery: https://hollisarchives.lib.harvard.edu/

of very distributed research computing infrastructure. This environment is decidedly more complex than twenty-five years ago, and from a researchers' perspective the hope of discovering useful data within it is slim.

Beyond the institution: The LDI focused on building an institutional digital library and offering coherent services primarily for the population of the institution, while also opening its digitized collections to the world. Today, our aspirations have broadened to contributing to the development of an interoperable global network of resources. In a sense, we have shifted our gaze, from looking out at the world to being of the world. That has meant thinking more deliberately about diversifying the digital collections we curate to include perspectives missing in the past. It has meant considering how we can make our open resources more discoverable outside our gates. It has also meant that in addition to stewarding vast collections and open digital content, we want to see communities in all parts of the world empowered to produce, share, and retain ownership of local research outputs and cultural resources. From a digital infrastructure perspective, this means investing in technologies that support equitable, sustainable models for scholarly communications and open knowledge, building upon interoperable repository networks and providing discovery across them.

Fragmentation of Digital Systems

The late 1990s and early 2000s marked a period of rapid growth in digital library programs and departments, building on the emergence of library systems in earlier decades. As the digital revolution gained momentum, university libraries began to evolve, hiring specialized technology staff to support the increasing demands of managing and delivering library services online. These hires were dictated by budgets and priorities, with varying levels of investment across institutions.

For many libraries, this transformation began with the addition of systems librarians focused on implementing and maintaining a new generation of library catalogs and integrated library systems. As digitization initiatives expanded, libraries also brought in project managers and digitization staff to scan physical content into digital formats. At the upper end of the resource spectrum, as exemplified by Harvard's experience, well-funded institutions were able to hire highly technical staff, including software developers, systems administrators, and IT specialists, to build custom solutions and manage their growing technological infrastructure.

For institutions unable to develop their own systems, an emerging market of vendor offerings and commercial products provided alternative solutions. These products ranged from digital repository platforms to integrated library systems and content management systems. Over time, libraries became increasingly dependent on complex technology systems to manage their core functions. This reliance on technology required not only new skill sets but also entirely new staffing models and organizational structures to manage and maintain these systems.

The implications of this shift were profound. Libraries experienced significant cultural changes as they adapted to functioning as technology-dependent organizations. Staff who were accustomed to traditional library workflows were required to work alongside technologists, introducing both opportunities for collaboration and challenges in communication and alignment. Organizational structures had to evolve to accommodate the integration of IT teams within libraries, and financial models needed to account for the ongoing costs of technology, including infrastructure maintenance, upgrades, and staffing.

These developments were not uniform across the field. Well-funded institutions, national libraries, and consortia that could collectively organize and pool resources often emerged as leaders in digital library innovation. They built and implemented large-scale systems, often setting standards that influenced the broader library community.

Conversely, less-resourced institutions frequently struggled to keep up, relying heavily on vendor solutions and external support to meet their needs.

Looking back, the proliferation of technology in libraries brought with it unintended challenges. The rapid growth and accretion of technical systems often resulted in an ecosystem of siloed tools, each developed to solve specific problems but collectively creating new barriers to integration and efficiency. Many of these systems were developed in isolation, leading to fragmentation that still affects discovery, access, and preservation today.

Since the days of the LDI, Harvard Library's approach to building our digital infrastructure has embraced various technology offerings, including a major vended system for our integrated library system and other systems for digital and archival collections discovery, as well as the bespoke, homegrown system for our preservation repository and open-source systems for open access described above. These markets and choices have evolved alongside the information explosion accompanying the internet age, the dual expansion of open access and commercial publishing regimes, and the ever-changing expectations of academic knowledge seekers.

Harvard's digital infrastructure currently consists of approximately sixty systems addressing both our internal operational needs and the user-facing services that support Harvard's teaching and research enterprise. To maintain these systems, the library has built a highly professionalized IT group, in collaboration with the university's central IT organization, that must at once maintain operational integrity, business continuity, and information and data security while also innovating to keep pace with the changing needs of students and scholars. The most dynamic of these systems enable discovery and access to both the digital content we collect and the broader universe of digital and analog data available through our peers and the broader universe of knowledge.

The Limits of Library Discovery

While the challenges noted above are significant, clearly there have been significant advances in library digital infrastructure as well over time, particularly through a range of collaborative projects and innovations. Early examples include Google Books (and the resulting HathiTrust[5] collaboration), global collaborations like the International Image Interoperability Framework (IIIF)[6], and efforts such as the Digital Public Library of America[7], RightsStatements.org[8], and web archiving initiatives through the Internet Archive[9]. In a recent example, the international COAR Notify Initiative[10], Harvard Library demonstrated opportunities to link research data to publications in institutional repositories. These projects show a shared effort to expand access and improve discovery through innovation and collaboration.

A prime example of how far our digital library infrastructure has come, and yet has fallen short, is the case of IIIF, a collaborative initiative embraced by institutions worldwide to create open standards and API-based infrastructure for image access and delivery (Snydman, 2015). With its adoption spanning libraries, archives, and museums, it is estimated that IIIF enables over a billion cultural heritage objects to be accessed globally. Harvard was an early adopter of IIIF and leveraged it to integrate image discovery and access across organizational boundaries, including the library, museums and academic technology. Despite the success of IIIF in enabling institutions to adopt a common standard and technical framework for access to a vast global treasure trove of

[5] HathiTrust Digital Library: https://www.hathitrust.org/
[6] International Image Interoperability Framework (IIIF): https://iiif.io
[7] Digital Public Library of America (DPLA): https://dp.la/
[8] RightsStatements.org: https://rightsstatements.org/
[9] Internet Archive: https://archive.org/
[10] COAR Notify Initiative: https://coar-repositories.org/what-we-do/notify/

digital images, a persistent challenge remains: discovery. While IIIF was designed to be open with an eponymous goal of interoperability, these cultural heritage image resources often reside in institutional silos and are discoverable only through local systems or commercial search engines. This fragmentation underscores the limitations of existing digital infrastructure and the need for transformative approaches.

While these efforts and others have scaled and advanced access globally, they reveal the entrenched nature of our field's aging digital discovery methods. We believe now is the time for a more disruptive and transformational change that challenges the ways in which we connect users with the vast universe of cultural and scholarly data.

Opportunities

Harvard Library sees the current moment in our field as an inflection point. In the early days of library digital infrastructure and in developments since then, managing specialized content and its preservation were necessarily a primary focus. Now, across each of the realms of digital content that our users seek, the most challenging imperative and possibly the most opportunity lies in enhancing discovery and engagement. This increased focus on the user is stimulated in part by the emerging normalization of natural language and semantic search, as well as the potential of generative AI to transform how information is created, organized and used. We have by no means solved the myriad of challenges and complexities in building the modern digital library infrastructure, and there is work to do on many fronts. With that said, we think now is the time to re-evaluate the aging approach and assumptions to library search and work towards its next generation.

The Potential of AI

In 2023, Harvard Library embarked on a set of initiatives to explore the library-related opportunities of generative AI, in concert with university explorations under the headings of teaching, learning, researching and administration. We framed our initiatives by saying that generative AI and research libraries share a fundamental promise: the ability to draw upon a broad corpus of existing information to answer questions and generate new information. In this equation, Harvard Library brings the fundamental value of access to information, meaning access to trustworthy information spanning centuries, regions, and voices around the globe.

The library's AI initiatives have spanned numerous aspects of library expertise. Centering user interests, Harvard Library's User Research Center undertook research exploring how generative AI tools are impacting students' search and research habits. Staff in other units volunteered to experiment with using new tools to enhance metadata practices and have adopted new workflows. Technologies enabling speech and handwriting recognition and transcription have been readily tested and adopted, speeding the creation of accessible digital content. And, to support a wide range of research, teaching, and creative endeavours—including innovative applications such as training large language models—the library has publicly released a dataset of approximately one million public domain books digitized from its collections in the Google Books project[11]. Most importantly, the library's AI explorations have included a focus on reimagining discovery, as described below.

[11] https://library.harvard.edu/services-tools/harvard-library-public-domain-corpus

Reimagining Discovery

Harvard Library's *Reimagining Discovery* initiative represents a bold and innovative vision for the future of library discovery systems. Initially, the initiative aims to fundamentally transform how users interact with the vast collections of materials held at Harvard, using emerging technologies, including AI, to usher in a new paradigm of discovery. This effort is both a response to the changing expectations of users and an acknowledgment of the opportunities offered by cutting-edge advancements in AI.

The guiding challenge of academic research libraries has always been to connect researchers and students with the resources they need to conduct impactful research and high-quality instruction. However, this role is increasingly complicated by the sheer volume and complexity of available information and the fragmentation of digital systems as noted above. And Harvard's vision goes beyond merely improving access to its own vast holdings—it seeks to provide tools that support discovery in a broader, global context, while maintaining the core values of trust and authenticity.

The *Reimagining Discovery* initiative is driven by feedback from users, which underscores key challenges in the current discovery environment. Many users have described the process as overwhelming, citing the fragmented nature of discovery systems and the inconsistencies in how information is presented. Comments such as, 'Finding materials at Harvard is onerous' and 'I'd rather use Google' reflect the urgency for a more intuitive, user-centered approach.

As described earlier in this paper, Harvard Library's discovery ecosystem, like many academic institutions, has been shaped by decades of bespoke systems and specialized tools. While these systems were designed to meet specific needs, they often create barriers to seamless access and navigation. Recognizing these limitations, *Reimagining Discovery* envisions a cohesive and integrated platform designed to meet the evolving expectations of its users.

A pivotal moment for the initiative came in late 2022, when the release of OpenAI's ChatGPT illustrated to a brand new audience the transformative impact of generative AI tools on how users interact with online information. Harvard Library embraced this opportunity, reimagining our approach to discovery systems by experimenting with natural language interactions, semantic search, and AI-powered recommendations. The foundation for this vision was laid through an innovation grant project, *Talk with HOLLIS*, which explored the integration of a generative AI chat interface into Harvard's online catalog. The project aimed to enhance user experience, test new ideas, and build technical expertise within the library team.

The success of *Talk with HOLLIS* informed a broader three-year vision for *Reimagining Discovery*. This initiative encompasses a comprehensive reevaluation of all existing discovery systems at Harvard and a commitment to experimenting with AI to improve search and navigation. The goal is to create a discovery environment that is intuitive, efficient, and effective for both casual users and expert researchers.

A cornerstone of this vision is *Collections Explorer*, an AI-driven platform that focuses on Harvard's distinctive and special collections. These materials are often unevenly described, scattered across various systems, and difficult to locate. *Collections Explorer* leverages AI to enable natural language search, provide contextual recommendations, and generate item-level summaries. For example, semantic search technology allows users to discover relevant materials even when their queries lack specific keywords. Additionally, large language models power features such as explanations of search results and suggestions for related prompts, enhancing the user's ability to navigate and refine their searches.

This ambitious initiative reflects Harvard's commitment to openness and transparency in the integration of AI. For example, through an early collaboration with Mozilla.ai, we sought to design a system that aligns with ethical principles and user-centered design. Throughout the development process, extensive user research, usability testing, and

stakeholder engagement have informed the design and functionality of the platform, ensuring that it meets the diverse needs of its users.

As libraries increasingly adopt AI and other emerging technologies, Harvard Library emphasizes the importance of collaboration and sustainability. Building AI systems at scale presents significant financial and environmental challenges. Harvard's approach acknowledges the necessity of shared infrastructure and partnerships within the library community to address these challenges collectively. While we recognize that we cannot single-handedly solve the broader issue of AI's carbon footprint or fully mitigate the environmental impact of our work, we are committed to being mindful of these challenges. For example, where possible, we choose technology solutions that minimize environmental impact and work with vendors who prioritize carbon neutrality. We also leverage Harvard's use of the Massachusetts Green High Performance Computing Center[12] for compute resources. This energy-efficient data centre, powered in part by renewable energy, is an integral part of Harvard's sustainable IT strategy.

Through *Reimagining Discovery*, Harvard Library seeks to redefine the role of discovery systems in academic research. By embracing innovation and leveraging AI, the initiative aims to create a future where library users can engage with collections in ways that are seamless, meaningful, and transformative. As libraries navigate this rapidly evolving landscape, Harvard's vision sets a precedent for integrating cutting-edge technology with the enduring values of trust, authenticity, and accessibility.

Implications for Libraries and Digital Curation

While these retrospective and aspirational views on library digital infrastructure have built on the experience of one large research library, they speak to several themes that resonate across our professional communities. These themes are offered as food for thought as we contemplate lessons and directions in digital curation.

1. Stay True to our Values

   In the wake of dizzying technological change and the temptations of the latest innovations, we must remain steadfast to our core values and double down on the principles that define us. This requires prioritizing systems and practices that reinforce the role of libraries as a trustworthy partner and maintain our integrity and the authenticity of the resources we steward. We must ensure that access to knowledge and resources is inclusive and equitable, and promote global collaboration and open science. And we must commit to practices and technologies that minimize environmental impact while acknowledging the energy-intensive nature of storage, compute, and data infrastructure.

2. Be Fiercely User-Centred

   In our profession, rooted in decades of theory and experience, we can become overly invested in our own expertise at the expense of addressing users' broader needs. Our systems are often too complex and difficult to use, reflecting professional egos more than user desires. This complexity drives academic users toward simpler commercial systems, despite the bespoke library systems we painstakingly develop. We must listen to users. While it is not always the best approach to give individual users exactly what they ask for—this approach can easily do more harm than good to user experience—investing in research on user

[12] Massachusetts Green High Performance Computing Center (MGHPCC): https://www.mghpcc.org/

behaviour and preferences is critical to designing systems that meet their needs. Follow the data. It tells a compelling story.

3. Define 'Data' Broadly

   Research, teaching and learning rely on a broad range of high-quality, open, and securely accessible data. That includes library-like objects—often called digital collections—as well as a wide range of other kinds of digital resources acquired or generated in the course of research. Though these different types of assets involve similar management concerns, from security and access control to curation and discovery, they are often organizationally siloed in different parts of the library or the university, in large part for historical reasons. We must consider not only how these different types of digital assets are distinctive but what they have in common across them, and develop seamless services and infrastructure for the sake of users who need all of them.

4. Invest in Interoperability

   A plethora of bespoke systems have arisen over time to manage access to different types of digital assets, but often at the cost of information security, operational sustainability, and user experience. We need to move away from this fragmentation of digital systems, while continuing to effectively steward distinctive collections across a multitude of distributed repositories. Whether a repository is institutional, multi-disciplinary, or domain-focused, we should consider its potential as a node in an international network. We must continue the standards-based approaches that have enabled interoperability in recent decades, and invest in sustaining distributed repositories that meet local needs while enabling global access.

5. Embrace Risk and Change

   To truly leverage the opportunities of technology, we need to question old paradigms and embrace change with courage. Organizational cultures rooted in tradition may resist or approach such changes cautiously, but innovation requires adaptability. It also requires calculated risks. While risk can be expensive, collaboration and partnerships can help distribute costs and reduce individual institutional burdens. Today, embracing risk and change is an imperative as we consider the unprecedented opportunities of AI. Caution, scepticism, and ethical considerations are warranted, but as our academic communities engage with these technologies, we need to be leaders in leveraging their potential to address our aims in the knowledge ecosystem.

6. Collaborate with Each Other and with Industry

   The collaborative nature of the global knowledge commons is key to advancing research and scholarship; academic production as an individual or even institutional endeavour is long past. Efforts like open metadata initiatives, which explicitly encourage the sharing of data across institutional boundaries, should be celebrated and potentially emulated to encompass all forms of data. In terms of resource demands, the requirements for storage, compute, and technical expertise are too vast and complex for any single institution to handle alone. Open source is not free, and vendor solutions alone cannot meet our needs. Institutions of all sizes face these challenges, and building effective partnerships within academia and with industry will be critical to our success. Circling back to staying true to our values, our industry partnerships will leave data in the control of the communities who produced it while providing technical solutions to support it.

Conclusion: Past is Prologue

As we embark on the project of reimagining discovery, we also must take a moment to step back and reflect on the lessons of the past 25 years, and the prospects for the future. With the emergence of generative AI, are libraries bound for an existential moment? Certainly, if we have learned anything from the past, it is that evolution in the midst of broader societal technological change is normal and cyclical. The move from the analog to the digital card catalog, and then the emergence of digital library initiatives, are but two moments in our history that feel ever so familiar today. Our local online catalogs remain critical pieces of academic infrastructure, even as our analytics tell us that over 50% of traffic to our systems comes from commercial search engines[13].

We are right, if not obligated, to continually reflect on the library's role in academia and society as a steward of information and enabler of the creation and dissemination of knowledge. In moments of profound change in the nature of information and knowledge, it is also important to remember that libraries are one of the most enduring institutions in the history of society. The key is to determine where the library's considerable assets, in the form of the trust we have of users, the expertise we have in our people, and our vast and diverse information resources, can best be deployed in the emerging milieu.

References

Crosas, M. (2011). The Dataverse Network®: An open-source application for sharing, discovering, and preserving data. *D-Lib Magazine, 17*(1/2). doi:10.1045/january2011-crosas

Flecker, D. (2000). Harvard's Library Digital Initiative: Building a First Generation Digital Library Infrastructure, *D-Lib Magazine*, *6*(11). Retrieved from https://www.dlib.org/dlib/november00/flecker/11flecker.html

Harvard Library (2020). *Advancing Open Knowledge.* Retrieved from https://library.harvard.edu/advancing-open-knowledge

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research, 36*(2). doi:10.1177/0049124107306660

Snydman, S., Sanderson, R., & Cramer, T. (2015). The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. *Proceedings of IS&T Archiving 2015*, *12*(1), 16–21. doi:10.2352/issn.2168-3204.2015.12.1.art00005

Suber, P. and Whitehead, M. (2020). A brief history of open access at Harvard. *Harvard Library Office for Scholarly Communication.* Retrieved from https://osc-harvard.pubpub.org/pub/2m1q3hm6/release/2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., …, Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3.* doi:10.1038/sdata.2016.18

[13] Based on web analytics compiled for Harvard Library discovery environments between May 2024 and May 2025.