ALONG CAME GOOGLE

CONTENTS

ACKNOWLEDGMENTS

This book had a long gestation period, and Peter Dougherty and the staff of Princeton University Press supported us throughout. Instead of complaining about our failure to meet milestones, they offered encouragement. Instead of critiquing what we submitted in early drafts, they offered additional ideas for our consideration. We count ourselves sublimely fortunate to have had an opportunity to work with Peter. We are also grateful for the excellent reviewers he found for our manuscript. They, too, offered great ideas for strengthening our arguments, and the result is a greatly improved work.

We acknowledge at the outset that we have played a part in many of the events we chronicle in this book and have been friends and colleagues of many of the key figures. Deanna Marcum has served in leadership roles at the Council on Library Resources, the Council on Library and Information Resources, the Library of Congress, and Ithaka S+R. Roger Schonfeld has been with the Andrew W. Mellon Foundation and ITHAKA for his entire career, currently serves on the board of the Center for Research Libraries, and wrote the history of JSTOR. Although we have tried to remain impartial tellers of this story, we recognize that our professional histories no doubt color our perspectives.

The real strength of this history is that so many key figures in book digitization were willing to talk with us so candidly. We are forever grateful to our colleagues and friends who made time to tell us how they experienced the early days of digitization and to reflect on the implications for scholarly communication, broadly considered. This book is the story of those pioneers: Alan Adler, Ivy Anderson, Heather Christenson, Dan Clancy, Dan Cohen, Paul Courant, Robert Darnton, Laine Farley, Dale Flecker, Mike Furlough, Kevin Guthrie, Brewster Kahle, Michael Keller, Paul LeClerc, Tom Leonard, Mark Sandler, Richard Sarnoff, Donald Waters, and John Price Wilkin.

We are also grateful to our many colleagues who talked with us about the framework for the book, the positions we have taken, and their thoughts about the implications of book digitization. That list of valued friends is too long to acknowledge all of them properly, but we owe special thanks to Clifford Lynch, Lorcan Dempsey, Jessica Gardner, Kevin

Guthrie, and Oya Rieger for taking time to talk again and again about our work.

Thousands of individuals, far too many to name, contributed to the initiatives discussed in this study. And many more digitization experiments were pursued than we could cover in this project, even though we fully recognize that so many projects made important contributions. Each library that launched a digital project helped build the digital future. We owe an apology to the great number of librarians, technologists, and publishers who made invaluable contributions to book digitization whom we have not discussed in this book.

Finally, we are grateful to one another. When we first discussed this manuscript with Peter, he reminded us that it is not easier to write a book with a coauthor. In fact, he said, it is probably harder. For us, it has been an invaluable example of how collaboration can work. We remain good friends at the end of this long process, and we are convinced that our joint project is better than what either of us could have produced individually.

ALONG CAME GOOGLE

Introduction

On January 3, 2020, the *Washington Post* published a story about two graduate students working to save the University of Virginia's card catalog. Literature doctoral candidates Neal Curtis and Sam Lemley learned that the four million cards in the library's catalog that had not been updated in two decades would be discarded to make way for a massive renovation of Alderman Library. All the library's current holdings were included in an online digital catalog, so the outdated card catalog was understandably used by very few. Library administrators had determined that, at a cost of $750,000, it would not be worthwhile to scan the cards and create a digital surrogate of the outdated catalog. Instead, it seemed sensible to discard the card catalog, as so many other libraries have done since the 1970s when libraries began to create machine-readable descriptions of their collections instead of creating iconic cards that represented each book in the library. The dedication of the two graduate students prompted volunteers to help pack the catalog cards into 798 boxes and store them in an off-campus facility. They have bar coded each box for retrieval so that students and faculty will be able to recall a box of cards and look at the entries and notes about specific books. This charming story of students volunteering to pack boxes to preserve what Sam Lemley described as "an accurate, preserved-in-amber view of what the library was in the twentieth century" is a good introduction to the current challenges: what will be the library of the future?

The Virginia students recalled a time when the university library built a collection of books that served the needs of scholars and students. But the university librarian, John Unsworth, faced a new set of challenges that propelled him to raise money for and undertake a massive renovation of the library. Part of the challenge was to bring the building up to fire, safety, and accessibility codes, but a much bigger challenge was that most students and faculty wanted more than print collections from the library. They wanted access to the galaxy of information resources that exist not only at the University of Virginia and but also everywhere else, not just in print form but digitally as well.

There is no card catalog for today's information universe.

The end of the twentieth century and beginning of the twenty-first marked the transformation of libraries from builders and preservers of collections to information nodes that connect information seekers with resources from all over the world. This book focuses on what is perhaps the signal milestone in that transformation: the entry of Google into the library arena with promises of making all the world's information available to everyone.

With news of Google's plans, a shock wave went through the academic library community. Some librarians, eager to see an acceleration of digital activity, embraced the concept of a universal digital library and began advocating for change. Others argued that librarians were experts in locating and validating information resources; they did not appreciate other players moving into their domain. At its core, the Google digitization project challenged the definition of "library." A large literature has developed over the past decade in the field of "Google studies," with scholars seeking to examine the effects of consumer technology companies, pursuing a combination of business growth and societal disruption. Within this field, there are many episodes where Google dipped its toes into a new sector and left an entire ecosystem spinning in disruption. Our goal in this book is not to offer a final judgment of Google but rather to explore deeply one example of its efforts to target an information space, in this case the important legacy of published materials held by libraries, and the results on an existing sector and ecosystem.[1]

Ultimately, the rapid change in user expectations and professional expertise with digital technology led to intense conversations within the library and academic communities about the roles and responsibilities of both libraries and corporate entities, but meaningful organizational change in academic libraries was slower. The story of Google's digitization ambitions telescopes the dramatic changes in libraries, readers' research habits, and, perhaps, even reading itself.

Research libraries in particular came under pressure to adapt to this emerging reality. The notion that any library, no matter how large, could collect comprehensively the knowledge that was being produced was clearly not possible. With digital technology, many of the quality control mechanisms that had been in place for decades, for example, peer review of both journal articles and books through publishers with established reputations, now had to compete with preprints, open access publications, and start-up publishers with an array of review practices (some of them predatory). Libraries, no longer focused on collecting the best of the published record, began to think of their mission as wayfinding for their users. What is the universe of material on a particular topic? How does the reader find out about it?

In the midst of this transition from collection building to providing information services, Google made its dramatic announcement that it planned to digitize published books, which would be discoverable along with the websites Google was rapidly adding to its search capability. It knew, in a way that many others would only later recognize, that the layers of gatekeeping needed to produce publications and for the great research libraries to collect them would add significantly to the quality of the information available online.

In some respects, the Google project to digitize millions of books might have relieved research libraries of their stewardship responsibilities for legacy collections, allowing them to make the transition to digital libraries more quickly. But at least some librarians and a few scholars hesitated to entrust a corporation with digital library development. The story we tell here is how Google attempted to enter, and in some senses disrupt, the traditional scholarly communications systems that served the universities, their scholars and students, and their libraries for decades. We describe the competing forces that bolstered or fought against Google's efforts, as well as the fallout after the Google book digitization project fell into a legal quagmire. Finally, we describe the attempts to achieve some of the goals of the Google book digitization project in other ways and speculate about other possible scenarios that will benefit the scholarly community.

Looking back on the development of mass digitization and the efforts to thereby unlock access to our legacy of published books, it is clear that while many individuals and organizations played vital roles, none was more significant than that of Google. Even though the project that resulted and the impacts that it had were ultimately limited relative to the vision, millions of books have been digitized, the information they contain was made more discoverable, and access to many of them improved dramatically.

Google was able to lead because it was bold and agile. Larry Page had been interested in digitizing books since his student days at Stanford in the late 1990s. In 2002, he and Marissa Mayer determined that it would take forty minutes to digitize a three-hundred-page book. At-scale progress began to be realized when Dan Clancy was appointed to head the digitization project for Google. The team soon developed partnerships with publishers and then large research libraries in the United States, the United Kingdom, and several other countries. Paul

Courant, the university librarian and former provost at the University of Michigan, and his colleague John Price Wilkin, then Michigan's associate university librarian, would provide especially important leadership for both the library digitization efforts and later preservation initiatives.

For nearly a decade, Google and its partners aggressively pursued the dream of a digital universal library. When, on March 22, 2011, the U.S. District Court for the Southern District of New York rejected the legal agreement that had been proposed by Google after being sued by publishers and authors, the utopian library fizzled into little more than dreamy aspirations.

Looking back on the failed agreement in 2017, *Atlantic* journalist James Somers reflected on what had been lost:

> You were going to get one-click access to the full text of nearly every book that's ever been published. Books still in print you'd have to pay for, but everything else—a collection slated to grow larger than the holdings at the Library of Congress, Harvard, the University of Michigan, at any of the great national libraries of Europe—would have been available for free at terminals that were going to be placed in every local library that wanted one.[2]

But this highly desirable digital library was not realized. Somers wrote, "When the most significant humanities project of our time was dismantled in court, the scholars, archivists, and librarians who'd had a hand in its undoing breathed a sigh of relief, for they believed, at the time, that they had narrowly averted disaster."[3]

The library community was not as monolithic as Somers seems to suggest. For some portion of librarians, at least, for some scholars, and for some futurists, the Google project promised a vision that they had been dreaming of for years. For the advocates, the Google book digitization project was the strategy for libraries.

For several decades, multiple individuals and organizations have seen book digitization as the best strategy for creating a universal library. This is our analysis of how the Google book digitization project developed, how other organizations and individuals responded to the advent of large-scale book digitization, and the implications for libraries, publishers, and the scholarly community.

Google's dream of a universal library was a technology-centric version of an old idea. Throughout history, scholars, librarians, and others who yearn for knowledge and learning have dreamed of building a comprehensive library that is accessible to all. The Great Library of Alexandria, beginning in 288 BC, aspired to collect all of the papyrus scrolls that had been written. The Ptolemaic rulers intended the library to be a collection of all extant knowledge. They sent agents to many different places to purchase as many texts as they could. Because Alexandria was a port city, they searched incoming ships for texts and made copies of them for the library.[4] In modern times, the great research libraries such as Harvard, the British Library, and the Library of Congress, at least until recently, described themselves as "libraries of record," and they aspired to collect as much of the important scholarly and cultural record as possible.

As academic research expanded after World War II, publishing exploded, and libraries realized they could never acquire all that would interest their readers. Yet, the technological revolution inspired a great many library leaders to imagine how they would transform their organizations into the "universal library." In the 1960s, Library of Congress giants William Welsh and Henriette Avram believed that the enormous bibliographic database of that institution would become the core of the universal electronic library. Later, OCLC founder Frederick Kilgour would argue that a network of institutions could do that job more effectively. Computer scientists would question if we needed librarians at all if we focused instead on computational power to provide access to the entire corpus of knowledge.

But the digital transformation of our economy and society in recent decades has given rise to unbearable tensions—between global and hyperlocal, between universal access and filter bubble, between freedom and control, between openness and truth. During the industrial age, the library served as one of the greatest democratizing forces in American society. The network of public and research libraries was built on an aspiration (even if inequitably achieved) for any book to be available to any American without payment, yielding rich rewards for the economy and citizenship. A similar model for libraries was adopted in a number of other countries as well. And, no less than publishers and journalists, libraries too have been forced to wrestle with the tensions of the digital transformation.

Past generations of librarians focused on the needs of their own communities—their students and faculty members, not only those of the present but those of the future, in the case of the academic research libraries that feature prominently in our story. They spent handsomely to develop their collections, pushing aspirationally toward comprehensiveness in many cases, to provide access for local constituencies.

At the same time, they recognized that it was not possible to meet all of the research needs of their scholars and from the late nineteenth century began building sharing networks that made the academic library not a stand-alone provider but part of a network linked by lending. The pressure on research libraries to provide timely and comprehensive access to scholarly resources grew dramatically with the onset of World War II as the federal government became much more interested in the nation's scholarly capacity in a global environment.[5]

To achieve this end, libraries have developed mechanisms for building what Lorcan Dempsey has called a collective collection.[6] They have shared information about their collections with one another as a mechanism for coordinating their collecting activity. They developed a robust, frequently used, and increasingly streamlined interlibrary loan system to provide access to one another's holdings.

But, ultimately, libraries have responded more to local needs than national imperatives. And, perhaps more importantly, US libraries have lacked a vehicle to coordinate and prioritize their work.

Even before Google developed an interest in book digitization, research libraries had recognized the importance of digitizing their collections. And the dreams of librarians began to shift away from individual library comprehensiveness toward a vision of providing free, open, and public access to all material in digital form. But as with the effort to build a collective collection, libraries found coordination difficult and resources scarce. By 2004, they found themselves with strong third-party interest in their work: an outside technology company in growth mode with seemingly unlimited engineering and financial resources to support their aspirations. When Google stepped into the picture, digitization took off like a rocket.

In this book, we have set out to tell a story about how the vast intellectual heritage of our civilization has become (or will come to be) universally accessible. It is the story of how librarians, scholars, technologists, and entrepreneurs have imagined a global, accessible knowledge source and the extent to which they have succeeded or fallen short in realizing it. This is a story of how digitization has been viewed as the best hope for making our scholarly and cultural heritage universally accessible, and also a story about a sector not yet prepared to leap into the future. It is a story about the limitations of disruptive techno-solutionism in the face of well-coordinated incumbent market leaders, and a story in which some librarians have limited the dream because of financial restrictions and failure of will. It is also a story of the validated knowledge that is still all too absent from an online ecosystem filled with disinformation. And it is a story of how corporate America made the dream palpably real by using computer engineering to productive ends. In this story, there are many actors, all of good intentions. Inevitably, it is also a story of limitations and failures to collaborate. It is a story of how comprehensiveness exists only at a scale greater than any individual organization. Finally, it is a plea to fulfill the dream of making knowledge universally accessible to a world drowning in data and information.

We call this a history of digitization, even though large-scale digitization efforts have been under way for only slightly more than a decade. Digital technology has resulted in such rapid change that libraries and scholarly communication have been transformed in that short period. In viewing the revolutionary decade, we trace the history of library initiatives to digitize and make accessible their legacy collections; we describe the individual efforts to harness digitization for the public good as well as the collaborative efforts to achieve the goal. We look at successes, disappointments, and failures. And throughout, we continue to see possibilities and call on libraries to redouble their efforts to contribute to the massive digital library that can open doors to knowledge for students, scholars, and citizens of the world.

In this book, we examine different perspectives on this ideal future. In the first chapter, we trace the history of quests to provide broad access to knowledge and their relative success or failure in fulfilling the dream. We explore the print-based attempts to make scholarly resources more widely available; we follow with those efforts made possible first through automation and later with digital technology.

Chapter 2 goes into detail about the technologists' aspirations for digital technology. Brewster Kahle, researchers at Microsoft, and faculty at Carnegie Mellon University, in particular, had firm notions of societal changes that technology could enable.

Google and its brash rhetoric burst on the scene in chapter 3. Two brilliant computer scientists begin to make the case for a universal digital library. Google was new and not that well known when Sergey Brin and Larry Page first made this argument, and it was frequently met with skepticism. But they had financial resources and they worked fast. Google became a force to be reckoned with.

Chapter 4 deals with the public's expectations for access and how enthusiastically Google's announcement of plans to digitize books was received.

In chapter 5, librarians and scholars begin to organize to respond to the threat or the opportunity of Google. Some of the initiatives were short-lived, but others have had a transformational effect on the nature of scholarship and recorded knowledge.

The lawsuit and the aftermath of the Google settlement are the centerpiece of chapter 6. How did the case develop and why did the proposed settlement fail? More importantly, what opportunities were missed and, now in hindsight, what have been the lasting effects of the Google book digitization initiative?

In chapter 7, we trace some of the efforts to fill the void after the Google project. We examine the possible role HathiTrust may be able to play in building a universal collection.

In the final chapter, we make our own observations about what book digitization in particular and other efforts to provide digital access to scholarly information more broadly have contributed to universal access. Where has there been progress? What else remains?

Finally, in an epilogue, we acknowledge the many changes that emerged in the COVID-19 era, when a greater reliance on technology became a principal strategy for protecting public health, not least in the provision of library services. Though faint, a picture of the future of libraries begins to come into focus.

In addition to capturing an important aspect of scholarly history, we raise a lot of questions about the digital future for the scholarly and information communities. We expect—or at least hope—that university administrators will engage their faculty in discussions about the implications for scholarship, teaching, and the broader public good. And library leaders will renew their efforts to complete the digital agenda that Google started more than a decade ago.

[1] See, for example, Siva Vaidhyanathan, *The Googlization of Everything (And Why We Should Worry)* (Berkeley: University of California Press, 2012); Ken Hillis, Michael Petit, and Kylie Jarrett, *Google and the Culture of Search* (New York: Routledge, 2013); Ken Auletta, *Googled: The End of the World as We Know It* (New York: Penguin, 2009); Amy Langville and Carl D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings* (Princeton: Princeton University Press, 2012); Jean-Noël Jeanneney, *Google and the Myth of Universal Knowledge: A View from Europe*, trans. Teresa Lavender Fagan (Chicago: University of Chicago Press, 2007); and Elad Segev, *Google and the Digital Divide: The Bias of Online Knowledge* (Oxford: Chandos, 2010); as well as broader treatments such as Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018); Christian Vandendorpe, *From Papyrus to Hypertext: Toward the Universal Digital Library* (Champaign: University of Illinois Press, 2009); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019); Evgeny Morozov,

other libraries in the network to borrow materials from their collections. Even though the process was time-intensive, it gave scholars and researchers an opportunity to request materials that would otherwise be unavailable to them.

As time went on, groups of libraries based on size or type or region entered into interlibrary loan agreements that facilitated speedier delivery. In the 1960s, interlibrary loan took on added significance when the Ohio State University Library and others in the state began to create what amounted to a single library system. State funds supported the development of the Ohio College Library Center, a collaborative that presupposed interdependence of libraries in the state, and borrowing materials from one another was a chief benefit. As we will describe in later sections, this "one library" concept, coupled with automation of bibliographic records, gave rise to OCLC, Inc., now a worldwide collaborative that supports thousands of library members.

The Farmington Plan

When Germany invaded Poland in 1939, it was a wake-up call for all research librarians. The great treasures of European libraries upon which so many American scholars depended for their research were threatened with massive destruction. High-level meetings of representatives from the Library of Congress, the American Council of Learned Societies, the Social Science Research Council, the Board on Resources of American Libraries, and the Association of Research Libraries met, beginning in 1939, to develop a plan to ensure access to scholarly resources. The consensus of this group was that library groups and learned societies should develop desiderata of European materials that should be microfilmed by the Library of Congress and made available to the research community. They also concluded that it was essential that the Library of Congress complete the work on building a National Union Catalog so that libraries across the country could locate the research holdings that they might wish to borrow from one another. Wars and natural disasters could easily erase the accumulated knowledge in major research libraries, and American research libraries agreed that they should take responsibility for preservation.

On October 9, 1942, the Executive Committee of the Librarian's Council of the Library of Congress met in Farmington, Connecticut (thus the name of the plan), to discuss next steps.[4] The emerging plan called for a comprehensive collection of currently published materials with individual libraries accepting cooperative responsibility based on subject

Library of Congress to acquire, index, abstract, and deposit library materials from designated countries. This plan to gather scholarly resources from other parts of the world with federal funds proved to be a far more popular method than a voluntary, self-funded Farmington plan.[7]

The Center for Research Libraries

The aftermath of World War II transformed major research libraries in other ways as well. The GI Bill brought thousands of new students to campuses across the country, and the U.S. government began investing heavily in research and knowledge creation, recognizing that the insular approach of the prewar years could not be repeated. Libraries grew rapidly, and space for collections was a problem. Collections coming from the Farmington Plan or PL 480 had to be retained in the national interest, but they were infrequently used on the local campus.

Again, research libraries sought a cooperative solution to the space problems created by building research collections "just in case" they were needed one day. Ten midwestern universities developed a partnership in March 1949 to create the Midwest Inter-Library Corporation (MLC) that allowed participating institutions to send their materials that were little used, but still had research value, to be stored and retrieved when needed from the MLC. The cost was shared based on a formula of the library's acquisitions budget and its university's number of doctoral programs.[8]

In the early 1960s, the MLC became a national organization, the Center for Research Libraries (CRL), under the leadership of Gordon Williams, who launched several national programs in collaboration with the Association of Research Libraries to collect foreign newspapers as part of a permanent, shared collection. CRL also worked with the National Science Foundation to identify and collect international scientific journals for the benefit of the broad scholarly community.

From Scholar to Manager: Change Comes to Research Libraries

Academic librarians pursued broadened access because they worked closely with scholars and researchers. Particularly after World War II, when plentiful research dollars led to an explosion of published literature, scholar-librarians recognized that there was no hope of building local collections that would meet the needs of their faculty and graduate students. While library budgets expanded in the postwar period to provide for better coverage of the research literature, it was clear that