

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

The Open Book: Digital Form in the Making

Permalink

<https://escholarship.org/uc/item/1px1s12t>

Author

Murrell, Mary

Publication Date

2012

Peer reviewed|Thesis/dissertation

The Open Book:
Digital Form in the Making

By

Mary E. Murrell

A dissertation submitted in partial satisfaction of the

Requirements for the degree of

Doctor of Philosophy

In

Anthropology

In the

Graduate Division

Of the

University of California, Berkeley

Committee in charge:

Professor Paul Rabinow, Chair
Professor Charles K. Hirschkind
Professor Corinne P. Hayden
Professor Mary C. Gallagher
Professor Ramona Naddaff

Fall 2012

The Open Book: Digital Form in the Making

© 2012 by Mary E. Murrell

Abstract

The Open Book

by

Mary E. Murrell

Doctor of Philosophy in Anthropology

University of California, Berkeley

Professor Paul Rabinow, Chair

This dissertation attempts to put anthropology in conversation with “the book.” It does so through an empirical specification of mass book digitization, the industrial-scale, retrospective conversion of books into digital form. Such mass digitization is the work of research libraries, Internet companies, non-profit organizations, national governments, and the computer scientists, digital librarians, and the lawyers and administrators who advise, encourage, and support them.

The dissertation approaches mass digitization as a venue for plumbing the turbulent waters of what I consider the “contemporary book”: an arena of experimentation arising from the productive, seismic encounter of the modern book apparatus with an emergent assemblage in motion around the production, authorization, storage, preservation, circulation, and production of knowledge. By the “contemporary book,” I refer to the modern book apparatus (*dispositif*) as it is being worked upon, reconfigured, and called into question in the early twenty-first century.

The activities of the digitizers I studied are based upon a shared conviction that the book and its institutions are “closed” and need to be “opened”: books are “inaccessible” or “locked up” by virtue of their materiality (their printedness); by the institutions that store and keep them (physical libraries); and by the state’s often misguided regulation (copyright). For them, mass digitization is an important development in moving from the closed nature of the modern book apparatus to a remediated knowledge infrastructure, a future book. If it could be achieved, mass digitization would be a breakthrough in the long-anticipated opening of the book.

Primary research for the dissertation was conducted at the Internet Archive in San Francisco, CA. Additional research took place in a variety of locations in the United States, in specific relation to the public controversy around the Google Book Search Settlement.

**The Open Book:
Digital Form in the Making**

List of Illustrations

Acknowledgements

Introduction: Opening the Book

Chapter 1: Pathways to Digitization

Chapter 2: The Matter of the Digitized Book

Chapter 3: Books as Data

Chapter 4: Books as Orphans

Conclusion

Works Cited

List of Illustrations

Chapter 1

Figure 1: Card files of Universal Bibliographic Repertory, Brussels.

Figure 2: Card files of the Universal Bibliographic Repertory, Brussels.

Chapter 2

Figure 1: The “red boxes” that used to form the Archive’s machine cluster. This photograph was taken in its former data center in downtown San Francisco. Courtesy of the Internet Archive.

Figure 2: Screenshot of the Jedi Library from *Star Wars: Episode 2: Attack of the Clones*. The ideal for Brewster Kahle’s redesigned machine cluster.

Figure 3: “Books that glow.” Part of the Archive’s new data cluster, situated in a niche in the back of the former sanctuary of the Christian Science church that now serves as the Archive’s headquarters in San Francisco. Photograph courtesy of the Internet Archive.

Figure 4. The list of files that makes up an Internet Archive digitized book as digital object. All are viewable or downloadable at <http://ia600404.us.archive.org/3/items/genealogyofmoral00nietuoft/>

Figure 5: The Internet Archive’s Scribe image capture machine. Photography courtesy of the Internet Archive.

Figure 6. A scanner operating the Archive’s Scribe image capture machine. Photography courtesy of the Internet Archive.

Figure 7. Drawing in Google Patent depicting their infrared page de-warping system.

Figure 8: An inadvertent image of a Google worker’s hand. “The Inland Printer – 164.” Courtesy of Andrew Norman Wilson.

Figures 9 and 10. Stored, “raw” images of pp. 114 (top) and 115 (bottom) of *Genealogy of Morals*. Figure 11 below will show these images after rotating, de-skewing, cropping, and reassembly into a facsimile of the book. I have rotated them 90 degrees here.

Figure 11: Page image highlighting the search term—“bribery”—located through a search carried out on encoded OCR text.

Figure 12. A view of pages 114-15 of *genealogyofmoral00nietuoft* in the Archive’s Book Reader, derived from Figures 9 and 10 (above).

Figure 13: Roughly the same two pages from Google’s *Genealogy of Morals* as in figure 12 (above), though this a different English translation.

Chapter 4

Figure 1: A snippet from the 1954 English translation of Marcel Mauss’s *The Gift*. The search keyword was “hau.”

Acknowledgements

The following organizations have financially supported the research and writing of this dissertation: the Anthropology Department of the University of California, Berkeley; the Graduate Division of the University of California; the National Science Foundation; the Doreen B. Townsend Center for the Humanities at UC Berkeley; the University of California Systemwide Humanities Network Society of Fellows; the American Council for Learned Societies; and the Andrew W. Mellon Foundation.

I have incurred a great debt to Brewster Kahle of the Internet Archive for his openness, friendship, and trust. During my time at the Archive (and since), June Goldsmith was also a constant source of help and good cheer. Other employees of the Internet Archive, all of whom I have kept anonymous in the dissertation, were more than generous with their time and expertise. Without their help, this dissertation would not have been possible.

I am also especially grateful to Pamela Samuelson who was gracious enough to include me in various meetings around the Google Book Search Settlement in 2009. She also allowed me to audit her course in the Berkeley Law School on the Settlement in the spring of 2010.

Thanks are also due, in no particular order, to many people who took the time to help me understand their areas of expertise: James Grimmelman, John Wilkin, Paul Courant, Jon Orwant, Rick Prelinger, Karen Coyle, Clifford Lynch, Bob Stein, Ben Vershbow, Dan Visel, Raj Reddy, Tom McCarthy, Don Waters, Gloriana St. Clair, Denise Covey, Michael Lesk, Ivy Anderson, Steven Rosenberg, Laura Brown, Roger Schonfeld, Kevin Guthrie, Paul Conway, Kenneth Crews, Anne Kenney, Sanford Thatcher, Suzanne Calpestri, Tom Leonard, Gregory Crane, Mark Weber, Roy Tennant, Jason Epstein, Whitney Dorin, Juliet Sutherland, Carole Moore, Fred von Lohmann, Bill Janssen, Jessica Litman, Josh Greenberg, George Kerscher, Peter Brantley, John Ockerbloom, Joe Esposito, Colin Koopman, Steven Hammersley, Paul Aiken, Mike Shatzkin, and the Read2.0 list members. Although I spoke to him only in passing—in lobbies or corridors—I need to acknowledge the benefit I gained from the tireless and spirited spokespersonship of Daniel J. Clancy, formerly of Google Book Search. Again, I hope I have not misconstrued or misapplied the expertise of any of these good folk in what follows. If I have, I beg their forgiveness.

My greatest thanks go to the supportive members of my faculty committee, each of whom has made distinct contributions to my thinking as well as gracious accommodation to my shortcomings: Charles Hirschkind, Cori Hayden, Ramona Naddaff, and Catherine Gallagher. Each has offered me the example of his or her scholarly integrity and collegiality. Above all, however, I owe an ever-growing debt

to Paul Rabinow for his uncompromising intellect, his extraordinary example, and, most importantly, his support and care. This dissertation only begins to scratch the surface of the gifts he has given me.

Introduction

Opening the Book

“A book is ... a mute space of unrealizable dreams and manifest desire for form.”

-- Johanna Drucker

For many decades the future of the book has been worried over. Whether elegiac or celebratory, the observations of scholars, artists, librarians, journalists, and others have presented the fate of the book as a threshold for humankind, the immense significance of which can be assumed if not specified. From the foundations of media studies—whether Walter Benjamin (1928, 1935) or Marshall McLuhan (1962)—the book has provided a foil for “modern,” “mass,” and “new” media. It is made to characterize an epoch before and never quite a part of an ever-modernizing modernity. The computer engineers, digital librarians, lawyers and activists—who form the group of people on which this dissertation focuses and whom I term “digitizers”—are designing this post-book epoch, which is both already here and yet never quite arriving. Rather than focus on rupture, however, they are specifically focused on forms of continuity, namely, the conveyance of the past into the future. Books, to them, are not a mere archaism but a crucial component of a digital future in that books, when aggregated, represent and convey the past as no other cultural artifact or form does. To some digitizers, books contain “knowledge”; to others they contain “information” but, in every case, books are highly esteemed repositories that need to be adapted for a future in which electronic mediation will be dominant. The research topic of this dissertation—mass digitization—is the means by which the legacy that are books will be transferred from one epoch into another. Digitizers are a bold techno-elite who assume a vanguard position of cultural stewardship.

“Digitization” is often used in a processual sense, referring to broad changes in a culture, economy, or in a specific medium (film, music, photography, etc.) as its production and distribution shift from analog to digital technologies. This dissertation, however, does not seek to describe or explain any such large-scale *process*; instead it is an empirical specification of digitization (mass book digitization) as a contemporary *practice* of uncertain outcome. It is a fairly recent undertaking in which a relatively small number of organizations are involved. By *mass digitization* I refer to the industrial-scale, retrospective conversion of books into digital form. It is “mass” in that it involves not hundreds or even thousands but millions of books. Its scale is “industrial” in that it requires very significant capital investment, large data management and workflow systems,

expensive customized scanning equipment, and a great deal of human labor. It is “retrospective” in that it specifically involves already existing print books such as those that fill library shelves. Such mass digitization is the work of research libraries, Internet companies, non-profit organizations, national governments, and the computer scientists, digital librarians, and lawyers who run, advise, encourage, and support them. Although mass digitization is occurring in many locations in the world—China, the Netherlands, France, England, Egypt, India, Japan, to name a prominent few—this dissertation is limited to mass digitization practices in the United States for reasons of manageability but also because, despite earlier efforts in France, mass digitization as the event I take it to be has emanated from the United States.

I came to the topic of mass digitization while designing an anthropological investigation into “the book”—that complex, dense and saturated cultural object of central importance not only to Western modernity but also to my own life. As a student, an avid reader, a professional editor, an undisciplined consumer, and now as a scholar and teacher, books have engaged me—and I have engaged them—in varied ways through most of my life. Before entering the doctoral program in socio-cultural anthropology at the University of California, Berkeley, I had worked in book publishing for many years as an acquiring editor. In the course of that career, however, it slowly became apparent to me that the centrality of the book in U.S. culture broadly was diminishing and that the role of the publisher—as was the role of the editor within it—was narrowing. What was worse was the sense I had of a *futurelessness* to the enterprise, the lack of a horizon. Of course, to many (mostly outside of book publishing) there was a beckoning horizon and it was electronic. Yet when the half-hearted suggestion was floated that my colleagues and I might need to consider acquiring electronic titles or projects, I would roll my eyes dismissively as I winced inwardly with a sort of grief. The angst of futurelessness could sometimes feel as if it was not only a condition of the book or the industry that supported it but a condition of my own. Perhaps, in retrospect, it should be no surprise that I ended up in California doing research among people so *futureful* that I find myself here, in this dissertation, seeking a middle ground.

After arriving in Berkeley in the fall of 2005, I had three formative experiences. First, I found myself in conversation with a variety of people who were involved in digital publishing initiatives, library digitization, and Web development. In these conversations, I perceived an atmosphere of intensity around books and their futures that I had not felt in New York, the center of the U.S. publishing industry. Simultaneously, I confronted the almost obscenely abundant “information environment” enabled by the extraordinary library at Berkeley and the University of California system generally. During the years I was busy publishing books, the measure of a great library had become not how many volumes were shelved in its library, but how many “resources” and “services” one could take advantage of without ever having to go to the library (or even to change out of one’s pajamas). I

had not been a full-time student for twenty years, and, among the many adjustments I had to make, the future shock that my privileged, local “infotopia” brought was no doubt the least difficult. Finally, I found myself among a group of scholars in and about the anthropology department charting what an “anthropology of the contemporary” could be. How, I was prompted to consider, might the book become an object of anthropological inquiry? And how might one approach the book as a contemporary problem space?

A year earlier Google had announced its mass book digitization project (now called Google Book Search). The project was a shock to what I will be calling here the “book apparatus,” and responses to it were forming around me. The University of California system was joining a competing digitization effort to Google’s—called the Open Content Alliance—to begin digitizing its print book collections en masse (Hafner 2005).¹ Between that effort and Google’s, some of the world’s largest libraries were stepping forward and allowing their books to be digitized. Much has happened around digital books in the meantime but before the Sony Reader, the Amazon Kindle, the Barnes and Noble Nook, the Borders bankruptcy, the *New York Times* e-book bestseller list, the I-pad, Google Play, or iBooks—or, it need be added, any major digital initiative from within the book publishing industry—Google’s catalytic move in 2004 made many involved in and around books believe that some long-anticipated shift had occurred or was just about to. John Wilkin, Associate University Librarian at the University of Michigan and one of the key actors in mass digitization, referred to December 14, 2004, the day that Google announced its project, as “the day the world changes.”² To many, whether pleased about it or not, a moment seemed to have finally arrived when books—*all books*—and their ancillary institution, the library, would become *digital*. An anticipated but ambivalent threshold had arrived. This dissertation situates itself within this threshold.

More formal than informal, more elite than “folk,” more Western than non-Western, the book has not coincided with the predilections of the discipline, which long ago distinguished itself once upon a time from history by the presence or absence of written documents among those studied. In the broad history of writing, often a context for discussing the history of the book, anthropologists have had more to say, such as the infamous “literacy thesis” put forth by Jack Goody (with literary historian Ian Watt), which argued that the development of logical thought depends

¹ The following August (2006), the University of California would become a Google partner, joining Stanford, Michigan, Oxford, the New York Public Library, and Harvard.

² MSNBC, “Google to Scan Books from Major Libraries,” December 14, 2004). <http://www.msnbc.msn.com/id/6709342>. Wilkin continued: “It will be disruptive because some people will worry that this is the beginning of the end of libraries. But this is something we have to do to revitalize the profession and make it more meaningful.”

on writing and that skepticism, “higher psychological processes,” and critical thinking result from literacy (Goody and Watt 1963). That essay along with a spate of work at the same time on “orality vs. literacy”—such as that by classicist Eric Havelock (1963), media critic Marshall McLuhan (1962), and literary historian Walter Ong (1958; 1982)—contributed to the mid-century creation of “the book” as a new scholarly object, but in anthropology it for the most part elicited a counter-literature exploring literacy as a widely divergent, contextually specific set of practices that operate within a standardizing, normalizing disciplinary power, against which groups and individuals often resist (e.g., Heath 1980; Street 1993; Boyarin 1993).

This perspective on writing as a technology of domination was perhaps most forcefully expressed by Levi-Strauss in *Tristes Tropiques*. There, after witnessing a chief’s attempt to wield the “power of writing” by scribbling meaninglessly onto a sheet of paper, Levi-Strauss reflects on writing and/or literacy. Dismissing any correlation between the existence of writing and profound civilizational changes, such as those presented by Goody and Watt, he wrote:

The only phenomenon with which writing has always been concomitant is ... the integration of large numbers of individuals into a political system, and their grading into castes or classes. The primary function of written communication is to facilitate slavery... [And] if we look at the situation nearer to home, we see that the systematic development of compulsory education in the European countries goes hand in hand with the extension of military service and proletarianization. The fight against illiteracy is therefore connected with an increase in governmental authority over the citizens. Everyone must be able to read, so that the government can say: Ignorance of the law is no excuse” (1955, 299-300).

In contrast to every platitude about access to knowledge promoting democracy, Levi-Strauss asserts that the development of literacy among the citizens of then newly independent states only encouraged “thinking in slogans that can be modified at will” so as to make people “easy prey to suggestion.” Access to knowledge in libraries would not, he argues, arm the liberated subjects but make them newly vulnerable to “the lies propagated in print” (300). Subsequent generations of anthropologists have considerably expanded and complicated this colonialist critique, turning the archive from a *source* into a *subject* (Dirks 2002; Stoler 2009). Expanding out from the colony and the colonial archive, anthropologists have also considered written documents as “artifacts of modern knowledge” in bureaucratic agencies, prisons, and universities (Riles 2008).

In more direct reference to matters “digital,” Arturo Escobar has, with the orality vs. literacy debate in mind, called for anthropological attention to the “hypothesized transition to a postscriptural society effected by information technologies” and to

“new ways of thinking determined by the operational needs of information and computation” (Escobar 1994, 219). This call for attention to “momentous” changes that such a transition would pose to theoretical and hermeneutical knowledges, including anthropology, comes much closer to the concerns of this dissertation than the literature on literacy. However, its horizon is still too far out, or at least beyond that which I have focused on here. It is, of course, the case that certain of my informants refer to the contemporary moment as one of “the transition” of books from one epoch to another, but that other the “new ways of thinking” that would arise from such a transition—should it come to pass—are before us not in front of us. My work shares some overlap with anthropologists of media interested in the “materialities of communication” (e.g., Larkin 2008). With regard to digital media, in particular, recent anthropology inquiry into free and open software overlaps with my work—some of the same people, places, and motivations appear and overlap—and I consider them sister projects, even if our central concerns diverge (Kelty 2008; Coleman 2012).

But in the broad field known as the history of the book—which has branched out to include contemporary book studies and which is dominated by historians and literary scholars but also includes bibliographers, sociologists, librarians, legal scholars, and art critics—cultural anthropologists are largely absent.³ Indeed, from Febvre and Martin’s *The Coming of the Book* (1958) to Ted Striphas’s *The Late Age of Print* (2009), the best example of an “anthropology of the book” I have found is by a historian of science. I speak of Adrian Johns’s *The Nature of the Book: Print and Knowledge in the Making* (1998). Johns’s book explores how the modern book came to be a form for the reliable circulation of authorized knowledge in early modern England. Challenging a technological determinism that has construed the printing press itself as “an agent of change” (Eisenstein 1979; McLuhan 1962), he shows how the qualities of fixity and reliability did not inhere in the print itself or in the “printing press” but rather accrued to the material instrument of the book through the coordination and rationalization of heterogeneous elements—regulatory decisions, administrative measures, bodily discipline, moral propositions, and legal instruments. In my reading, the book demonstrates how the printed book transformed into a *dispositif*, or the modern book apparatus.⁴ It is Johns’s particular focus on the contingencies of a *form in the making* that attract me to his book and leads me to label it an “anthropology.”

³ I need note, however, Andrew Lass’s work on “comparative” libraries (Lass 1999).

⁴ The English word “apparatus” conflates two French terms—*dispositif* and *appareil* (instrument or tool)—and the ambiguity of the single English term suits me because it captures how material instruments (*appareils*) transform into epistemological figures (Geohegan 2011, 99). Although Johns does not use the term or concept *dispositif*, he shows the modern book to be such an epistemological figure.

This dissertation attempts to put anthropology in conversation with “the book.” A step toward doing that is to put John’s historical work in conversation with the “anthropology of the contemporary,” formulated by Paul Rabinow and colleagues. In the use of the term *apparatus* above, I am following Rabinow who draws upon Michel Foucault’s use of *dispositif* (Foucault 1978; 1980; 2007)⁵ to develop his own definition: “Apparatuses are stabilized forms composed of heterogeneous objects that bring multiple aspects of domains together and set them to work in a regulated functional manner” (Rabinow and Bennett 2012, 61). Well known terms such as “bourgeois tradition” (Williams 1977); the “order of books” and *le circuit du livre* (Chartier 1994, 2004); “print culture” (Eisenstein 1979); “discourse network” (Kittler 1990; 1999); and the “modern literary system” (Hesse 1996) also attempt to describe domains working together in a regulated manner. However, it is the remainder of the definition that sets the term apart for me: “Apparatuses are long standing, long enduring specific responses to particular dimensions of larger problematizations.” When we see the modern book apparatus is then a stabilized response to an *earlier* problematization within European modernity, we can begin to see the developments I describe here as symptomatic of a *new* problematization, which it is too early to discern but to which is new generating responses. Such an approach to the book also importantly diverts us from a too-great fixation on “technology,” which leads to uninteresting back-and-forth about whether and when the printed book will be superseded by electronic forms.

So, put another way, in this dissertation I approach mass digitization as a venue for plumbing the turbulent waters of what I consider the “contemporary book”: an arena of experimentation arising from the productive, seismic encounter of the modern book apparatus with an emergent *assemblage* in motion around the production, authorization, storage, preservation, circulation, and production of knowledge (Rabinow 2003, 55f). In expanding Foucault’s “history of the present” mode to a “contemporary” mode, Rabinow directs us to look for such assemblages. He defines an assemblage as a “nascent organizational form that attempts to identify and associate elements from diverse domains (e.g. law, technology, government, media, science, spatial arrangements, etc.), in response to events that signal the insufficiency and discordancy of previous apparatuses in relation to emergent problems” (Rabinow and Bennett 2012, 62). Thus, by the “contemporary book,” I refer to the stable apparatus (Johns’s “the nature of the book”) *as it is being worked upon, reconfigured, and called into question* in the early twenty-first century.

⁵ Hurley (1978) translated *dispositif* as “deployment” in *History of Sexuality* (“Part 4: The Deployment of Sexuality), whereas Burchell (2007) chose “apparatus” (Opening Lecture, “The Security Apparatus”).

This emergent assemblage is certainly bigger and more complex than what I will discuss in this dissertation, but mass digitization, I contend, is an important element within it. It has been catalytic in a number of ways. Promulgated through a confluence of interests between research libraries, universities, and Internet-oriented technology companies or organizations, mass digitization operates if not entirely from outside of the book apparatus then at least from its edges. It is a controversial, contested, and fundamentally experimental practice of uncertain outcome. It straddles the old and new: the digitized book is a distinct new object with emergent properties, functions, and uses, especially with regard to its “machine” readability, but, at the same time, it is also “merely” a new type of copy, modeling and drawing upon the older familiar form of the book. (What kind of copy it can be, and how it will function as a copy, is a matter of dispute and maneuver.) And, finally, mass digitization efforts have become a flashpoint of controversy, attracting involvement from a wide variety of actors. The Google Book Search Settlement, for instance, elicited action from all branches of the U.S. government: the Copyright Office, Congress,⁶ the Department of Justice⁷; and the Federal judiciary. In sum, mass digitization has been a key vector in the problem-space of the contemporary book and, as such, it provides me with an empirical point of entry into it.

Inquiry: Here, Now, Unfolding

For nearly three decades, anthropologists have been grappling with how to adapt their research practices to both conceptual developments within the discipline and to contemporary problems beyond it (e.g., Clifford and Marcus 1986; Gupta and Ferguson 1997; Marcus 1999; Ong and Collier 2003; Rabinow and Marcus 2008). This literature has both guided and formed me in this study of timely events unfolding not simply in the West, or the U.S., but in the very institutions that employ, house, and engage me. How is one’s anthropological engagement simultaneously political, ethical, and conceptual? What constitutes “fieldwork” or inquiry? How does one anthropologist situate herself within the “writing machine” of such a contemporary milieu: the journalists, the multiple experts and their

⁶ Congressional hearing, Committee on the Judiciary, “Competition and Commerce in Digital Books.” September 10, 2009. Transcript and audio tape available at http://judiciary.house.gov/hearings/hear_090910.html

⁷ The Department of Justice Anti-trust Division made two highly influential filings in relation to the Google Book Search Settlement: 1) In response to the original settlement (“Statement of Interest of the United States of America Regarding Proposed Class Settlement”, September 18, 2009) – available at <http://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2005cv08136/273913/720/>; and 2) to the amended settlement (“Statement of Interest of the United States of America Regarding Proposed Amended Settlement Agreement”, February 4, 2010). Available at: <http://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2005cv08136/273913/922/>

overlapping networks, the websites, list-servs, conferences, white papers, reports, legal proceedings, and gossip? How does one negotiate the relationships that constitute anthropological inquiry? And, a question rarely posed, what role does a dissertation play within the overall project of a contemporary anthropology?

This dissertation is a “writing up” of some key episodes or points of entry into the much broader and complex field of activity that I am referring to as “mass book digitization.” It is not a full account (historical, political, legal, or technological) of book digitization, nor is it an “ethnography” or site study of the mass digitizing organization that I focus on. Rather, it is an informed, partial “take”—or set of takes—on the enterprise, based on four years of involvement in and attention to its (ongoing) experimentations.

Mass book digitization, as I’ve said above, is an arena of experimentation arising from the productive encounter among elements of the modern book apparatus with an emergent *assemblage*. I situated myself within the assemblage: that is, among the digitizers. A small number of organizations perform mass book digitization but a much larger network of people, whom I will refer to herein as “digitizers,” enables it. For instance, Google is the largest mass digitizing organization but its project could not have commenced without the consent and the continued partnership of large research libraries, and, quite specifically, the University of Michigan. Although many point to Google as a brazen corporation plundering the book system, in the case of mass digitization, Google is one actor (albeit an very important one) in a broad field of actors. John Wilkin of the University of Michigan captured this when he described Google to me as an “extraordinary vendor.” The Internet Archive is another mass digitizing organization and, though it does digitize books on its own premises, most of its digitization occurs in small-scale scanning centers situated inside libraries in the U.S., Europe, and beyond. In addition to the actual scanning of books, mass book digitization is enabled by a range of actors—lawyers, librarians, and others—who write, advise, consult, lobby, comment and so forth. To the extent that I judge their activities significant, I include all such individuals in my use of the term “digitizers.”

Because of this dispersed networked of actors, I conducted research in two different modes. First, in a more traditional mode, I situated myself in one node in the network of mass book digitization (the Internet Archive) over an extended period of time. From there, I branched out, at chosen intervals, to other locations, engaged in different activities among different actors, not to gain a comparative perspective on mass digitization, but as a continuation of my engagement at my main “field site.” My collaboration with that location connected me to others. But to speak of a “field,” with its suggestion of a clear delimitation in time or space, is inapt. As I suggested from the outset, this project began well before I had chosen a “site” of

formal research, and it is still not clear to me when (or even if) I have yet “left the field.”

The activities of the digitizers I studied are based upon a shared conviction that the book and its institutions are “closed” and need to be “opened.” Early in my fieldwork, I saw that this conviction was taken for granted, so self-evident as not requiring comment or explanation, even if, to me it was surprising. It was common sense: books are “inaccessible” or “locked up” by virtue of their materiality (their printedness); by the institutions that store and keep them (physical libraries); and by the state’s often misguided regulation (copyright). For them, mass digitization is an important development in moving from the closed nature of the modern book apparatus to a remediated knowledge infrastructure, a future book. If it could be achieved, mass digitization would be a breakthrough in the long-anticipated opening of the book.

The questioning of the form of the book extends back over a century and probably even further. Chapter 1, “Pathways to Digitization,” explores the activities and thoughts of a heterogeneous but interconnected group of men (no women) who saw the book and its institutions as inadequate, and even detrimental, to the intellectual needs of modernity and who sought to resolve the problems of the book through technological intervention. In particular, I investigate and chronicle the interest and enthusiasm for microphotography and microfilm, especially pronounced from the turn of the twentieth century up to the Second World War. By focusing on the period before the computer, I aim to work against a singularity too often attributed today to networked computers—that they are radically world-making and in the process of “changing everything.” This may be true, but the least we can do is to remove epochal presumptions from our analysis, so that continuities can be revealed and discontinuities made evident. Such is my method in this chapter: to pursue a resolutely “analog” history of digitization as an orientation to my inquiry into the contemporary problem space of mass book digitization. In so doing, I pursue what Lisa Gitelman has referred to as a “pre-digital” history of the digital: “an account of surprising continuities rendered against obvious and admitted discontinuities” (Gitelman, forthcoming; 2006). I pull out four key intersecting concerns: 1) the problem of too many books; 2) the inadequacy of the book and the library for science; 3) the limits that print publication exercised on accessibility and circulation; and 4) the perishability of modern paper. I follow the problem of perishable paper beyond the period of most of the chapter, into to the 1990s, charting a collective effort among research libraries to undertake the large-scale reformatting of books. Initially the impetus was to *preserve a minority* of deteriorating books but it changed over time into a mandate to *increase access to the majority* of books.

Part of the critique of the book, as noted above, is that it is imprisoned by its materiality (its printedness). At the same time, however, my work at the Archive

taught me that the digitized book is not simply a remediation in form that overcomes the book's constraints and limitations but also a re-materialization that itself produces new constraints and limitations as well as new conditions of possibility. In chapter 2, I provide an anatomy of the digitized book that enables the reader to grasp it not merely as a facsimile, model or surrogate of a printed book—a representation—but also as a new and different object that needs to be understood on its own terms. In order to describe the new book as a “digital object,” I adopt a three-pronged perspective, borrowed from digital preservation, that considers the digitized book as simultaneously a physical, logical, and conceptual object. Although I will provide considerable (and perhaps overwhelming) detail in this chapter about the physical materiality of the digitized book, the expertise and labor involved in making and maintaining them, and the infrastructure on which such digital objects depend, the point is not simply to change the terms of discourse—to redirect the tendency to think of digitization as a de- or immaterialization—but rather to show *how* precisely the materiality of the digitized book matters and what the re-materialization of print books into digital form produces. I continue this through the remaining chapters.

In addition to producing a facsimile reproduction of a book to be read online, digitization also creates a new dimension to the book: what I call in chapter 3 “books as data.” Digitization “datafies” books in a double conversion: they become data literally (as ones and zeros) and conceptually (as new sources of latent knowledge). It reconstructs and reconstrues library holdings into “data sets” or “text corpora”; words become “data points”; and machines become “readers” of books. In this chapter, I explore two things: 1) I describe the “ethos” of data that pervades the Internet Archive; and 2) how digitizers and their allies have reconceived books as containing new spaces within them that are out of the reach of their copyright. Referred to by a series of negations—non-consumptive, non-display, non-expressive—new machinic uses for the digitized book are poised to legitimate the legal “gray zone” of digitization. In the search for points of leverage, digitizers and their allies have stumbled upon this newly present “data” within books as a possible method of forcing their way through blockage of copyright regulation. Whereas chapters 1 and 2 explored varying aspects of the first two prongs of the critique of the book apparatus—materiality and institutions—chapter 3 begins to extend our analysis into the third prong—regulation—as it analyzes how “data” has become a strategic wedge with which to attempting to pry open the book apparatus.

Chapter 4 turns to another strategic wedge: the “orphaned book.” “Orphan” has a variety of meanings, but, broadly, it refers to a copyrighted work that someone would like to make use of but for which the owner cannot be consulted for authorization. “Orphans” are not strictly speaking “new,” but they have been brought newly to the fore by the desires of digitizers, libraries, archives, and the likes of Google. In order to show how mass digitization has put orphans into motion, the chapter traces the Internet Archive's direct involvement in the

expansion of the term from a local concern of film preservationists into a broader national effort at copyright reform. I then analyze the Google Book Search Settlement as yet another failed attempt to solve the orphan problem, and conclude with a recent pulling away from the “orphan” metaphor, which to some has misstated the problem. Orphans have been a means to critique expansionist U.S. copyright policy but they are more than that. The conflicts around them and around digitization—especially the Google Book Search Settlement—reveal a resilience within the book apparatus that argues against an easy belief in any inevitable process of digital evolution. In the Conclusion, I briefly reflect on the current state of mass digitization, as litigation proliferates and both Google and the Internet Archive pull back from their former investments. I end by recounting how my main informant and I have arrived at a similar interest and perspective on the future of mass digitization.

What follows almost to the end of this Introduction is a contextualization of my inquiry: stage-setting, background about my fieldsite, a brief chronicle of mass digitization from one point of view, and a broad outline of the research I conducted.

Sites of Inquiry

Mass digitization is most often associated with the search and advertising company Google. In October 2004 Google launched a service (then called “Google Print”) that enabled a user to search through the textual content of currently commercially available books. They called it a “virtual card catalog” in that it was like consulting a catalog of books but with the extra benefit that, in addition to the author, title, and other bibliographic information, you could also search through the entire book itself. Google Print wasn’t original. It resembled a service that Amazon.com had announced the year before, “Search Inside the Book,” which also digitized publishers’ books and made them searchable on the company’s website, although only one by one, not as an aggregated corpus of books. Then, a couple months later, in December 2004, Google announced a much bigger and more remarkable project, then called “the Library Project,” in which they would digitize parts of, but in some cases the entirety of, the collections of five major research libraries: the New York Public Library, Oxford’s Bodleian Library, Harvard University, Stanford University, and the University of Michigan. The project, which had been in the planning since at least 2002, had been kept so secret that, to a person, all those to whom I spoke about its announcement told me that it was a surprise (or even a shock) when they learned about it through the *New York Times* (Markoff and Wyatt 2004). One key difference between the two products—Google Print and the Library Project—was that of size. The former would be limited to books that were currently “in print” or commercially available from publishers, whereas the latter expanded Google’s efforts outward toward the age-old fantasy of a universal library.⁸ The

⁸ Google Print and the Library Project are now known today simply as “Google Book Search.”

company described its ambition thus: “Our ultimate goal is to work with publishers and libraries to create a comprehensive, searchable, virtual card catalog of *all books in all languages* that helps users discover new books and publishers discover new readers” (emphasis mine).⁹ Another key difference was that while Google Print proceeded with the permission of the copyright owners, the Library Project proceeded *without* permission. Although many books in the library collections would be out of copyright, the majority would be in copyright.

As an event, the Library Project structures the anthropological inquiry that is the basis of this dissertation. The December 2004 announcement produced a variety of responses that continue to have ramifications today, two of which are central to this dissertation. One response was the Open Content Alliance (OCA), a coalition of libraries, technology companies, and other organizations that came together to provide an alternative and competitor to Google’s Library Project. The Internet Archive spearheaded this effort. The Archive hosted the annual OCA meeting, implemented the digitization, recruited new members, nurtured collaborations, and more. The OCA was “administered” by the Internet Archive but “governed” by the member libraries.¹⁰ From the time of its founding in September 2005, the OCA cast itself into the role of a public-spirited alternative to Google’s book digitization project. Whereas Google’s was proprietary, closed, private, centralized, secretive, and a perceived threat to the library system, the OCA would be the opposite: open, non-commercial, de-centralized, public-spirited, and made up predominantly of libraries. That said, the OCA was most significantly funded by the software giant Microsoft, a fierce competitor of Google, before it withdrew its financial support in May 2008.

Happening at the same time of the OCA’s formation was another response to the Library Project: two lawsuits, one filed in September 2005 by the largest U.S. author trade group (the Authors Guild) and another the following month by the largest U.S. publisher trade group (the Association of American Publishers or AAP). Both suits charged Google with massive copyright infringement and were eventually consolidated into one massive class action suit, *Authors Guild et al. vs. Google*. Then, three years after their filing, and shortly after I began my research, the Authors Guild, the AAP, and Google announced an agreement, which they had been secretly negotiating for over two years. The extremely complex agreement sought to resolve their differences and end the litigation. To do so, the key negotiators¹¹ had to think through every facet of mass digitization,

⁹ www.google.com/googlebooks/library.html

¹⁰ <http://www.opencontentalliance.org/participate/>

¹¹ The Authors Guild’s in-house attorneys Jan Constantine and Paul Aiken; its outside counsel Michael Boni and Joanne Zack; the AAP’s Allan Adler and its outside counsel Jeff Cunard;

which is also an exercise in thinking through “the book” as it confronts its computerization. Google’s partner libraries were not among the negotiators, though Google reportedly consulted with them and attempted to represent their interests in the negotiations. The Settlement sorted out not only the authors’ and publishers’ concerns with Google but persistent nagging problems they have had among themselves: one of the Settlement’s fifteen attachments, the “Author Publisher Procedures,” hammered out without Google in the room,” contained a compromise between publisher and author representatives over electronic rights to titles published before such things were contemplated (i.e., before the mid 1980s).

The complexity of the Settlement could fill a dissertation (or two or three) itself and, indeed, it has spawned a voluminous legal literature that continues to grow (and to which I refer through the dissertation).¹² Before the Settlement, the original lawsuit had already attracted a good deal of scholarly and journalistic attention.¹³ With the Settlement, the case became even more complex and fascinating to close observers of digitization, copyright, digital libraries, and electronic books, but especially to copyright scholars. One copyright scholar has called Google’s book project and its legal travails “one of the most significant copyright developments of our time” (Tushnet 2012) and another has described the case as “the most interesting thing that’s ever happened in my field” (Samuelson 2009).

These two responses to Google’s Library project—1) the Open Content Alliance; and 2) the legal controversy and public contestation around the Google Book Search Settlement—formed my two main sites of inquiry, and the rough period of the Settlement’s announcement and the Settlement’s rejection (October 2008 to March 2011) marks the period of my anthropological inquiry into mass digitization. Thus, what I provide here is thus not an exhaustive treatment of mass digitization and its various ramifications, but a focused glimpse into mass digitization from one significant and specific location.

The Internet Archive is, by its own description, “a 501(c)(3) non-profit ... building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.” By my description it is an experimental

publishers Richard Sarnoff and other executives from the named company plaintiffs: Simon and Schuster; John Wiley & Sons; Pearson; and McGraw Hill; Google Book Search’s chief engineer Dan Clancy and product counsel Alexander MacGillivray; and Google’s outside counsel Joe Gratz and Daralyn Durie.

¹² Just the official legal filings are voluminous, but they have been made easily accessible at the Public Index (thepublicindex.org), a project of the New York Law School.

¹³ Charles W. Bailey, Jr. keeps an online bibliography: <http://digital-scholarship.org/gbsb/>

organization of the Internet. Brewster Kahle and his partner Bruce Gilliat founded the Archive in 1996 as the non-profit ancillary to the commercial company Alexa Internet, which provided Web traffic analysis. As part of its operation, Alexa regularly “crawled” the Web. A Web crawler is an automated program that methodically scans or “crawls” the Web to create an index of the data it is programmed to look for. Also called “spiders” or “bots,” crawlers are the basic technology of a search engine. At its beginning, the Internet Archive was essentially a “dark archive” made up on the web crawl data collected by Alexa Internet. “Dark archive” is a collection preserved for future use but to which no current public access is granted. It wasn’t until October 2001 that the Web archive was made publicly accessible, through an interface called the Wayback Machine (a playful reference to the WABAC, a time travel machine on the 1960s TV cartoon *The Rocky and Bullwinkle Show*). Through the Wayback Machine, a user can see a website as it existed in the past, or, to be more specific, as it existed on one of the dates on which a Web crawl was performed by Alexa. The Web archive and the Wayback Machine pioneered a method for making a good bit of the ephemeral web “enduring,” and thus to make it an accessible historical record. Although is not a “total” record of the Web, it nonetheless has been a boon to individual researchers, scholars, lawyers, and the government with an interested in Web-based information post 1996 (Rosenzweig 2007; Howell 2006; Nakashima 2008). The Internet Archive has been a leader in Internet preservation, including being the only non-governmental charter member of the International Internet Preservation Consortium, founded in 2003. Although the Wayback Machine is free to use, for-fee Web archiving services are an important part of the Archive’s “business model”: it provides “contract crawls” to libraries, governments, educational institutions, and others. In addition to the Web archive, the Archive acquires collections of varying types of “content”—live music, moving images, texts, and audio recordings—and in addition, allows individuals to upload their own “user-generated content” a la YouTube (which started in 2005, well after the Archive began its hosting services). From 2002, books have been part of the “text archive,” including collections such as Project Gutenberg the International Children’s Digital Library, and the Million Book Project (which I will discuss in some detail below), among others.

Beyond its collection and hosting activities, the Archive is an organization that self-consciously seeks to influence the shaping of both policy and practice around the World Wide Web. Its director sees himself as a warrior for the Open Web, and he focuses the Archive’s policy-type work on what he refers to as the “content layer.” He sees the Internet as having three distinct layers: at the bottom is the “physical” layer (the wires or “pipes” that link computers on the Internet); in the “middle” is the software or “logical” layer (the code that makes the hardware run, including Internet protocols); and, on the top, is the “content” layer (the actual “stuff” that gets transmitted such as the Archive’s text, audio,

and image collections).¹⁴ Kahle sees the history of the Internet as three wars, one corresponding to each layer. The forces for openness won the first two “wars of the Internet” (over the network and the software) in the 1980s and 1990s by and yet, as he sees it, pro Internet advocates remain mired in a continuing battle over the content layer. Although the Archive’s advocacy extends from expanding municipal wi-fi networks to resisting National Security Letters and intrusive state regulation of the Internet, my attention here will be limited to its specific activities related to books. Advocacy around books centers quite significantly on copyright but not exclusively as it also seeks to nurture an “open Web,” a decentralized system that has no central points of control. At the Archive these concerns have been most significantly confronted in the case of books. Part of why I describe the Archive as an “experimental institution” of the Internet is that it cultivates a supple responsiveness to events, taking on urgent problems and doing what it can, given its limited capacities, to address them.

In a sense, everything the Archive does is related to moving forward the development of “the future library.” Mass digitizers—a term I use to encompass all those spearheading mass digitization (library leaders, the Internet Archive, Google, various legal advisers, intermediaries, and others)—see themselves as architects and builders of future libraries. For Google and the Internet Archive, the point of central concern is the Web, as they are most organization fundamentally “of” the Web; for existing “traditional” libraries, the point of central concern is their community of users and their legacy collections—both of which are changing under the influence of Web-based technologies. Across all mass digitizers, the Web is understood to be the means by which culture will be recorded future of recorded culture and they see themselves as designing a proper infrastructure for doing so. The Web is young, still in formation, and they understand themselves, generationally, to be nurturing, maintaining, and shepherding it as they actively add “content” to it.

Although this dissertation will move across the differing activities of various actors, here I will set the stage for the detailed explorations in the subsequent chapters through a narration of Brewster Kahle’s (unanticipated) involvement in mass book digitization. Kahle narrates his professional life as having been motivated by wanting to do something positive with technology. From a young age, he says that he knew he wanted to “do good,” to make a positive contribution to the world, and to devote his life to “something big.” Kahle grew up in Scarsdale, New York, which, during his youth, was the most affluent suburb in the United States. Carol O’Connor titles her history of Scarsdale, *A Sort of Utopia*, to evoke it as a “capitalistic version of the ideal community” (O’Connor 1983). Kahle’s paternal grandfather had helped establish Standard Oil’s presence in Venezuela in the early

¹⁴ I presume that Kahle derives this scheme from Lawrence Lessig, who in turn derives it from Yochai Benkler (2001, 23).

twentieth century. Kahle's father (Robert) also worked for Esso, a Standard Oil subsidiary, as a research engineer, and later also negotiated the reparations from the nationalization of the Venezuelan oil industry in 1976 and continued to manage Esso's relationship with Venezuela thereafter.¹⁵ Like his father, Brewster became an engineer, though a computer engineer (at MIT) rather than a mechanical engineer (at Cornell).

One day in 1979, as Kahle tells it, an MIT friend and classmate challenged him, while they walked across the Charles River Bridge from Cambridge to Boston: "You're an idealist. Tell me something you can do with your technology that's actually good. What is a positive future with technology?" The nineteen-year old had to ponder it and recalls that it was a "harder puzzle than one would think." He came up with just two possibilities: cryptography (for the protection of people's privacy) and building the Library of Alexandria. The latter he thought was "too obvious." The idea had been around a long time, and others would no doubt take that "big" task on. So he pursued cryptography in the hopes of protected people's privacy, until he realized that anything he made would be at a cost out of the reach of the average "working man," and thus of use only to governments, the military, or corporations. So he went back to the idea of the Library, which was "one of those things we had promised to the world and hadn't done yet." Kahle hadn't read Vannevar Bush's "As We May Think" or J.C.R. Licklider's *Libraries of the Future*, nor had he been (then or since) a reader of science fiction. The idea of the Library was just "one of those things" that had been imagined as possible and that had informed and inspired computer engineers for some time. It had, Kahle says, "always been the idea, specifically of the computer science guys, to digitize the Library of Congress." The Library from college on became his life's work, indirectly while he pursued his business career and directly once he devoted himself full-time in 2001 to the Internet Archive.¹⁶ When questioned, Kahle's library becomes synonymous with the Internet itself. As he told me: "I think of the Internet . . . in general, as the Library. So I'm trying to build the Library. I'm not trying to build it in a little building. It's got to be a big, distributed, joint thing." Daniel J. Clancy, the chief engineer and executive in charge of Google Book Search during my fieldwork, uses precisely the same language: "[T]he Library of Alexandria is this thing we call the Internet We are seeing it being created today, and it will be created as a distributed entity. . . . Google Book Search . . . is a repository of *some* content that contributes to this broader initiative."¹⁷

¹⁵ Robert Vinton Kahle. May 2, 2001. <http://www.nytimes.com/2001/05/02/classified/paid-notice-deaths-kahle-robert-vinton.html>

¹⁶ I provide details about Kahle's early career in chapter 3.

¹⁷ Daniel J. Clancy, Engineering Director, Google, Presentation at "The Google Book Settlement: Implications for Scholarship" conference (May 7, 2010), available at http://www.acls.org/publications/audio/annual_meetings/2010/default2.aspx?id=5584.

At the beginning, however, books were not of particular interest to Kahle. He assumed someone else—publishers or university libraries—would take care of books. Instead, he focused more on “born-digital” material, most obviously the archiving and preserving of the World Wide Web (Kahle 1997). But to think about how to archive the Web was also to think about the infrastructure necessary to make that possible. In 1998 he and one of the early Internet Archive board members, political theorist and former university librarian Peter Lyman, called for a “critical mass digitization project” that could further the goals of digital libraries in general:

While the Web is often the information resource of first resort, print publication has been the medium of record for quality, and print libraries are far more comprehensive and reliable. If a large-scale library were to be scanned and offered on the Internet, then we would be confronted with a potential testbed for new models of copyright and royalties and might then develop new economic models for digitization of print (Lyman and Kahle 1998).

A “critical” mass digitization might chart a future for digital libraries.

Such a project came along in 1998 in the form of the Million Book Project, spearheaded not by libraries but by the computer scientist and Turing Award winner Raj Reddy at Carnegie Mellon University. Twenty years or so Kahle’s senior, Reddy has had a lasting effect on Kahle; for instance, he credits Reddy with inspiring his dedication to the goal of “universal access to all knowledge.” At Carnegie Mellon Reddy had his own universal library project—the Universal Digital Library—with a specifically global vision that imagined the American public library system available on the Internet to everyone in the world. Its mission was “to create a Universal Library which will foster creativity and free access to all human knowledge. ...The result will be a unique resource accessible to anyone in the world 24x7, without regard to nationality or socioeconomic background.”¹⁸ Like Clancy and Kahle, Reddy’s universal library is coterminous with the Internet.

In 1998 Reddy appealed to Michael Lesk, then the program manager for the National Science Foundation’s division of Information Science and Intelligent Systems (part of the Computer Science directorate), for financial support.¹⁹ As recounted to me in an interview, Reddy told Lesk that “the Web was being given a hard time” because people believe “everything on Web is junk,” and that publishers weren’t doing anything out of fear and inertia to put “quality” information on the Web. Remember that the early Web was associated with

¹⁸ <http://www.ulib.org/ULIBAboutUs.htm#visionBkMark>

¹⁹ Kahle refers to Michael Lesk as the “father of digital libraries.”

misbehavior, unreliable information, and pornography. In 1999, computer scientist and critic Joseph Weizenbaum described it this way: “The Internet is like one of those garbage dumps outside of Bombay. There are people, most unfortunately, crawling all over it, and maybe they find a bit of aluminum, or perhaps something they can sell. But mainly it’s garbage” (Hafner 1999). Book digitization, Reddy thought, would solve both problems: it would provide high-quality “information content” for the Web and also “shake publishers up and break the bottleneck.” Google would later express its motivations in a similar fashion years later. In 2007, Google co-founder Sergey Brin explained: “We really care about the comprehensiveness of search. And comprehensiveness ... is about having the really high-quality information. You have thousands of years of human knowledge, and probably the highest-quality knowledge is captured in books. So not having that—it’s just too big an omission” (Toobin 2007, 31). Google executive Marissa Mayer provided a similar rationale: “If we provide access to books, we are going to get much higher-quality and much more reliable information. We are moving up the food chain” (Toobin 2007, 34).

The NSF’s Lesk, who shared Reddy’s general frustration, agreed: “We wanted to stir things up,” he told me. Although a book scanning project did not directly serve the NSF’s mandate to fund scientific research, Lesk and Reddy successfully construed the project as a type of “scientific infrastructure” that would provide a foundation for computer science research: “This 24x7x365 resource would ... provide an excellent test bed for further research in database creation, OCR technique, machine metadata creation, and textual language processing research such as machine translation, summarization, intelligent indexing, and information mining.” In 2001, the NSF gave the Million Book Project over \$3 million—Kahle was on the review committee—for purchasing Minolta flatbed scanning equipment to send to China and India, the governments of which, through the rallying efforts of Reddy, had joined the Million Book Project as partners and would provide and pay for the scanning labor. Although the main Million Book Project team was at Carnegie Mellon in Pittsburgh, PA, Kahle lent his time, interest, and resources. When the Project was having trouble convincing U.S. libraries to send their books out of the country, Kahle bought 100,000 de-accessioned books from the Kansas City public library system and donated them; and, later, when money was running low, he provided servers and staff support. Some, though not all, of the books were included as part of the text collection of the Internet Archive. The Million Book Project began Kahle’s now long involvement with the mass digitization of books.²⁰

²⁰ Though the Million Book Project managed to scanned nearly two million books, it was beset with problems: finding books to send to the foreign scanning centers (whether persuading libraries to loan their books or obtaining copyright permission from publishers); custom inspection and delays; received scans back from their partners; and bandwidth problems that made it hard to serve the books. But for 30,000 available from the Internet Archive, at the time of writing, none of the MBP books are viewable on line. Nonetheless, those involved in the project consider it to have succeeded as a “proof of concept”—proving that “it was practical to scan books, OCR them, bring

By the early 2000s, Kahle had become an evangelist for mass digitization, not just of books but everything. In 2002, at the Library of Congress, he delivered for the first time what can best be described as a stump speech—usually simply entitled “Universal Access to All Knowledge.” He estimates he has since given the talk in one form or another “maybe 100 times.”²¹ The talk can vary in some details, depending on the audience, but the message and form is basically the same: there is no good reason why “we” aren’t digitizing everything and putting it on the Web. It is important even urgent that we do so; it is possible; and it’s affordable. Computer speed and storage presents “our” generation with a opportunity to build the new Library of Alexandria anew: this time it will not be geographically isolated and it will exist in redundant copies so that it can’t be destroyed. The reality, however, is that very little has been digitized, a situation he appeals to his various audiences to help rectify. He proceeds through a series of “media types” presenting simple calculations—which one journalist has dismissively described as “back of the envelope” (Stross 2008)—demonstrating how manageable and affordable digitization actually is. Books always come first in the presentation. How hard would it be to put books online?, he asks. His argument in answer goes thus: The largest print library in the world, the Library of Congress, has around 30 million volumes and a book is about 1 megabyte in bits (as straight ASCII text, no images), this means the Library of Congress book collection is about 29 million megabytes, or 29 terabytes, which with today’s disk storage fits on a bookshelf-sized rack of computer servers that costs about \$60,000 to build and maintain. {This \$60,000 figure has not changed in the past ten years.) That’s storage costs. In terms of digitization, it costs \$.10 a page to digitize a book. If we assume the average book is 300 pages in length, which amounts to \$30 a book. A ten million-book library would then cost \$300,000,000 million. Given that libraries spend \$3-4 billion a year on books, it is certainly conceivable that this money could be squeezed out of library collection budgets. His answer, then, to the question “how hard would it be?” is “Not that hard.” It is doable. The reasons that digitization is not happening at the scale it should are political, cultural, and legal not technological. The talk shares a lot with the typical TED talk: an out-sized idea; a simplifying tendency; and a feel-good, rousing appeal to the audience.

them to the Web, and have them searched. People didn’t believe it could be done or that it would be a good idea to do it.” When Google announced its book-scanning project in 2004, the team felt that “its vision had won.” Google credits MBP on the Google Books website <http://books.google.com/googlebooks/history.html>

²¹ Here are some online examples of the talk: at the Long Now Foundation (2011): <http://archive.org/details/brewsterkahlelongnowfoundation>; at the 2006 Wikimania meeting: http://archive.org/details/wm06_kahle_plenary_5aug2006; TED Talk (2007): http://www.ted.com/talks/brewster_kahle_builds_a_free_digital_library.html; with the SD Form (2004): <http://archive.org/details/SDForumBK>; at the Library of Congress (2004): http://archive.org/details/cspan_brewster_kahle.

As part of the stump speech he gave to the Coalition for Networked Information in 2004, Kahle invited library leaders to step forward and to make their book collections available for scanning even though he had no large-scale scanning system in place. He would figure that out later. Having become impatient with the progress of the Million Book Project, Kahle started moving forward on his own. Carole Moore, then head of the University of Toronto's libraries, took Kahle up on his offer. Google had approached Toronto to become a scanning partner but Moore refused because of the restrictions they insisted on putting on Toronto's use of the books (even though all would have been in public domain). She told Kahle she wanted the scanning done inside her library, and so he gave up on the notion let over from the Million Book Project that scanning needed to take place outside of the U.S. to keep labor costs down. He first set up a scanning center in Toronto using a \$130,000 page-turning machine from the Kirtas scanning company, but it proved too expensive and undependable, so Kahle eventually hired an outside design company to build an affordable image capture machine, which became the Scribe scanner that the Archive has continued to use since (and which I describe in chapter 2). The Archive started to scan books in Toronto in late 2004.²²

Meanwhile, Google had been quietly pursuing its own book digitization project. Details surfaced as sketchy rumors. As early as 2002, Google executives had been secretly involved in discussions with the library leadership at the University of Michigan, where Larry Page had been an undergraduate, to digitize its library holdings (Carlson and Young 2005). Michigan's library had been a leader in digital library innovations, including the "incubation" phase of JSTOR, the Making of America book digitization project, and TULIP and PEAK, two experiments with Elsevier in the early electronic journal licensing, and they were also a public university, which would afford them sovereign immunity under the Eleventh Amendment, should they be charged with copyright infringement. Google was also involved with Stanford University in what the *New York Times* described as a "secret project" known as Project Ocean, to digitize all of its library's public domain holdings (Markoff 2004). These early dealings with Michigan and Stanford would grow, by 2004, into the Library Project.

After Google Print (the publisher service) was officially launched at the annual Frankfurt Book Fair in October but before the secret Library Project was revealed, Kahle welcomed the news because it meant someone was "stepping up to the plate" to get books onto the Web. He had always expected publishers to handle the digitization of in-print books and libraries to handle the rest. Between October and December 2004—when Google would surprise everyone with the news of the Library Project—Kahle and Google co-founder Larry Page discussed how the publisher program could be expanded to include older and out-of-print books (i.e.,

²² The University of Toronto became the largest Internet Archive scanning center.

books held by libraries). One option they considered was having the Archive (and others) manage the library digitization side of things, with Google funding it (on the condition that it not benefit their competitors). Kahle visited Google's headquarters during this time, where he presented the same general stump speech, with Larry Page and Sergey Brin present throughout the talk. Nonetheless, no one mentioned to him that they already had a commitment from the University of Michigan to scan its collection and were very close to having others. Kahle would learn about the Library Project from the *New York Times* like everyone else.

As details about the Library Project emerged, Kahle became alarmed. The secrecy and use of nondisclosure agreements with the libraries was the first problem. Details of library agreements with Google were not known until an advocacy group (googlewatch.org) filed a Freedom of Information request with the State of Michigan to request a copy of the University of Michigan's contract with Google.²³ From that document, it was learned that Google had placed restrictions on Michigan's use of public domain books—including not being able to share them with third parties such as the Internet Archive. The second problem was that Google was going to keep the book content for itself and not make it part of the open Web. Passionately, even obsessively, invested in keeping the Internet a decentralized, distributed network that operates on the principle of open systems, open protocols, and reciprocal benefit to those using it, Kahle identified Google's book project as a move against the open Web. As a search engine, Google lives most profitably off the open Web, and yet by not allowing Web crawlers access to the digitized books, they contradicted that very underlying rule of the Web. Hoarding and "walled gardens" preclude search engines from crawling and indexing and, as Kahle sees it, such closing off of content contradicts the Web. "They want everyone else to be open but themselves," Kahle told me. Google's Sergey Brin has recently lodged this same complaint against Facebook, which thrives on networked users but closes itself off from the wider Web: one cannot access content or information on its site without being a member (Katz 2012). The upshot, to Kahle, was that Google, with its Library Project, threatened to become a centralized point of access and control to books.

Google's leadership had violated what Kahle had assumed the Internet Archive and Google shared: *the ethos of the Web*. That ethos of the Web involves an "open and neutral" web platform that enables interoperable web services from many independent entities. To continue as such, it cannot be dominated or controlled by one (or three or five) companies or governments. If Google, the dominant search engine, was going to become a content repository, restricting access to that content (even when in the public domain), the company would be violating the implicit

²³ The Agreement can be found here: www.lib.umich.edu/files/.../um-google-cooperative-agreement.pdf; the University of California's can be found here: http://wayback.archive.org/web/20090915000000*/http://cdlib.org/news/ucgoogle_cooperative_agreement.pdf. None of the other universities has made its original agreement publicly available.

terms of the Web. That compact is necessary to keep the character and potential of the Web intact. Google had betrayed Kahle but the Web itself. And, in so doing, they had quickly gone from being an ally in building the Library to a major obstacle. Whereas the Archive and Google have a great deal in common—both organizations are fundamentally “of the Web”—Kahle’s sometimes surprising antipathy to Google comes from the fact, as opposed to the other major tech companies—Apple, Amazon, Microsoft—Google is in a position to do more to harm the future of the open Web than the others. Google has the ability to get under Kahle’s skin as only a familiar can.

In early 2005, Kahle appealed to Jesse Ausubel, a program director at the Alfred P. Sloan Foundation for support—financial and otherwise. Kahle approached Sloan in particular because its history of science and technology program had in 2003 supported the Archive with a \$1 million grant to “augment its capabilities” in preserving the Internet. Using the Human Genome Project as an analogy, Kahle pleaded with Ausubel to fund a public competitor to Google’s project. Google, in Kahle’s analogy, corresponds to Craig Venter’s Celera, and “human knowledge” corresponds to the human genome. Even the costs he estimated would be about the same. “We need to do this again,” Kahle told him, “we need to do a public human genome project around this.” In 2005, the Sloan Foundation gave the Internet Archive \$2 million to foment a counterpart project to Google’s. According to Sloan’s 2005 Annual Report:

The best-placed organization to unite the holders of content in a counterpart massive digitization effort built around the concept of open access is the Internet Archive. This grant enables the Archive to take steps to build the universal digital library, beginning with an archive of millions of books. The project requires more than \$200 million in funds by 2010, to be obtained from other foundations, libraries and other archival institutions around the world, and from private sources. The best technical advice suggests that today’s technology for scanning, data compression, storage, and distribution puts the goal within reach. The main hurdles are funding, political will, cooperation from libraries, copyright restrictions, and the competing commercial effort.²⁴

Kahle’s appeal to Ausubel had a particularly lasting effect on Doron Weber, the program officer for the Public Understanding of Science and Technology, the purpose of which is to “break down the barriers between the two cultures.” Weber considered Kahle “ahead of the pack” for “immediately understanding that the Internet was *the* new global matrix of universal culture and the documentary record

²⁴ Available here: <http://www.sloan.org/pages/15/annual-reports-financial-statements-of-the-alfred-p-sloan-foundation>.

of that culture for our time.”²⁵ He developed a subprogram to the Public Understanding of Science and Technology whose title indicates Kahle’s influence: “universal access to knowledge.” In 2006, Sloan gave the Archive another \$1 million, emphasizing again the project as a counter-effort to Google’s: “Unlike Google, OCA supports an open-access, non-proprietary online library in which no single entity can exercise exclusive control” (Sloan 2006).

Sloan also that year gave \$2 million to the Library of Congress for a “pilot demonstration of the feasibility and utility of an approach to mass digitization that may allow the Library of Congress to go to Congress with strong evidence that this is an investment that should be taken on by the federal government” (Sloan 2006). Herein lay the germ of disagreement between Kahle and Weber. Kahle didn’t see the OCA as an effort to building something that would eventually become part of the state. He wanted an open Internet library ecosystem not a Bureau of Digital Information. I witnessed this first at the OCA’s third annual meeting in San Francisco in October 2007. In a small breakout session, a group, which included Weber, discussed how the OCA could grow and become a stronger force. An Internet Archive employee who coordinated OCA activities explained, in response, that the purpose of the OCA was to facilitate collaborations around “openness.” The OCA, she said, was part of “the openness movement.” Weber expressed a different view: to him, the point of the OCA should be to create a universal digital library. Weber urged that the OCA become a full-blown organization with legal status: that it hire people, that it travel, talk to people, and coordinate efforts. He wanted to “turn this into the big thing that I think it is.” Eventually, he hoped, such efforts could lead to Congressional funding: “\$1 billion is nothing to Congress,” he said. “And it’s an easy sell. We’re at the beginning of the transformation into the digital age.” To this end, Sloan granted the Archive an additional \$500,000 to fund “a hands-on, day-to-day executive director for public advocacy, fundraising, and development of new partnerships” (Sloan 2006). Kahle and Weber’s differing perspectives on how digitization should proceed—fostering a loose confederation of decentralized efforts vs. building a universal digital library by means of a central organization—would grow into a rift by the next OCA meeting in October 2008, which just happened to coincide with the announcement of the Google Book Search Settlement.

As soon as he knew the rough outlines of the Settlement—at the center of which was a subscription database for selling the Library Project books back to university libraries—Kahle decided to do everything he could to stop it from being approved. He set himself the task of “finding friends” and fomenting as much opposition to it as he could. The debate around it grew slowly but heated up throughout 2009 and lasted until February 10, 2010, when a final “fairness

²⁵ Weber made these remarks at the inaugural meeting of the Open Content Alliance in October 2005. Video available at: <http://archive.org/details/ocalaunchevent>.

hearing” was held in Federal court to determine whether the Settlement was “fair and adequate” to members of the two classes that had joined to bring the suit: the author class and the publisher class. Between October 2008 and February 2010, a robust opposition had grown up against the Settlement, and included foreign publishers, foreign governments, literary agents, library groups, privacy advocates, author groups (including “academic authors”), Google’s competitors, and many individuals—even songwriter Arlo Guthrie. The most prominent opponent was the Department of Justice (or, as it calls itself in legal filings, “The United States of America”). Many who had supported Google in the original lawsuit became strong critics of the Settlement (e.g., Pamela Samuelson, Lawrence Lessig, the Electronic Frontier Foundation). Journalist Steven Levy, author of the geek class *Hackers* (1984) as well as a generally admiring book about Google entitled *In the Plex: How Google Works, Thinks and Shapes Our Lives* (2011a), observed that you knew something had changed for the worse in Google’s generally favorable perception among geeks if the Science Fiction and Fantasy Writers of America and Cory Doctorow were opposing the Google Book Search Settlement (Levy 2011b). According to Cynthia Arato, a specialist in copyright and class action law, the 377 objections and 13 amicus briefs files against case represented an “unprecedented” number of filings in opposition to a class action settlement (Arato 2009). In March 2011, the Federal judge overseeing the case issued his judgment rejecting the settlement because it “would simply go to far.” The staff of the Internet Archive and many others around whom I had worked considered the rejection of the Settlement to be a victory. The Archive saw itself as a “mouse that roared.”²⁶

I worked at the Archive from August 2008 through 2009, where my activities were varied but fundamentally involved the typical activities of an anthropologist: participant and other forms of observation; interviews; and the collection of documents. From these activities, I produced fieldnotes, transcripts, and an rather large archive of documents—the bases of this dissertation. In 2010 to early 2012, I continued to be involved with the Internet Archive though not regularly physically present there. The Archive is a very open and relaxed place. Not only does it have a lunch every Friday that is open to the public, but it is also very open to visitors and guests, including a note-taking, tape-recording anthropologist. As I write this, it is hosting an “artist in residence,” who was not invited but merely offered his services. Ted Nelson, the famed coiner of the term “hypertext,” has become a regular fixture as well. The atmosphere at is non-profit Northern California: relaxed, geeky, “alternative”; a *mélange* of aging Burning Man creatives, “makers,” and Silicon Valley veterans; open source advocates, copyright activists, Wikipedians, collectors, pro-drug advocates—and a cross-section of the Northern California counterculture as threaded through a devotion to the open Internet.

²⁶ In the Conclusion, I relate what has happened since the Settlement’s rejection.

Among my activities at the Archive headquarters, the most time-consuming was my work as a researcher, editor, and ghostwriter for Kahle in his efforts to stymie the Google Book Search Settlement. With Kahle, I had become caught up in the dispute. New to the Archive and eager to get involved, I could be put to work. The Archive has a small staff, mostly engineers, and, then at least, it had nothing like a policy, research, or communications department.²⁷ Just as Kahle saw an opportunity in me, I saw an opportunity for myself. The anthropologist is always alert to opportunities for “access,” and shortly I was in phone conferences with lawyers, activist groups, tech company executives, library leaders, and many others—all potential allies against the Google Book Search Settlement. The scope of my research grew from studying simply the Archive’s book digitization activities to encompass the expansive controversy around the Settlement, as the Internet Archive became a key opponent and critic of the Settlement.

I worked closely with Brewster Kahle to figure out the implications of the Settlement. This was no simple task. Even the most adept legal observers had to consider it for some time before knowing how to think about it. We both kept our ears to the ground—he to his network and I to the Web and other media—consulting with many folks over Skype and phone conferences.

Beyond the Archive, I participated in call-ins sponsored by various organizations such as the Authors Guild (clued in by a literary agent friend) and the Copyright Clearance Center; I participated in an online workshop on the Settlement organized and hosted by copyright scholar Peter Jaszi; I audited Pamela Samuelson’s course on the Settlement at UC Berkeley’s law school; and I attended conferences on the Settlement (at Columbia University, Harvard, and Berkeley) and participated myself in another (at New York Law School). I also attended two hearings in the Federal court in Manhattan, including the final Fairness Hearing where supporters and opponents lined up to say their bit in front of the Judge. I interviewed people important figures in mass digitization in Ann Arbor, Ithaca, Cambridge, New York, and various parts of California. With others, I Skyped. In short, I became one of a small group of “Settlement groupies” who followed it from the outside with a special tenacity. Once, as I walked to yet another talk on the Settlement on the Berkeley campus, I ran into Pamela Samuelson, perhaps the leading critical voice against the Settlement, and remarked: “You must think I’m stalking you.” To which she replied: “No, I know that you are simply among the obsessed.”

My inquiry, however, was not confined to the Google Book Search Settlement or the Archive’s involvement with it. I scanned books alongside the professional scanners in the Archive’s scanning center in downtown San Francisco. I attended the regular Monday morning “books meeting.” While it existed, I was also part of

²⁷ The Archive hired Peter Brantley to combat the Settlement full-time in spring 2009.

the Archive's "access group," which was formed to expand the Open Library project; to convert the Archive book collection into the industry standard E-pub format for distribution as e-books, and into the DAISY format, an accessible format for the "print disabled." (My note-taking was on occasion useful to the Archive staff.) When asked, I helped some Archive employees with grant applications, press releases, or other public-facing documents. I monitored the Archive forums to see how users were using the Archive. I was the "help desk" for Open Library. I was happy to be regularly included in meetings, impromptu and not, both within the Archive and without: in shiny venture capital offices along the Embarcadero, in overcrowded nonprofit headquarters in the Mission, in warehouses in a desolate neighborhood of Richmond, CA, or elegant high-rises overlooking the San Francisco Bay. And, of course, in the best anthropological tradition, I also spent a lot of time "hanging out," whether in the Archive offices, or on afternoon "beach walk" along the Presidio's San Francisco Beach Trail, chatting with the engineers about everything from the difference between code "sucking" and "being evil" to whether Blue Bottle coffee beans were or were not the best around.

* * * * *

I consider this dissertation a punctuated exploration of recent events in mass digitization with sensitivity both to the past and the near future. I did not have a research protocol, even if my human subjects review demanded such a document from me. I proceeded in response to the environment I found myself within. As in conventional accounts of anthropological fieldwork, it took a while to get my bearings, to find a place for myself, to figure out a way to work with the people around me. I also had a lot to learn myself about books, computers, engineering, libraries, and the law. Even when I arrived at one plateau of competence, a new challenge would appear. I was working in a complex environment, to which I was required to respond and to negotiate on the fly, improvising in the face of opportunities as they presented themselves or to events that occurred and perspectives that changed. There was no clear path from the beginning of my research to the end, and I stopped not because of developments in "the field" because of the time constraints of graduate education. Things are happening as I write this paragraph that may well render some facts or issues reported here out-of-date. The field is in motion. But, of course, the facts are not the point, and good coverage of book digitization can be found by looking into the many resources indicated in my footnotes and bibliography. What I hope to have accomplished here is the first step in "pausing, reflecting, and putting forth a diagnosis" of how contingent elements within a broader assemblage (within an even broader problematization) are coming together and pulling apart (Rabinow and Bennett 2012). One researcher cannot accomplish such a diagnosis on her own, and I hope

through this work to open up the book as contemporary problem space so as to attract more anthropologists into it.

Chapter 1

Pathways to Digitization

“As far as the book is concerned, despite admirable technological progress since the fifteenth century, all is far from perfect.”

-- Paul Otlet and Robert Goldschmidt, 1906

The main character of Kurd Lasswitz's (1848–1910) science fiction fabulation, “The Universal Library” (1901), describes a hypothetical library made from an irreducible set of twenty-five orthographic symbols, whose recombinations would be capable of forming every possible verbal expression—past, present, or future—in any language.¹ Such a library would contain any text that had ever been lost and every text that had yet to be written, as well as everything that is correct and everything thing that is not. It would also be larger than the known universe and be unusable. Four decades later, in clear reference to Lasswitz, Jorge Luis Borges (1899-1986) expanded upon this mathematical phantasmagoria in his now famous “The Library of Babel” (1941).² In it, the Library (coextensive with the Universe) comprises an infinite number of identical, hexagonal galleries, which are lined on four sides with shelves full of books of identical size and shape (the other two are for human necessities). The books contain, as in Lasswitz, all possible combinations of twenty-five irreducible characters. Despite its meticulous structure, its organization and order, the Library is chaotic, disorienting, exasperating, and enervating.

These fables evoke what Friedrich Kittler has analyzed as “discourse network 1900,” a new episteme usurping a previous network based on texts, literature, universities, paper archives with one based on technology (media, data flows and streams, information networks, machines) (Kittler 1990). The typewriter, with its mechanization of handwriting, had in the late nineteenth century detached writing from any human trace or subjectivity so that, in the place of the writing hand were

¹ Originally published in a collection *Traumkristalle* (1901). My reference is to the English translation, “The Universal Library,” in *Fantasia Mathematica*, Clifton Fadiman ed. New York: Simon & Schuster, 1958, p. 237-43.

² Originally published in *El Jardín de los Senderos Bifurcados* (1941). I am using the translation in *Collected Fictions of Jorge Luis Borges*. Trans. by Andrew Hurley, p. 112-18. New York: Penguin Books. Borges's 1939 essay “The Total Library” discusses Lasswitz as well.

only mechanized parts—the twenty-six finite letters of the Latin alphabet—with no spirit, no essence, just naked, recombinable material signifiers. The hoary book, which had been the paradigmatic discursive vehicle of the hermeneutical enterprise and the embodiment of Spirit extended into form, lost its “magic” amid an undifferentiated background of noise (1990, 178). Borges’s beleaguered, suicidal librarians were searching not for order in the chaos but for a signal in the noise.³

Between the years these two stories roughly mark out—1900 to 1940—bibliographers, social critics, scholars, librarians, entrepreneurs and science advocates were anticipating, in their own ways, a universal library that was also modularized, mechanized and mobilized. Their enabling techniques were index cards and microphotography. These individuals, beginning at the end of the nineteenth century, together saw the book as inadequate, and even detrimental, to the intellectual needs of modernity, and they looked to new means of mechanical reproduction that would mechanize, modularize, and mobilize the book as a solution to its inadequacies. Viewing the book as a set of problems that can be resolved through technological intervention is also a feature of mass digitization and, as such, this chapter is what Lisa Gitelman has referred to as a “pre-digital” history of the digital: an account of surprising continuities rendered against obvious and admitted discontinuities” (Gitelman, forthcoming).

The lens through which I look backward is that of one particularly popular technological “solution” to the problem(s) of the book: microphotography. As early twenty-first century library administrators, computer engineers, journalists, scholars, and Internet activists are enamored of digitization, early twentieth century bibliographers, scholars, science advocates, and librarians were enamored of microphotography. By focusing here on this old (and largely outmoded) practice, I provide neither an exhaustive history of the twentieth-century problematization of the book nor even an exhaustive history of microphotography and photoduplication. What I can provide, however, is what I call, after Paul Rabinow, one “pathway” for understanding mass digitization and the field of thought and action around it. Rabinow develops the concept of “pathway” as an “orientation to inquiry, picking out and connecting elements across a heterogeneous and dynamic contemporary problem space” (Rabinow 2010). The concept provides a means of thinking about the significance of the past that avoids narration in a teleological mode of historical inevitability or explanation. I take my pathway, the early twentieth-century fascination with microphotography and its perceived promise as a solution to serious problems within the book apparatus, as a key feature of a “path-connected set of nodes” between the past and the present.

³ Walter Benjamin captured this thus: “And before a contemporary finds his way clear to opening a book, his eyes have been exposed to such a blizzard of changing, colorful, conflicting letters that the chances of his penetrating the archaic stillness of the book are slight” (Benjamin 1928, 172).

Microphotography: A Brief Introduction

A microphotograph is an image taken through a microscope lens so as to render it many times smaller. Herman Fussler, a pioneer of microfilm use in libraries, defined it in 1940 thus: “A microphotograph is a photocopy of any object ... which has been copied on an unusually small scale or, as we more commonly say, at ratios of reduction such as to make reading difficult or impossible with the unaided eye” (Fussler 1940, 9).⁴ The image, stored on film, is later viewed by means of an enlarging machine or, in the case of documents and textual material, a reading device or “reader.”

Those of a certain age will know microphotography more familiarly as microfilm (or, perhaps, microfiche). For many researchers today, microfilm is the only way to read certain material, especially newspapers, because the originals were discarded after microfilming. It is stored as a reel, like other film, and one reads it by winding a crank that threads the film through the machine until the desired frame is located and then viewing the magnified image on a lighted screen. Microfiche—a flat film—requires moving the hovering lens above the fiche, something like consulting a Ouija board, until one locates the frame containing the desired document or passages. It is a cumbersome and often unpleasant method for consulting books (or newspapers) because in order to find one’s desired place, one has to scroll back and forth on the reel of film that could contain much besides the consulted document. In his screed against microfilming, Nicholson Baker notes one archive that even took the step of taping an airsickness bag to its microfilm reader in case the user of the machine suffered from motion sickness while attempting to track something as the scrolling text whizzed by (Baker 2011, 40). But, in stark contrast to the nuisance microfilm may conjure up today, a century or so ago the same technology inspired grand visions, as extravagant as those used more recently to describe the transformations surrounding the Internet and the World Wide Web.

Microphotography followed right on the heels of photography in the nineteenth century. Just months after Louis Daguerre introduced his photographic method in 1839, English optician John Benjamin Dancer produced microphotographs using daguerreotype plates and a camera with a microscope lens. His first micro-reproduction was a miniaturized document reduced at a ratio of 160:1 (Luther 1959, 16). The new technique became popular as a novelty, especially in the form of tiny portraits, but its future would be in more practical applications. In 1853, scientist and photography pioneer J. F. W. Hershel saw in the new method the

⁴ A photomicrograph, conversely, is a photograph taken of something tiny in order to make an enlarged image of it.

potential for publishing “concentrated microscopic editions of works of reference.”⁵ In 1857, Scottish scientist Sir David Brewster, who toured Europe enthusiastically exhibiting Dancer’s microportraits, speculated that microscopic copies of “valuable papers” and “secrets” could be transmitted by post.⁶ Indeed, the first large-scale practical use of microphotographed documents was the “pigeon post,” developed during the 1870 Siege of Paris in the Franco-Prussian War, when Paris had been cut off from the rest of France. Carrier pigeons were able to transport thousands of miniaturized letters at a time into Paris.⁷ Microphotography would find itself useful once again in World War II. The score of Shostakovich’s *Leningrad Symphony*, written during the Siege of Leningrad in 1941, made its way out of Russia to the United States on microfilm. And microfilm companies, such as Eugene Power’s University Microfilms, worked for both the British Foreign Office and the predecessor to the Central Intelligence Agency, the Office of Strategic Services (OSS), filming intercepted German mail, newspapers, and other printed materials that its foreign agents were collecting (Power 1990, 133-42).

Once film from the motion picture industry and cameras from banking (used for recording and storing cancelled check images) proved commercially viable, an industry developed around microphotography for broad application. To those who cared in one way or another about books in the 1920s and 1930s, microphotography appealed as an alternative to paper and as an exciting field of experimentation for copying, publishing, distribution, and storage. Microphotography offered an immensely attractive reformatting technique because it provided compact storage; was easily reproducible, light and easily transportable; and it was permanent—or so it was believed at the time.

In this chapter, I trace the interests of a disparate group of individuals who, together, form a modernist reform movement seeking a remediation of the book apparatus with microphotography as the enabling technique.⁸ Bibliographers (such as Paul Otlet) reimagined universalized information and knowledge as not only benefiting scholars and researchers but also bettering all of humankind. Individual scholars

⁵ J. F. W. Herschel, Letter to Editors of *Athenaeum*, July 6, 1853, no. 1341, p. 831. Reprinted in Veaner 1976. The potential was so self-evident that Herschel added, “The details are too obvious to need mention.”

⁶ Sir David Brewster, “Microscope,” in *The Encyclopaedia Britannica, or Dictionary of Arts, Sciences, and General Literature*, Eighth Edition, Vol. 14, Edinburgh: Black, 1857, p. 801-2. Accessed March 17, 2011 on Google Books.
http://books.google.com/books?id=EmlBAAAACAAJ&source=gbs_navlinks_s

⁷ Luther 1959, chaps. 5-7, 14-15. For fuller accounts, see J.D. Hayhurst, *The Pigeon Post Into Paris 1870- 1871* (Ashford, 1970). See also John Stirling Fisher, *Airlift 1870: The Balloon and Pigeon Post in the Siege of Paris* (London: M. Parrish, 1965).

⁸ See Rayward 2008. I extend the group identified by Rayward to include figures in the United States.

(such as Robert C. Binkley), scholarly societies (such as the Social Science Research Council and the American Council of Learned Societies), and philanthropies (such as the Rockefeller Foundation) saw the potential of microcopying techniques for improving, expanding, and modernizing research. Entrepreneurs (such as Eugene Power) devised new modes of publishing in microformats. And science advocates in the United States (such as Watson Davis and Vannevar Bush) saw in microfilm a crucial component of a new infrastructure appropriate to the expanding ambitions of American science.

“Far From Perfect”

What is important is not the technology chosen, but the problems for which solutions were sought. Microphotography was part of a response to a larger problematization about knowledge, its forms, its dissemination, and its reproduction, as mass digitization is today. This chapter thus is not a history of microphotography as applied to books but rather an exploration of the motivations and thoughts of those who sought to address the problems of the book with microphotographic solutions.

Around the time Lasswitz published *“The Universal Library,”* Belgian bibliographer Paul Otlet (1868-1944) began his career-long effort to conceive and build a universal knowledge system that he called, among other things, the Universal Book. Here I will pay extended attention to the Belgian Paul Otlet (1868-1944) and his new *“science of the book”* because he was a widely connected advocate for the reform of the book apparatus. He had cross-cutting ties across Europe and the United States, collaborating with a wide array of cultural actors, as this chapter will show. I take Otlet as a sort of index to the preoccupations around the book in the early twentieth century.

Otlet was the son of a wealthy industrialist and financier, Edouard Otlet, who had built railways and tramways throughout Western Europe before becoming a financier with interests in Africa and South America. In 1886, he even funded a (failed) expedition to the Congo, from which he and others had hoped to form a *“museum of Africana”* (Rayward 1975, 15). Paul grew up part of a close-knit upper class elite, and as a young man he was obsessed with *“the necessity of performing in life some magnanimous and useful task for society”* (Rayward 1975, 12). Enamored of positivist philosophers August Comte, Herbert Spencer, Alfred Fouillée, Otlet developed a sense of purpose out of a belief in science, a rejection of traditional metaphysics and a conviction that a synthesis of knowledge was possible (Rayward 1975, 28).

As a young disaffected lawyer, Otlet developed a devotion to bibliography while working as a clerk for Edmond Picard, a prominent socialist politician and self-styled legal philosopher who promoted a *“positivist bibliography”* (Rayward 1975,

30-31). Part of Otlet's work as clerk involved assisting in the compilation of a massive legal bibliography, the *Pandectes belges*. The thrust behind the project was Picard's idea to apply the "procedures common to the natural sciences" to law:

Facts, observations, then more observations and facts, to deduce afterwards general truths is the way to proceed, a procedure which, formulated by Bacon, has gradually established its Domain and has become the rule for all serious study. The human mind is no longer considered as an organ which produces the sciences, but rather as an apparatus for enregistrement, whose unique role is to observe the laws which emerge from carefully collected facts and from scrupulously carried out experiments" (quote in Rayward 1975, 30).

Otlet was soon to begin articulating his own vision of a reformed bibliography along similar positivistic lines. While working for Picard, in 1892, Otlet had met and befriended another upper-class Belgian, the utopian socialist internationalist, Henri LaFontaine, with whom he would have a life-long professional relationship.⁹ Together they conceived and founded, in 1893, an institutional complex for bibliography that would eventually become the International Institute of Bibliography (IIB).¹⁰ With the support of the wealthy individuals and the Belgian government, they and colleagues developed a range of institutions, tools, and practices in the service of the international organization and dissemination of knowledge. Its purpose was to build a science of society through bibliographic organization on the model of the natural science and, from such a unified science of society, they believed they might unify societies across the globe.

Otlet was particularly concerned about the potential for a new bibliography to help the social sciences because they, he felt, had so much catching up to do when compared to the natural sciences. Science, he believed, was approaching a great synthesis because it had succeeded so well at documentation: the collection of facts for observation and analysis (Otlet 1892, 11). The proliferating mass of monographs in sociology, on the other hand, was an "unclassified mess," difficult to find or make use of because of the problems of books.

⁹ LaFontaine, who as a lawyer specialized in international arbitration, would win the Nobel Peace Prize in 1913 for his role in the popular peace movement in Europe.

¹⁰ Its name would change in 1931 to the International Institute for Documentation and in 1937 to the International Federation for Documentation (Rayward 1975, 325, 338) and finally in 1980s to the International Federation for Information and Documentation. For ease of comprehension throughout this chapter I will refer to it by its original name, the IIB. (It was finally dissolved in 2002.)

In a 1903 essay Otlet outlined a series of obstacles books presented to the demands of modern science and “society.” First, books obscured knowledge by hiding it. He sought to “release” the knowledge hidden and confined in books and to rescue original theses, novel observations, and important results from the “superfluities” and “dross” in which they “are submerged and disappear” (Otlet 1903, 84). Second, books were also too time-consuming and difficult to work with, and they needed to be made easier to work with: “The experiments we are now witnessing are to make the book easier to consult and easier to handle so that it is more effectively and more quickly informative—in a word, more documentary” (Otlet 1903, 85). Third, books existed in isolation from each other such that he sought means for remediating them into a vast interconnected system that would not simply enable readers to identify and locate them—a conventional goal of bibliography—but would also connect *the books (and the ideas that they contained) to one another*. This required a move from wholes to constituent parts: “Just as the chemists have moved from analyzing molecules to analyzing atoms, and biologists from tissues to cells, even so must the bibliographer having completed the inventory of written works, attempt an inventory of the *contents* of these works” (Otlet 1903, 78; emphasis mine). Fourthly, there were also too many books, and since there was no way to stop authors from writing them, some method of managing them, of identifying the usefulness of each, was required. Authors were also a problem. They stuffed their books with so much irrelevance, rhetorical flourish, and personal idiosyncrasy that the actual “knowledge” in their books was frustrating to locate. And, finally, books were suited to a different mode of reading, which was out of step with the times: “Once, one read; today one refers to, checks through, skims. The trend is no longer slavishly to follow the author through the maze of a personal plan which he has outlined for himself and which, in vain, he attempts to impose on those who read him... ” (Otlet 1903, 79). In short, both books and the science of books (bibliography) were inadequate to the times: “The old forms of the book will no longer be maintained; they must give way before the abundance and the variety of matter” (Otlet 1903, 84).

What was required was a new science of the book, and Otlet spent the rest of his life developing it. For his renewed science—which he would dub “documentation”—he would formulate and put into practice the mechanisms for making not just books but all documentary materials properly useful. Documentation was intended to replace bibliography, which had traditionally concerned itself, as its etymology makes clear, with books: their collection and description, their printing history of printing, and cataloging practices. Documentation, instead, would subsume books into a greater category—documents—that included any material form of communication whether text, image, or sound. Documentation would be a universal science of documents—“the assembling, classification, and distribution of documents of all sorts in all fields of

human activity.”¹¹ The new science of documentation, which flourished until becoming subsumed into “information science,” is properly seen as a modernist movement—a search for new forms for new times—and Otlet was the most committed of reformers.

Otlet, with a wide variety of colleagues, set out to build a new infrastructure, with its center in Brussels, which would result in a worldwide “radiated library.” Books would be de-centered in the new library; they would become one sort of document among many. Books would also require disaggregation: from solitary, idiosyncratic crypts for knowledge into expertly winnowed collections of cards, atoms, or “facts.” Books, in their new form, would be put into motion, reduced to their essence, and made recombinable. A new user was imagined—the scientist or specialist who sought extractable pieces of knowledge or “information,” not extended arguments that depended on nuanced language, rhetoric, or the individuality of an author’s expression. And a particular form of reading—efficient, effective consultation—was privileged over bookish notions of intensive attention. The new library would also be impersonal, mechanical, systematic—optimized for mobility and cross-referencing—and standardized to serve a universal humanity.

He called it the International Institute of Bibliography (IIB) comprising three elements: 1) “repertories,” or systematized collections of information from particular fields of knowledge; 2) a universal classification system; and 3) new local institutions called “offices of documentation.” The first element, the repertory, was a massive storage system for “knowledge” culled from books and other documents. The goal of the Repertory, Otlet wrote, was: “To detach what the book amalgamates, to reduce all that is complex to its elements” (Otlet 1918, 149). Such “detachment” would be accomplished through what he called the “monographic principle”: a method of extracting the essential knowledge from inside books, separating it out from the “dross” and that which was merely “fine language, repetition or padding” or “any aspect of the author’s personality” (Otlet 1918, 149; Otlet 1891-92, 17).¹² Essential knowledge belonged to one of four categories: facts, interpretations of facts, statistics, and sources (Otlet 1891-92, 16). The many cards that resulted would be collected in cabinets of drawers, resembling a library’s card catalog, or pieces of paper and kept loose-leaf binders. This way, new cards or leaves could be inserted and old ones removed, so that the Repertory was always

¹¹ Quotation from the letterhead of the Institut International de Documentation (a subsequent name of the IIB). Cited in Schultz and Garwig (1969, 153). See also, Buckland (1997).

¹² Otlet is likely to have borrowed the term “monographic principle” from Wilhelm Ostwald (1853-1932), a Nobel-prize winning Latvian chemist and co-founder of The Bridge, an international institute in Munich, similar to the IIB, dedicated to the organization of intellectual work (Rayward 1994). Ostwald was one of the many international colleagues with whom Otlet collaborated (see Hapke 1999; Krajewski 2011, ch 7).

up to date. By 1914, it had over 11 million entries and by April 1934 nearly 16 million, as seen in figures 1 and 2.



Figure 1: Card files of Universal Bibliographic Repertory, Brussels.



Figure 2: Card files of the Universal Bibliographic Repertory, Brussels.

Individual disciplines or subjects would have their own repertories, which would be aggregated into the “universal bibliographic repertory,” a comprehensive collection of all knowledge. It would be a “kind of artificial brain by means of cards” (Otlet 1892, 17) and a “machine for exploring time and space” (Otlet 1903, 86). Books, once re-formed as a series of removable cards would deliver up the knowledge they keep hidden.

Media archaeologist Markus Krajewski has recently argued in his book *Paper Machines* (2011) that the modularization of knowledge through the monographic principle and the card index instigated a “war of modern writing systems,” wherein the card index competes with the book for domination—and wins—at least according to a variety of German commentators in the 1920s and 1930s, including Walter Benjamin (Krajewski 2011, 135ff).¹³ The basis of the card index’s victory is its “functional superiority of mobility” (Krajewski 2011, 127). Cards—uniform in

¹³ Benjamin wrote in the late 1920s: “And today the book is already, as the present mode of scholarly production demonstrates, an outdated mediation between two different filing systems. For everything that matters is to be found in the card box of the researcher who wrote it, and the scholar studying it assimilates it into his own card index” (Benjamin 1928, 172). The book is only the thing into which one scholar reassembles the notes he has collected, and its fate is to be mined and made into another scholars’ notes—a book being a mere mediation between filing systems.

size¹⁴—are efficient “mobile carriers” of information, making the card index into the “paper machine” of Krajewski’s title. The rigid form of the book, in contrast, prohibits such motion: “Bookbinders are the enemies of mobility” (Krajewski 2011, 117), and mobility is the highly desired function for modern knowledge.¹⁵ The book unbound enables unlimited insertions, constant extension, reconfiguration and continuous updating—those requirements for an intellectual system proper to modernity (Krajewski 2011, 127).

But how to actually make all of these discrete bits of knowledge “universally accessible and useful” (to quote from Google’s mission statement)? The second element of the Otlet’s bibliographic structure was a classification system (the Universal Decimal Classification or UDC), which would make the repertory’s mass of “facts” searchable. Boyd Rayward likens the UDC to a “highly sophisticated software package” for providing access to the repertory/database (Rayward 1990, 4). Otlet and LaFontaine devised the UDC after discovering Melvil Dewey’s decimal system in 1891 and formulating their own elaboration of it. Whereas Dewey had developed his classification system for shelving books in the American public libraries, such that they could be easily located and retrieved, Otlet saw in the Dewey Decimal System a potential tool for describing the relations among *the ideas inside the books* and not just among the books themselves.¹⁶ It would provide the structure of an “ideal library” by encoding concepts in its theoretically infinite decimal numbering system and thereby representing the totality of what was known while still allowing one to find “one’s way through the labyrinth formed by all these objects” (Otlet 1918, 150-1). Ultimately, the combination of the UDC classification scheme with the monographic principle collected in the repertories would establish connections across and among all of the knowledge in books, and, in so doing, Otlet believed the “chaos” of the innumerable, individual, idiosyncratic books would come together into one entity, which he repeatedly called the Universal Book--“a book which will never be completed but which will grow unceasingly” (Otlet 1903, 84).

So the UDC would index the information but, still, how would one actually get to the information? This is where microphotography comes in. The final component of the IIB’s institutional complex were “offices of documentation.” More than a mere “museums for books”—Otlet’s dismissive description of the traditional library—offices of documentation would bring together all types of documents (photographs,

¹⁴ In 1908 Otlet’s IIB decided on the 7.5 x 12.5 cm (3” x 5”) as the internationally valid format for index cards (Krajewski 2011, 92; Balsamo 1984).

¹⁵ Benjamin, too, characterizes the book as immobile when he refers to the “archaic stillness” of the book (Benjamin 1928, 172).

¹⁶ Today the UDC is known as a “faceted” classification system. A faceted classification system attributes to an object multiple attributes so that the classification can be ordered in multiple ways, rather than according to a single, predetermined, order (such as, say, alphabetical order).

films, audio, television, reports, statistics) and “achieve all that is lacking in the library” (Otlet 1918, 154). Located throughout the world, the offices would, through cooperative sharing arrangements, serve up the organized and classified information to patrons. Combining bibliographic, image, and textual databases, the offices of documentation would become nodes in a worldwide system of information dissemination—the “radiated library”—with the IIB at its center (Wright 2007, 265n2).

Microphotography would be that mode of diffusion. To Otlet and many others, microphotography would make it possible that “all of Human Thought could be held in a few hundred catalogue drawers ready for diffusion” (Otlet and Goldschmidt 1925, 93).¹⁷ As such, microforms would provide the storage mechanism from which one, from a workstation in any office of documentation, could query and search across the systematized repertoires. Microforms could also become, effectively, miniaturized libraries, or what Otlet terms “microphotolibraries” and “encyclopedia microphotica.” Microphotography formed, then, a fundamental component of Otlet’s expansive vision, even if it was the least realized.

Otlet’s expansive plans to reorganize knowledge fit into what historian of science Geoff Bowker has called the “drive to database” (Bowker 2005, 109). He argues that the rise of statistics, the development of imperial and state archives, and much humbler innovations such as carbon paper, manila folders, and, indeed, the 3x5 index card combine toward “database ... the most powerful technology in our control of the world and each other over the past two hundred years” (Bowker 2005, 108). Therefore, the nineteenth-century archive and the twentieth-century database exist on a continuum of technologies for the administration and ordering of social worlds in one shared “memory epoch,” characterized fundamentally by a move from narrativized record-keeping (the book) to non-narrative record-keeping (the card index, the repertory) (Bowker 2005, 30).¹⁸ The reform of the book is an important, and ongoing, part of this reordering, and what in 1925 looked possible with microphotography now looks possible, to the likes of Brewster Kahle, through digitization.

¹⁷ Robert Goldschmidt (1877-1935) was a Belgian scientist and inventor with whom Otlet had worked on ideas for a microform reading system (see Otlet and Goldschmidt 1906). He continued to working on reading machines and microfilm processes.

¹⁸ Media theorist Lev Manovich makes a related argument about narrative vs. database: “As a cultural form, the database represents the world as a list of items, and it refuses to order this list. In contrast, a narrative creates a cause-and-effect trajectory of seemingly unordered items (events). Therefore, database and narrative are natural enemies. Competing for the same territory of human culture, each claims an exclusive right to make meaning out of the world (Manovich 2001, 225).” He differs from Bowker in that the two opposed forms (Bowker’s “memory practices”) continue in synchronous “combat.”

Four Key Problems

The rest of this chapter considers the specific means by which microphotography, from the 1930s onward, became fundamental to reformist ideas around the book apparatus in Europe and, eventually, even more so in the United States. I have chosen to focus on four key intersecting problems: 1) the problem of too many books; 2) the inadequacy of the book and the library for science; 3) the limits that print publication exercised on accessibility and circulation; and 4) the perishability of modern paper. While I intentionally limit myself to the period before the computer with the first three problems, the final problem—paper perishability—I follow up to the “end” of microfilm in the 1990s because library concerns over the preservation of “brittle books” serve as a direct bridge to mass digitization.

Multitudo Librorum

“Information overload” is a recent term for an old complaint, even if people consistently consider their own circumstances more extreme than those in the past (Blair 2010). Intellectual communities with strong textual traditions have consistently expressed anxieties about book (over)abundance. Even well before printing, the fact of too many books has persistently bedeviled scholars and others. But, the obverse pertains as well: a strong desire for, and pleasure in, accumulation and access to greater and greater numbers of books. Given its Janus-faced character, it should not surprise anyone to find contradictions fundamental to the modern research library. As ambitious libraries acquire vast collections, often in competition with one another, those very collections aggrieve their keepers, with the demands of organizing, storing, and taking care of them.

As far back as 1890, British Prime Minister William Gladstone, who was also a book collector and library founder, anticipated a population explosion among books: “A book ... is smaller than a man; but, in relation to space, I entertain more proximate apprehension of pressure upon available space from the book population than from the numbers of mankind... Already the increase of books is passing into geometrical progression.”¹⁹ The head of the University of Chicago Libraries, M. Llewellyn Raney, would later claim (in 1940) that American university libraries were doubling in size every twenty years and that “no institution can indefinitely support such a geometric ratio of growth.”²⁰ In 1944, another

¹⁹ William Ewart Gladstone, “On Books and the Housing of Them,” *Nineteenth Century* 27:157 (March 1890) 384–396. Page numbers refer to reprint edition: New York: Dodd & Mead, 1890, p. 6.

²⁰ M. Llewellyn Raney, “A Capital Truancy.” *The Journal of Documentary Reproduction*. Vol. 3, no 2 (June 1940), p. 83. Herman Fussler, also of the University of Chicago, asserts the same figure of twenty years in “Microfilm and Libraries,” in William Madison Randall, ed., *Acquisition and Cataloging of Books*.

prominent librarian, Fremont Rider (1885-1962), famously proclaimed in his *The Scholar and the Future of the Research Library*—a classic in the library literature—that research libraries double even faster, every sixteen years:

Every scrap of statistical evidence that we can gather shows that, as far back as we can reach, the story is exactly the same. It seems, as stated, to be a mathematical fact that, ever since college and university libraries started in this country, they have, on the average, doubled in size every sixteen years.... at a rate so uniform over so many years, and so uniform in so many different libraries, that it might almost seem as though some natural law were at work (Rider 1944, 8, 15-16)

Although his calculations were disproved, such Malthusian sentiments continued over the course of the twentieth century, and library growth came to be seen as an inexorable and oppressive force that threatens to overwhelm the stewards of book collections.²¹

Indeed, for Rider, accumulation posed great danger not just to libraries but to all of humankind:

This is far more than a library problem; far more than merely an educational problem; it is a problem—and a problem to the nth degree complex and baffling—of civilization itself.... We seem to be fast coming to the day when ... civilization may die of suffocation, choked in its own plethora of print (Rider 1944, 13-14).

Too many books will make it impossible to sort the good from the bad or to make proper use of any of it.²² This irony—that the book could invert its civilizing mission and, through improperly managed accumulation, de-civilize—continues a theme in European history (Blair 2010, chap 1). Ecclesiastes, Seneca, Erasmus, Descartes, Bacon—all complained about the overabundance of books. Leibniz worried in 1680 about a “return to barbarism” from “that horrible mass of books which keeps on growing” (Blair 2010, 58). Adrien Baillet, a biographer of Descartes, wrote in 1685: “We have reason to fear that the multitude of books which grows every day

²¹ See Robert E. Molyneux, “What Did Rider Do? An Inquiry into the Methodology of Fremont Rider’s *The Scholar and the Future of the Research Library*.” *Libraries & Culture* 29:3 (Summer 1994), 297-325; and Robert E. Molyneux, “Patterns, Processes of Growth, and the Projection of Library Size: A Critical Review of the Literature on Academic Library Growth,” *Library and Information Science Research* 8 (January-March 1986). Molyneux shows that growth was indeed exponential from 1830-1938 and from the mid-1950s to 1968, but at no other time.

²² It seems worth nothing that Borges, at the time of writing “The Library of Babel,” worked in the basement of a municipal library in Buenos Aires, cataloging its books.

in a prodigious fashion will make the following centuries fall into a state as barbarous as that of the centuries that followed the fall of the Roman Empire” (Blair 2010, 59).

Microphotography presented what appeared a plausible strategy for mitigating the conflicting pressures of the *multitudo librorum*: libraries could be miniaturized book by book. As early as 1851, a photography jury at the London World’s Fair, for example, responded to microphotographs by anticipating “miniatures of printed books” and a condensed storage technique for archives whose “enormous mass of documentary matter ... more and more defies collection” (Luther 1959, 12). *Photographic News* imagined that “whole archives of a nation might be packed away in a snuff-box.”²³ German photochemist Hermann Vogel reported, in 1874, on an idea then in circulation for “microscopic libraries” that could bring “the substance of books filling entire halls ... within the compass of a single drawer.”²⁴

By the 1940s, Rider had dedicated himself to the task of convincing the library establishment that growth was an urgent problem.²⁵ He divided *The Scholar and the Future of the Research Library* into two halves. The first (“The Problem”) described the growth problem summarized above. The second half (“The Solution”) outlined his proposed solution—the “microcard.” A 3x5 card of sensitized paper, the microcard would contain microphotographs of a book’s pages on its reverse side, with the ordinary catalog information on the front. If you had found the record, you had also found the book. The library thus became “endless aisles” of such file cases (Rider 1944, 165). Rider argued that his microcards would actually occupy “no space whatsoever,” because they would be taking advantage of the “wasted” space on the unused back of cards already in the card catalog (Rider 1944, 102). Even if little about his solution worked as he had envisioned, including the fact that little space was saved, Rider was himself so persuasive that by 1948 Barnes and Noble, H. W. Wilson, Mathew Bender, the American Council of Learned Societies and others were publishing microcards, libraries were buying them, and it

²³ M. H. Garbanati, quoted in *Photographic News*. February 4, 1859.

²⁴ Hermann Wilhelm Vogel, *The Chemistry of Light and Photography*. New York: D. Appleton & Co., 1875, p. 210. He continues: “...a circumstance which, with the enormous increase of material that has to be swallowed by our libraries, may be of importance.” The full text of this book is available here: http://openlibrary.org/books/OL2669383M/The_chemistry_of_light_and_photography

²⁵ Nicholson Baker (2001) devotes a chapter to Rider. He considers Rider responsible for having successfully instilled an unfounded, long-lasting fear in library leaders about growth: “a fear of the demon Growth that was alive in the stacks, doubling relentlessly, a monstrous exploding pustule of cellulose” (81).

remained the primary form of flat sheet micropublication through the 1950s, when something better came along (microfiche).²⁶

An important aspect of Rider's enthusiasm and perhaps the source of his persuasive power was his clear belief that "micromaterials" were revolutionary for libraries, if people would only see them the right way: "No one seems to have realized," he wrote, "that for the first time in over two thousand years, libraries were being offered a *chance to begin all over again*" (Rider 1944, 93 [emphasis in original]). Admitting that microforms had been disappointing for libraries, he attributes that to library leaders' failure to see them as a "brand-new form, an utterly and completely and basically different form, a form that demanded ... an utterly and completely and basically different library treatment" (Rider 1944, 93). Microforms were not books but something new, and to see a microform as a book was to see it as problem making instead of problem solving. Microforms were not books but a solution to the problems book pose.

The Scientist and the Problem of the Library

Shortly before the turn of the twentieth century, Canadian radio engineer Reginald Fessenden tinkered with microphotography as a way to supplement his own overtaxed memory. He wrote: "The engineer sooner or later comes to the point where he can no longer rely upon his memory for the collection and preservation of his technical and scientific data."²⁷ An alternative to copying by hand, the photographic method he devised—from exposing, developing, fixing, washing, to filing away and labeling—made it possible, by his calculation, to copy 2,000 words in less than a minute. What might now seem thoroughly laborious, Fessenden experienced in 1896 as faster, easier, cheaper, and of greater quality than the other copying method he had at his disposal—his own hand. His experiment led him to speculate about broader applications of the method. Calculating that a box one-foot cube could contain a library of 50,000 volumes—or even more, if greater magnification ratios were employed—he concludes: "It is well within the bounds of possibility that the scientific student of the future will do his book work with the aid of a small projection lantern and a library of small positives" (Fessenden 1896). As such, microphotography could extend the capacities of the individual scholar or scientist and help in the management of accumulated literatures.

Four decades later, Vannevar Bush (1890-1974) expressed the same concerns but conditions had only worsened. Preoccupied with the overabundance of specialized

²⁶ Martin Jamison, "Fremont Rider's Pre-computer Revolution," *Libraries & Culture*, vol. 23 No 1 (Winter 1988), p. 10.

²⁷ Reginald A. Fessenden, "Use of Photography in Data Collections," *Electrical World* 28, no. 8 (August 22, 1896): 224. Also reprinted in Venear 1976, 96-99.

literatures for scientists, he sought to apply science to the “accumulated record” in order to find new, more adequate methods of organizing, sharing, and making use of scientific findings. “The summation of human experience,” he wrote in his famous essay “As We May Think,” “is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships” (Bush 1945).²⁸ He saw potential for microphotography much like those who had first witnessed it in the 1850s. The *Encyclopedia Britannica*, he envisioned, could be reduced to the size of a matchbox; and a “library of a million volumes...[could be] compressed into one end of a desk.” Taking almost for granted microforms as an effective means of large-scale storage, Bush wanted to build on them to find a way through their labyrinthine accumulation. To that end, Bush had dedicated himself to developing the microfilm “rapid selector” for the MIT library in the 1930s—a (failed) attempt to build a mechanized information retrieval system for searching through microforms.²⁹ Bush believed in its “revolutionary” potential to address the “information problem” that libraries faced, and to which they were ill equipped to manage. Krajewski suggests the Memex is an inheritor of the cross-referring card index (Krajewski 2011, 63).

Bush was a strong critic of the traditional library, with its orientation toward books. Knowledge was growing at a phenomenal rate and mechanization was necessary to keep up. It was imperative to remediate the means of storing “our findings and our thoughts.” His strongest criticism was that the hierarchical indexing schemes that made up library classification schemes were rigid, restricting, and incompatible with science: “No system of coding or of indexing under established subject leads can efficiently serve to guide us through the written thoughts or findings of scientists when the very science in which they work will not submit to definition” (Bush 1953). As when the Royal Society and its *International Catalog of Scientific Literature* four decades earlier, Bush wanted, in the place of static categories created by librarians, systems based on the scientist’s creativity and mental associations: “a kind of fluid indexing ... the pursuit of paths through a complex record by means of association of ideas” (Bush 1953, 151). At the very least, scientists should be able to specify subject terms for their catalogs and indexes. All of this can be seen in his description of the fictional Memex, a “mechanized private file and library,” that would draw upon mechanized microlibraries organized by means of the individual scientist’s own “associative

²⁸ Bush drafted his famous essay in 1939 but published it, essentially unaltered, only after the end of World War II. On its publication history, see Wright (2007), 196-97.

²⁹ Bush was not the first to develop a rapid selector. See Michael K. Buckland, “Emanuel Goldberg, Electronic Document Retrieval, and Vannevar Bush’s Memex,” *Journal of the American Society for Information Science*. 43, no. 4 (May 1992): 284-294. The fullest treatment of the rapid selector is Burke 1994.

trails" (Bush 1945). It is a dream machine for Bush by which the scientist might bypass the library all together.³⁰

In the 1930s and 1940s Bush and other reformers were involved in efforts to enhance the status of science in America. They knew that this required increased support for research, and libraries were crucial to that. European libraries were then superior in most disciplines, with American scholars cut off from research materials they needed. There was, as yet, no library system for science, not even an inter-library loan system. The public library system was extensive but decentralized, its professions weak, and oriented toward the "general user" rather than the elite researcher. There was no national library. The Library of Congress, the closest thing to it, was also weak and slow moving. The National Archives was in its infancy, having opened only in 1935. Scholars in the United States depended on European libraries. American libraries cooperated very little—no national cataloging system, only rudimentary interlibrary loan. Reformers wanted systems specifically designed for elite scientific researchers, whether they found themselves working in universities, corporate laboratories, or government agencies. They wanted alternatives to traditional library catalogs, their indexing methods, and, of course, the book itself, which the sciences moved away from in the nineteenth century as its central form for establishing authoritative knowledge.³¹

These reformers came from two groups, both outside traditional libraries: academic engineers like Bush; and entrepreneurial bureaucrats from the emerging federal science infrastructure. The latter would become known as the "American documentalists," after the European movement begun by Paul Otlet and Henri La Fontaine and discussed above (Burke 1994). Documentation took American roots in Washington, D.C., in the 1930s, led by Watson Davis (1896-1967), particularly with regard to the sciences, and would morph into "information science" after the Second World War. Davis, who had been trained as an engineer and was one of the first science journalists, became the second director of the Science Service, an organization founded in 1920 with money from E. W. Scripps, the newspaper publisher, to popularize science and to lobby for its support (Farkas-Conn 1990, 12). Davis wanted to build a new knowledge structure for science based on documentalist ideas. Those who joined in his effort—government librarians and science advocates—shared many of Bush's goals but worked for the most part separately from him.

³⁰ Even four decades later, Luigi Balsamo could assert: "Both the system of collecting and preserving the documentation that depends on the traditional library and the system of bibliographic information have become inadequate to satisfy the requirements of the sciences" (Balsamo 1984, 5).

³¹ Csiszar (2010) quotes Michael Foster, from 1903: "When a man of science writes a book, he writes, as a rule, either a textbook, in which original matter is out of place or even dangerous, or a lengthened essay in which he develops general views at a greater length than he is at liberty to do in a periodical; he rarely uses such a book as a means of making known his new results" (424).

Microphotography was of utmost importance to Davis. He had gone to the 1937 Paris Exhibition to spread the gospel of microfilm in Europe (Burke 1994) and inspired H.G. Wells, in particular (Wells 1938, 91).³² Davis had become excited by the new technique in the 1920s, as had so many others, and saw possible applications for science. He and chemist Edwin Slosson wrote in 1926 that a “method for putting books and manuscripts into compact and portable form by some miniature photographic process similar to motion picture films” was the most promising plan for improving dissemination of information (Schultz and Garwig 1969). He hoped to create a constantly updated world bibliography of science, with the abstracting and indexing done by authors—not unlike the scientists of the Royal Society and their *International Catalog of Scientific Literature*. He also wanted to reinvent scientific publishing. To that end, he initiated, or helped to initiate, a series of programs, including a “BiblioFilm service,” operated out of the Department of Agriculture library, which provided microfilmed copies of its holdings to scientists who had no direct access to its collections; and “auxiliary publication,” a micropublishing service to provide inexpensive but rapid publication of original research when necessary. Each of these services solved different problems of print publication: with the BiblioFilm service, the patron didn’t have to return anything (microfilm was a cheap duplicate); with the auxiliary publication, the author would no longer have to find a publisher or wait through a publication process.

³² From 1936 to 1938, in a series of essays, Wells articulated the barest outlines of an international, worldwide institution for the purpose of modernizing the “intellectual organization of the world.” He referred to it with a variety of names, such as “the new encyclopedism,” “superuniversity,” and “world brain.” As a supplement to the world’s existing institutions, this new institution would “hold the world together mentally” (Wells 1938, 21) and unify it. At the center of his vision, he, like Otlet, placed the new technique of microfilm as its enabling force:

[Microfilm] presages a real intellectual unification of our race. The whole human memory can be, and probably in a short time will be, made accessible to every individual. ... There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind ... (Wells 1938, 86-87).

This new supra-institution would be a “depot” where specialists would select, extract, and “very carefully assemble” all the world’s knowledge (Wells 1938, 20), enabling a synthesis of human knowledge in “continual correspondence with every university, every research institution, every competent discussion, every survey, every statistical bureau in the world” (Wells 1938, 69). And because microphotography made replication easier, cheaper, and more transportable, this “new all-human cerebrum” could have many instances throughout the world: “the direct reproduction of the thing itself can be summoned to any properly prepared spot” (Wells 1938, 86). It could exist fully in Peru, China, Central Africa, or wherever. Echoing Otlet’s universal book, Wells called his universal library “an undogmatic Bible to a world culture” (Wells 1938, 21).

His determination to build a new institutional infrastructure for science and his pursuit of microfilm-based library reforms led to the formation of a new organization, the American Documentation Institute (ADI) in 1937—specifically founded to “take advantage of micrographics” (Farkas-Conn 1990).³³ A joint effort of scientific administrators, librarians, archivists and scholars, the ADI concerned itself with building shared services, based on microfilm, to help solve problems then common to scholars and scientists. The group distrusted profit-motivated companies and sought control “solidly vested in America’s organized intellectual world” and not in commercially minded third parties (Schultz and Garwig 1969, 156).

Watson was, like Otlet, both a dreamer and someone who got things done. He believed it possible to build “one great library,” and like the European internationalists, he imagined an expansion of these services worldwide into an interconnected cooperative network: “All the libraries cooperating will merge into one world library without loss of identity or individuality.”³⁴ The big difference between him and the internationalists was that Davis was focused very specifically on science, whereas Otlet had sought to integrate all fields of knowledge. This focus on science proves to be a particularly Anglo-American determination: that the sciences required their own infrastructure, distinct and separate.

Circulation and Access

Davis’s innovative publishing programs were picking the fruit of decades of speculation. As early as 1880, in a patent application for a microfilm camera and reader, two California inventors expressed hopes for microphotography similar to those mentioned throughout this chapter, but with the twist that they imagined universal accessibility: Condensed, microformed copies “may be rendered universally accessible by reason of their cheapness, their small size and light weight, their imperishable character, and the facility with which they may be multiplied (Luther 1959, 95, and 108). Here we see, over 130 years ago, the very same wish that microforms, as a remediation of books onto film, would vastly improve the circulation and availability of books, indeed to such a degree that they might be not just more accessible but “universally” so.

Otlet had speculated as early as 1906 about the possible uses of microphotography to address problems of circulation (Otlet and Goldschmidt 1906), focused

³³ The ADI continued its work, eventually becoming part of the Library of Congress. In 1968 it changed its name to the American Society for Information Science.

³⁴ Watson Davis, “How Documentation Promotes Intellectual World Progress,” *The Science News-Letter* Vol. 32, No. 861 (October 9 1937): pp. 229-231; Watson Davis, “Microphotographic Duplication in the Service of Science,” *Science* Vol. 83, No. 2157 (May 1, 1936): pp. 402-404.

particularly on print publishing. The economics of printing books required that publishers print large runs of books ahead of time even though they cannot know with precision what the demand will be. This unavoidable problem of print publishing has long made inventory management a regular challenge and often leads to unsold books and financial losses. More important for Otlet, these factors combine to cause a book's premature unavailability if the publisher is not willing to reprint it. Access by way of a library might be considered a solution to this problem, but, he argues, it does not. Libraries are too few and too geographically dispersed. To make use of a book at a library also entails a lot of time, an often difficult journey, the risk of not finding the work once there, and, if that weren't enough, in order to make a copy of the parts of the book he or she might want to use, the researcher would have to do so by hand. The limitations of librarians, Otlet asserted, impeded not only research but progress: "Access to these libraries is not always easy and delays in securing works often discourage [even] the most tenacious workers, with great injury to scientific progress (Otlet and Goldschmidt 1906, 88). Libraries could not satisfy the problem of curtailed access that was a fundamental aspect of print publishing.

What was needed was a new method of publication, which microphotography seemed to promise. Otlet and Goldschmidt envisioned a new publication system of "microphotographic" books that were small and lightweight, uniform in size, permanent, moderately priced, easy to use, and available on demand.³⁵ Such a system could simply forego the "costly intermediate process of printing" and promised the "ideal copying machine" using film rather than paper (Otlet and Goldschmidt 1925, 206, 207). A remediated book that took full advantage of new microphotographic techniques would be a book that was easier to publish, cheaper to reproduce, and easier to find, buy, and generally make use of. The whole book system would function better with such an improved comprehensive method of publication.

Similar notions pervade the period, which saw rapidly proliferating photoreproductive techniques. The work of Robert C. Binkley and Eugene Power provide two important examples of people—one a scholar and one an entrepreneur—who found valuable new opportunities in new forms of reproduction. In late 1929, partly in response to the new attention given to microphotography, the Social Science Research Council and the American Council

³⁵ Although in 1906 Goldschmidt and Otlet were merely speculating based on some "experiments" they had conducted, in April 1911 the American magazine *Popular Mechanics* reported on what sounds like an actual, working prototype of the system proposed in 1906: "A Belgian scientist [presumably Goldschmidt] has constructed a microphotographic apparatus by means of which rare books, valuable documents, magazine articles and newspapers may be reproduced on minute photographic plates in which form they will require infinitely less space for storage than would be necessary for the originals." "Microphotographic Libraries," *Popular Mechanics*, April 1911, p. 518. Available online at <http://books.google.com>.

of Learned Societies jointly set up a Committee on Materials for Research.³⁶ The members of the Committee began with the shared belief that “our national machinery for collecting and preserving records is inadequate” (cited in Gitelman, forthcoming). Prominent among the Joint Committee’s concerns were the two vexing problems with publishing that I am outlining in this section (access) and the next (paper perishability). The Committee wanted to provide resources “to discover, select, edit, publish or otherwise reproduce basic data in the social sciences, which are difficult of access to students or likely to perish” (Binkley 1936, iii).

In 1930, the committee had historian Robert C. Binkley (1896-1940) compile a survey of methods for reproducing research materials. Binkley published it in two stages, a provisional report in 1931 and a final *Manual on Methods of Reproducing Research Materials* in 1936. The 1936 report is a remarkably meticulous compendium of then-current information about the various methods of textual production and reproduction along with cost analyses, samples of paper types discussed, and other assorted samples and exhibits, even small strips of sample film stapled to tipped-in pages.

Whereas the *Manual* lays out in precise detail the many available techniques of reproduction then available, his nearly concurrent essay in *The Yale Review*, “New Tools for Men of Letters,” explores the broad import of having so many different methods available to the researcher, whether professional or amateur. Binkley thought new techniques, most especially reprographic ones, would refurbish intellectual life. Fundamental to his view was a need to move beyond printing as the key technique for publication, because it imposed such significant limitations on reader and authors.

Binkley shared Otlet’s concerns about access to printed books. Although printing may have increased accessibility over the centuries, he argued, in the twentieth century it was having the opposite effect because the print publishing industry, technologically and organizationally geared as it is toward making profits on large print runs, requires a public or a buying readership inevitably larger than that addressed by the average researcher. In the end then, the demands of mass production diminish accessibility by excluding from possible publication those works that could or would not attract a sizable enough public. “Western civilization,” he complained, “now expects even poetry to fit the Procrustean bed of the publishing industry” (Binkley 1935, 185).

Another related issue for Binkley was that printing, like the telegraph, television, radio, and telephone, tended toward top-down cultural production. It

³⁶ On the Joint Committee, see also Peter Hirtle, “Research, Libraries, and Fair Use: The Gentlemen’s Agreement of 1935,” *Journal of the Copyright Society of the USA* 53:3-4 (Spring 2006–Summer 2006): 545-601.

“concentrate[s] the control of culture and professionalize[s] cultural activities” (Binkley 1935, 179). In contrast, what Binkley termed the new “graphic arts”—that is, forms of production based on the typewriter and photography—tend toward a decentralized and less professionalized culture. Microcopying, one of those graphic arts, makes possible the publication of specialized works because production would be cheaper and copies could be made on demand. Microcopying made it possible to “bring the Library of Congress to the small-town high-school teacher” and to “give the reader exactly what he wants, and bring it to him wherever he wants to use it” (Binkley 1935, 194, 184). Serving the writer who produces a “work of limited circulation” as well as the reader who seeks access to materials of interest to very few, microfilm could overcome the lumbering inefficiencies, the tendency toward massification, and the geographic inequities that had developed within the modern print apparatus. Furthermore, such a refurbished publication system would break the university’s monopoly on research, enabling a democratization of scholarship, a revival of the amateur researcher, and a fundamental “localization” (or de-centralization) of scholarly activity into what Binkley calls “local studies.” Such locally directed intellectual pursuits would revitalize education, improve critical self-consciousness in communities, and offset the attractions of the big city. He ends his essay hoping for a “scholar in every schoolhouse and a man of letters in every town” (Binkley 1935, 197).

It bothered Binkley that, when there were clear, superior alternatives, scholars had a “kind of fetishism” toward the printed page. He believed it was only tradition that protected the status of the printed book. Lurking within the *Manual* was his interest in promoting a wider range of methods for publishing that would not only more adequately serve both authors and readers but also nurture more authors and readers. He recommended forms of publication, summed up by Gitelman as the “typescript book,” that didn’t seek the same sort of “public” that the publishing industry required. He was advocating a new circulatory system for shared knowledge that was effectively “sub-print”—that is, more local, smaller scale, less mediated, and, in his opinion, better suited to the needs of scholars and researchers.

Access, as it concerns Binkley, has taken on a larger, more political set of meanings than it had in Otlet’s work. It is more focused on loosening the control over the production of knowledge from existing centralized institutions and expanding them to individuals and localities and, in so doing, suggesting a new public contrary to that developed in relation to what Benedict Anderson has termed “print capitalism.” Whereas Otlet wanted to “release” knowledge, Binkley wants to release individuals—especially those outside urban centers—into publication and/or authorship of a new sort.

Friend and colleague to Binkley, Eugene Power (1905-1988) was a pioneering entrepreneur in a new publishing niche industry known as “micropublishing”—i.e., microfilm used as a publishing medium—which would be brought to bear on some

of the key issues I have extracted from the writings Otlet and Binkley around access. In 1938, he founded the company University Microfilms, and, reversing the economic pattern of traditional publishing, made a successful business practice out of publishing “editions of one” (Power 1990). The inverse of print publishing, micropublishing produces single copies of a large number of titles as opposed to producing large quantities of a single title. Like Binkley, Power believed that print publishing, by deeming much research unpublishable, was hiding knowledge from human use (Power 1990, 17). He had gotten the idea for producing copies of academic material in small quantities—as few as one, on demand—at a 1931 meeting in Cambridge, Massachusetts, organized by none other than Robert Binkley, to discuss methods of producing scholarly material. They had met earlier that year, after Binkley sent a copy of his first 1931 report (discussed above) to the Ann Arbor printing company Edward Brothers, where Power worked as a salesman, asking for assistance. They quickly became friends, and Binkley a couple years later introduced Power to microphotography. He immediately saw microphotography as the technique that would make his “edition of one” idea possible (Power 1990, 25-27).

Power decided to try out the new publishing method on a selected list of books from Pollard and Redgrave’s *Short Title Catalog of Books Printed in England, Scotland, and Ireland from 1475 to 1640* (STC), which he had learned about while working on an Edwards Brother printing project with a University of Michigan English professor. In 1935, he took his custom-made microfilm camera to the British Museum, where the books would be photographed.³⁷ Then Edwards Brothers would sell subscriptions to libraries, which would pay \$500 a year for 100,000 pages of STC titles. Power eventually left to start his own company, University Microfilms, with the STC project as the base from which he would build his business.³⁸ His next big project was to develop an economic means of publishing doctoral dissertations, on microfilm. Dissertations were impossible to distribute commercially, but they contained valuable research. If it could be possible to publish them on demand, one at a time, then universities, researchers, and

³⁷ Power claimed that his microfilm book camera was only the second in existence (Power 1990, 29). The first was R. H. Draeger’s. A captain in the US Navy, Draeger wanted to take hundreds of books with him on his assignment to China and so designed a camera that held 100 feet of roll film that advanced automatically after each exposure. He mounted the camera over a flatbed, on which an opened book could be pressed flat beneath a glass cover. He could later enlarge the film and print on paper (Power 1990, 25). The camera was later given to the Department of Agriculture library, where Watson Davis was using it for his BiblioFilm service (on Davis, see above). Power also met Watson Davis (with his book camera) through Binkley. Power was among the founding members, with Davis and Binkley, of the American Documentation Institute.

³⁸ The STC project continued to grow over the years and itself deserves a chapter of its own. It was the first “ebook” corpus sold to libraries, as microfilm, and was eventually digitized to become Early English Books Online. On EEBO, see: McLean 2001; Williams and Baker 2001; Kichuk 2007; Gadd 2009; Keegan and Sutherland 2009.

scholarly authors would all benefit. Power microfilmed the dissertations at a small cost to the universities and sold printed copies of them as readers requested them. This effort too became successful.³⁹ Power had succeeded in finding ways to implement the expressed hopes of Otlet and Binkley, increasing the access American scholars had to European collections as well as enabling the very specialized research of Ph.D. dissertations to circulate much more widely. Power sits at the beginning of some long-term developments that today present new problems for library administrators and scholars, which the case of book digitization makes evident: 1) the library subscription-model business, which has come to dominate library acquisitions; and 2) the commercialization of library materials by capturing them in a new format and then selling them back to libraries.⁴⁰

Paper Perishability: From Preservation to Access

Power also put microfilm into the service of another important concern in the late 1930s, beyond that of increasing access, namely, the preservation of European culture from the destruction of war. As a measure to prevent the loss of European culture, should libraries be destroyed in the War, he traveled throughout Europe photographing important collections.⁴¹ On the brink of a second world war, the young American entrepreneur, with his new heroic technology had come to the rescue of a fragile, beleaguered Europe and its precious knowledge form: the book.

But it was preservation against a different threat that would come to play an important role among research libraries in the twentieth century: that threat was “perishable paper.” Concern for the problem had catalytic effects, reshaped old practices, and, as I will show, became fundamental to the development of mass digitization. What begins as a grave concern for the preservation of the past ends as an expansive future-oriented vision concerned with institutional renewal.

What was “perishable paper”? As paper production industrialized in the early nineteenth century, the much more abundant and cheaper wood pulp was devised as a replacement cotton or linen rag. Chemicals used in the wood pulp papermaking process (especially alum) react to humidity and over time this interaction, especially in less-than-optimal storage conditions, will weaken the cellulose fibers, leaving the pages dried out, cracked or crumbling at their edges.

³⁹ Today, ProQuest—a successor company to UMI—remains the pre-eminent publisher of doctoral dissertations.

⁴⁰ Brewster Kahle has characterized Eugene Power as a “pioneer in the locking up of the public domain.”

⁴¹ Queen Elizabeth knighted Power in the 1970s for this preservation work. See Power 1990. Appendix H, p. 420.

Another wood pulp problem in some papers was the presence of the fiber lignin, which causes yellow or brownish discoloration after exposure to light. Even before these explanations were arrived at, nineteenth-century scholars, preservationists and others around the world had begun to notice the deterioration and to worry about not just the durability but also the permanence of print.⁴² Indeed, in 1906, Otlet and Goldschmidt's essay "On a New Form of the Book" (discussed above) listed "permanence" among the desiderata for their "new form."⁴³ It was the now-evident perishability of paper that had brought into consciousness the newly important quality of permanence.⁴⁴

As Binkley identified in his essay on the topic—which was originally an address to the First World Congress of Libraries and Bibliography in Rome—there were two challenges, one prospective, the other retrospective: what to do going forward; and what to do with the legacy of publications that were now in need of rescue (Binkley 1929). For the former, new standards need to be developed and pressed upon publishers so as to make sure materials intended for preservation were produced on durable materials; and for the latter, the already existing impermanent records needed to be copied (i.e., reformatted). In 1929 Binkley called for institutional cooperation toward a technocratic solution that could forestall another looming civilizational crisis (like library growth):

When we know finally the best possible ways of preserving perishing materials, a vast task of organization will lie before us. It will be necessary to prevent the wasting of effort by unnecessary duplication of the work of preservation. A wise coordination of the salvaging efforts of the libraries of the world, counseled by the scientists as to the technique of preservation, and by the representatives of all scholarly and intellectual interests as to the selection of what is worth preserving, may then recover for civilization what a generation of thoughtless publishing practices have threatened to lose (Binkley 1929, 178)

⁴² See Binkley 1929 in Fisch 1948. Robert P. Walton, "Paper Permanence: The Physical and Chemical Factors which are Limiting the Life Span of Modern Books," *Bookmaking* September 7, 1929, pp. 979-83; Edwin E. Williams, "Deterioration of Library Collections Today," *The Library Quarterly* 40 (1) January 1970: 3-17; and Smith 1999.

⁴³ Other desiderata were that the "new form" be light, compact, of uniform size, cheaper, and easier to preserve, consult and reproduce.

⁴⁴ Though books had long had enemies. William Blades lists them as fire, water, gas and heat, dust and neglect, ignorance, bookworms and other vermin, bookbinders, and book collectors. William Blades, *The Enemies of Books*. London: Trubner & Co, 1880. The copy of this book that I found in Berkeley's Doe Library is a reproduction bearing this comment in the front matter: "LBS Archival Products produced this replacement on paper ... to replace the irreparable deteriorated original. 1989." The 1989 copy also includes some excised plates from the original edition.

Books were in need of rescue and remediation—and on a large scale. From this statement from Binkley in 1929 forward until the 1988 NEH-funded Brittle Books program, large-scale retrospective conversion of books using microfilm was contemplated, studied, planned, and pursued.

For reasons that are not clear, microfilm in the early twentieth century was presumed to be durable, permanent, even “indestructible,”⁴⁵ and so microphotography presented an alternative medium and possible substitution for wood pulp paper. But, in fact, early microfilm was plenty destructible. It employed the same film as early motion pictures, or nitrate-based stock, which is highly flammable. Its replacement, acetate film, too suffered an irreversible degradation sometimes called “vinegar syndrome.” It is now believed that pre-1980 film cannot be expected to last more than 100 years.⁴⁶ Nonetheless, the presumption of microfilm’s permanence irreversibly construed “modern paper” with its tendency to decay—in contrast to film—as “unstable,” impermanent,” and “of inherent vice.”

Library collections in the 1930s, obviously, held a great many books printed on wood pulp paper and they felt a responsibility to care for the continued integrity of their collections. As Binkley indicates in his 1929 essay, the concern was well observed but ill understood. For a variety of reasons—World War II prominent among them—it wasn’t until the 1950s that libraries launched a concerted effort to deal with the problem. The Council of Library Resources was founded, with Ford Foundation money, in 1956 to foster cooperation among libraries toward tackling the major problems facing libraries (such as library growth). It commissioned a study of the causes of paper deterioration (from William Barrow) and, on that basis of that, established one task force and then another. In 1962, a separate group, the Association for Research Libraries, commissioned noted librarian Gordon Williams, who would later supervise the completion of the National Union Catalogue, to do a large-scale preservation survey of its member libraries. In his 1964 report, Williams projected that more than 40 percent of the books printed after 1900 would be unusable within twenty-five years, and 90 percent would be unusable within seventy-five years (ARL 1964).

Counteracting this trend and making sure that these books remained available was, Williams wrote, “essential to the continued progress of society” (ARL 1964, 1). Although deacidification—the exposure of the deteriorating books to solvents that

⁴⁵ At the 1938 New York World’s Fair, a time capsule was buried, not to be unearthed until 8113 A.D., which contained 22,000 pages of “indestructible” microfilmed text from a variety of publications. Allen B. Veaner, “Micrographics: An Eventful Forty Years—What Next?” in Veaner 1977, pp. 467-81.

⁴⁶ Today, the only film considered archival is silver halide emulsion on polyester film, with a life expectancy of 300-500 years (under optimal storage conditions).

would neutralize the acids and retard decay if not restore the book—had been identified as one solution, Williams dismissed the labor-intensive process as too costly. The only real option he saw was large-scale reformatting—that is, transferring books into another medium (i.e., microfilm). The paper perishability problem began compelled a shift from traditional preservation, which had operated at the individual book level—which today would be closer in meaning to “conservation”—to preservation on a large, even industrial scale.

Williams proposed a plan that could be taken, in retrospect, as an early blueprint for mass digitization. Like Binkley three decades earlier, his report expressed the need for a large-scale multi-institutional effort, though the scope was now resolutely national in contrast to Binkley’s internationalism. The goal would be to preserve at least one copy of the “intellectual content” of every significant book. Williams recommended a national program administered through a central agency that would have a number of responsibilities: overseeing the microfilming of the books; disseminating microfilm copies out to libraries; maintaining a library of original books from which a copy had been made; and establishing and administering a system of bibliographic control akin to the National Union Catalog so as to prevent the duplication or omission of books. Finally, because the preservation of and access to books was “of national interest,” Williams urged federal support for such a program.

Williams’ call for national administration was heard but did not result in any federal initiatives (Bello 1986). Meanwhile, library groups continued to study the problem and commission further reports. The goal became a coordinated nationwide microfilm preservation program but Williams’s expectation that the federal government, namely the Library of Congress, would play a leading role was abandoned in the 1970s in favor of a decentralized, distributed effort. Preservation, meanwhile, became a new field of expertise. To combat what was a “primitive state of knowledge” about preservation practices, Columbia University opened the first formal training program for “preservation administration”—a new field in library science—in 1981 (Bello 1986). The Research Libraries Group (RLG), begun by four elite research libraries (including Columbia) in 1974, achieved the most in terms of coordination. They intended their Cooperative Preservation Microfilming Project as “a practical model for a nationwide, coordinated preservation program” and in the 1970s and 1980s they made very significant progress (Bello 1986, 14-37). And, finally, in response to continued calls from the library community, the National Endowment for the Humanities opened an Office of Preservation in 1985 to “help save the content of deteriorating humanities resources in the nation’s libraries” (Bello 1986).

In 1986, the CLR produced a report, *Brittle Books*, summarizing its conclusions from a series of meetings it had sponsored from 1982 onward. In that time, the mission of to save “brittle books” had, importantly, expanded from mere

preservation to encompass both preservation and access. In another CLIR-commissioned report, which offered the empirical data for determining the scope of a national effort, its author, Robert Hayes, noted at the outset an important qualification: "During the meeting at CLR, it was emphasized that the primary focus in the preservation program is on ACCESS. That means that the preservation of the 'artifact,' while perhaps important overall, is not the focus of attention at this time" (Hayes 1987, 1). The *Brittle Books* report itself explained that preservation alone was not a goal libraries could justify in light of the required resources for a national program: "While preservation, per se, is a valid goal, it is the prospect of providing wider and more equitable access to a growing collection of preserved material that fully justifies the cost and effort" (CLR 1986, 16). Preservation had been redefined as access.⁴⁷

Between the 1950s and the 1990, the research library community managed to foment a shared purpose from the germ of paper perishability, a vision for a cooperative national microfilm preservation project that would propel them forward, helping them both to manage their increasingly troublesome collections but also to re-imagine their futures. Much more was at stake than the preservation of decaying books. What had begun as the ARL/AAU Task Force on Preservation in 1979 became, in 1986, the Committee on Preservation and Access. Its first president, Patricia Battin, spoke much more boldly than the sober *Brittle Books*. She saw the paper perishability problem as an opportunity "of monumental proportions." Libraries needed to use paper's perishability "to redefine the preservation problem from a single-item technical solution to a broad strategy for providing access to the human record *as far into the future as possible*." And she went further: "The new technologies now provide us with a palette of multiple capacities and the opportunity *to re-examine all our assumptions ...* for the provision of access to knowledge" (Battin 1991, emphasis mine). This reconceptualization not just of preservation practices but also of library collection management continues to form a central thrust of mass book digitization. In the reformatting of books, libraries themselves, those "keepers of books," might find their future—indeed their own remediation.

Six decades after Binkley called for a "vast salvaging effort," libraries finally won, in 1988, a twenty-year commitment from the National Endowment for the Humanities (NEH) to fund the reformatting of at least three million volumes published from

⁴⁷ Recently, the administrator who oversees the University of California library system summed this up, somewhat derisively, as the "care and feeding of books": "Libraries got caught up in the care and feeding of books because, in the print environment, that was essential to delivering access: to get access to information, you had to be near it. But libraries are fundamentally about access." Daniel Greenstein, quoted in "The Library as Search Engine," *The Chronicle of Higher Education*, January 5, 2007.

1865 to the present.⁴⁸ But was the effort a Pyrrhic victory? Libraries had won recognition and support for their grand strategy of microfilm reformatting, in 1988, which was precisely the moment that the term “digital library” appeared and when their own attentions were drifting off toward digital forms of image capture, if only speculatively. The term appear in the late 1980s, around the time Robert Kahn and Vinton Cerf published their report, *An Open Architecture for a Digital Library System* in 1988 (Stefik 1996). Written as the Internet was being opened for broad use beyond universities, the computer scientists’ report envisioned a distributed network of personal, public, commercial, and national “digital libraries” sharing common standards and methodologies. New grand visions would soon accompany the World Wide Web. (By this time, Otlet had been forgotten.)

In 1993, three large federal funding agencies, the National Science Foundation (NSF), the Advance Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA), together launched the influential Digital Library Initiative (DLI) to fund research on digital libraries. It was successful enough that its Phase II began in 1998, funding in part the original Google search engine prototype.⁴⁹ In 1995 alone, there was an explosion of movement to solidify an institutional structure for thinking about and developing “digital” libraries. That year witnessed the launch of the Digital Library Federation, *D-Lib Magazine*, the Making of America book digitization project, JSTOR, the Library of Congress’s National Digital Library Program (also known as the American Memory Project), and mass book digitization was set in early motion. National strategizing over “national infrastructure” centrally involved digital libraries, bringing together Big Science and library science over a key term, which meant something different to the two groups. The librarians’ trouble with the book was of course not over but redescribed for a new era: “The book is a marvelous technology for use, but it is a cumbersome dissemination format and an increasingly frail storage format in this age of rapid telecommunications” (Battin 1991). With microfilm set aside, along with a certain narrow view of preservation, all things digital, with access as its new keyword, would become the locus of responses to the continuing problem of the book.

Conclusion

⁴⁸ The NEH website today reports that one million “brittle and disintegrating books [have been] saved by NEH-supported preservation microfilming funding.” <http://www.neh.gov/whoweare/divisions/PreservationAccess/WhatWeDo.html> Accessed January 9, 2012. The last of the Brittle Book grants was completed in 2007.

⁴⁹ Google co-founder Larry Page developed the Web page-ranking prototype, BackRub, as a graduate student working on the Stanford Integrated Digital Library Project, a DLI-funded project. See David Hart, *On the Origins of Google*, National Science Foundation (Aug. 17, 2004), http://nsf.gov/discoveries/disc_summ.jsp?cntn_id=100660&org=CISE.

As the Internet does today, microforms required new reading systems, which involved cameras, film, enlarging machines, lighting, bibliographic control, and so forth. Although all aspects of the microfilm reading system went through many changes and improvements from the 1920s to the 1980s, and although many different forms were devised (e.g., microprints, microcards, microfiche, microdot, etc.), microphotography presented persistent problems. Most significantly, it was very hard to use microfilms and, despite a great deal of evidence of this dissatisfaction, it took a long time for microfilm advocates, so strong was their commitment, to admit it.⁵⁰ It also failed to solve problems of what would become known as information retrieval, which had been prominent among the problems it was looked to to remedy.

By the 1970s, microforms had already disappointed most of the grand expectations for it and new visions, however incipient, had started to revolve around the digital computer, even if microfilming held on. Far short of early hopes and expectations, microfilm settled into providing three main functions with regard to books: an archive medium for deteriorated books or newspapers; a circulation medium for very old, fragile, or rare books; and a publishing medium for “editions of one” like Ph.D. dissertations. But even these functions have been or are being replaced by digital imaging.⁵¹

Indeed, the book persists and, for that reason, it is ever more the problem that it was, but for the fact that it is less imaginatively central as it was a century ago. It no longer represents “how things are” or even “how things were” but rather the force of tradition, of things that persist. Meanwhile, as the ragtag microfilm archive is slowly being reformatted yet again—into digital form—and journal articles have long passed the tipping point from print to electronic circulation, mass book digitization once again suggests that epochal threshold beyond which the book will be standardized and rationalized, as well as made permanent and universally accessible.

I have assembled here, in this partial, pre-digital history of digitization, evidence for a “critique” of the book and its institutions, especially the library, that well predates the computer, the Internet, and the World Wide Web. I have also tried to show how a variety of different actors came to see reformatting as a solution to these problems. In their abundance, books had, ironically, been construed as a threat to

⁵⁰ Venear, “Micrographics: An Eventful Forty Years—What Next?” in Venear 1976, 467-81. Harold Wooster, “Microfiche 1969—A User Study, (“Wooster Report”) July 1969; Steve Salmon, “User Resistance to Microforms in the Research Library,” *Microform Review*. Vol. 3, no 3 (July 1974): 194-199.

⁵¹ See the ARL report, *Recognizing Digitization as a Preservation Reformatting Method*, which tipped the scales toward digital preservation. 2004 paper prepared for the ARL. Available at: <http://www.arl.org/preserv/digitization/index.shtml>

those who cared about them, and, indeed, contrary to their purpose. Progress could become regress, if action were not taken. Burgeoning science had abandoned the book as a form but hadn't yet settled on what could replace it. The needs of this search among scientists and their advocates had broad effects on the book apparatus, which continues. Elites concerned with institutions of knowledge wanted to get more from what they had: they want to improve and reform. They wanted to get inside books and open them up to manipulation, reuse, and recombination. They wanted to facilitate greater, easier, and more creative access to what they contained. Others like Binkley and Power wanted more people to have access not only to books but also to publication and/or authorship—a trend Walter Benjamin was simultaneously bemoaning (Benjamin 1935, 33). They saw the problems in movement, distribution, and existing publishing structures as inhibiting progress and contrary to democracy. And, finally, libraries had a particularly strong attraction to microfilm as a means to solve one problem but also, with the echo of Fremont Rider reverberating, to open a door to institutional reform and rejuvenation. “Libraries were being offered a chance to begin all over again” (Rider 1944, 93).

Although it is tempting to wonder, as Nicholson Baker would have us do, whether image-based book digitization isn't just the next outfit for the emperor—another foolish flash in a technocrat's pan—it has not been the purpose of my inquiry in this chapter to arrive at such a judgment or to analogize the practices so utterly. Rather, I have sought to use the past to render the present both more complex and more visible.

Chapter 2

The Matter of the Digitized Book

Chapter 1 ended with problem of perishable paper and how it grew, over the course of the twentieth century, into a collective effort among U.S. research libraries to undertake the large-scale reformatting of books with the purpose not of *preserving* a minority of deteriorating books but of *increasing access* to the majority of books. But paper today remains a motivating concern for digitization. The problem, however, is no longer paper's perishability, as that problem has been solved,¹ but its environmental impact. In the past few decades, paper has been reconstrued as the product of a nasty industrial process—involving bleach and chlorine to whiten paper; energy consumption not only during manufacture but more significantly during shipping and shopping; water usage during pulp processing; petroleum-based ink; and, of course, the felling of forests—to which “greener” alternatives should and can be found. Whether it is a company asking us to “go green” with paperless billing or to “save a tree” by foregoing a print publication, part of the habitus of the middle-class consumer in the United States is to be conscious that paper consumption is bad for the environment and that one should act accordingly. Commonly, and especially and among technologists, terms like “dead-tree books” are used as a disdainful contrast in the face of digital alternatives. These contrasts seem meant to suggest that printed books involve physical matter and e-books somehow do not. And one can't help but sense in such phrasing “dead-tree book”—common among the engineers I worked among—a tinge of judgment as a cruel and unnecessary arborcide is evoked.²

Let us look at one example of digitization seen as environmentalism. In July 2009, the director of strategic web communications at a publisher gave a speech to trade gathering. The speech, entitled “Scholarly Publishing in the New Era of Scarcity,” urged its audience of publishers to embrace digital publishing as soon

¹ In 1984, a consensus national standard was established (ANSI NISO Standard Z39.48-1984) for the “permanence of paper for publications and documents in libraries.” The symbol for paper permanence is the mathematical sign for infinity. A bit later, in 1990, the Federal government required that all of its publications be printed on “acid free” paper. Indeed, today, much commercial paper is acid free, though mostly likely to do with new paper processing methods and not preservation concerns.

² Indeed, the book publishing industry uses 30 million trees a year to make the books sold in the United States. For more information, see The Green Press Initiative website: <http://www.greenpressinitiative.org/about/bookSector.htm>.

as possible, not just for the sake of their own survival as businesses but, more gravely, for the sake of civilization. I quote it at some length:

...[T]he lifecycle energy and CO2 costs of printing, shipping, storing, and distributing physical books must be radically curtailed. ...Within the context of a world in crisis, we *must* demonstrate that we're radically rethinking our relationship to the future. We must demonstrate that we are part of the solution, not part of the problem. We must seize initiative now, and start making changes as fast as we can.... If we don't make these kinds of changes, we will be knowing participants in the death spiral. ...We must ... brand ourselves as becoming part of the CO2 solution, to our administrators and institutions, as part of their external messaging campaigns. Brand ourselves with the public as a key part of a civilized world trying to save itself.³

In addition to an echo of the twentieth-century observers discussed in chapter 1, who found the proliferation of books a threat to civilization, we find in this inflated rhetoric a condensation of some key features of those who populated my field sites: a sense of urgency and the need for quick action; high-mindedness; a conflation of “innovation” with progress; an assumed purpose to educate the uninformed or technically naïve and pull them along; and a yearning for a “transition” away from the “heavy industry” of print to “CO2 solutions” of networked computers. In hastening the collapse of print publishing, we will stave off *environmental* collapse; in the creative destruction that digitization brings with it, we renew ourselves.

This chapter will not establish whether or not e-books are “greener” than printed books.⁴ But I do want to counter the tendency, on display in the speech excerpted above, to regard digital information as immaterial and digitization—in the slogan of MIT Media Lab director Nicholas Negroponte—as a move “from atoms to bits” (Negroponte 1995). Recent scholarship has started to address this dominant “mysticism”, “ideology,” or “trope” of immateriality (Van den Boomen et al 2010; Kirschenbaum 2008; Blanchette 2011), in reaction to the rather bizarre presumption that somehow matter has been overthrown—what Bill Brown calls the “dematerialization hypothesis” (Brown 2010). The

³ The entirety of the speech, with slides, can be seen here: <http://www.nap.edu/staff/mjensen/scarcity.html>.

⁴ The following documents report on the state of research on the matter: Green Press Initiative, “Environmental Impact of E-Books,” available at: <http://www.greenpressinitiative.org/>. Don Carli, “Is Digital Media Worse for the Environment than Print?” PBS Mediashift, March 3, 2010, <http://www.pbs.org/mediashift/2010/03/is-digital-media-worse-for-the-environment-than-print090.html>; Daniel Goleman and Gregory Noris, “How Green Is My iPad?” *New York Times*, April 4, 2010.

dematerialization hypothesis is not simply digital boosterism but, more significantly, as Brown points out, part of a long analytical tradition that characterizes modernity as a process of increased abstraction. (Even environmentally minded researchers refer to digitization as de-materialization.⁵) One history of media would be a history of repeated perceived dematerializations, wherein form is continually divorced from an originary materiality. The pessimistic versions of the dematerialization hypothesis presume a former materiality that once nurtured more intimate and more meaningful connections. In the case of books and digitization, this perspective has been widely expressed by literary commentators (e.g., Birkerts 1994; Gass 1999; Franzen 2012). The optimistic versions of the dematerialization hypothesis are progressivist, presenting digital forms, in particular, as promising to emancipate humans from the shackles and accreted burdens of older forms. In the realm of digital media, this has also been amply expressed (e.g., Kelly 2006; Negroponte 1995; Barlow 1996).

The people with whom I conducted my research would certainly fall into this progressivist camp, but they could hardly be said to believe that digitization is a de-materialization, laboring as they are in the utter materiality of the enterprise. My fieldwork at the Archive presented me with daily lessons in the material burdens of “the digital” at all levels: the physical network that keeps it connected and running; the software applications; and the content that it held and served to the public. One machine might be running slow, while another wasn’t working at all, and yet another “was sick.” Disks died, machines overheated, APIs overloaded, functions broke. “Evil” machines “threw red rows,” users confronted “404s,” Nagios alerts fired.⁶ Indeed, at times the Archive’s distributed cluster of spinning disks seemed like it might as well have been room of spinning plates: one (or two or three) is always just about to fall to the ground and catching it requires a skilled and unstinting vigilance. I don’t think that the Archive is unusual in this regard. Opportunistic metaphors such as “cloud” aside, digital networks depend upon a material infrastructure of human labor, built environments, extracted resources, expertise, institutional configuration, and much more.

⁵ For example, see: Asa Moberg et al. “Books from an environmental perspective—Part 2: e-books as an alternative to paper books.” *International Journal of Life Cycle Assessment* (2011) 16:238–246; Greenpeace, “Make IT Green: Cloud Computing and Its Contribution to Climate Change.” March 2010, p. 10. Available online at: <http://www.greenpeace.org/international/en/publications/reports/make-it-green-cloud-computing/>; and Kris De Decker, “The Monster Footprint of Digital Technology,” *Low-Tech Magazine*, June 16, 2009. <http://www.lowtechmagazine.com/2009/06/embodied-energy-of-digital-technology.html>

⁶ Indications of errors. 404 (or “error found”) is the HTTP standard response code indicating that the client was able to communicate with the server. A red row is an internal indicator in the Archive’s catalog that the processing of a book (or other item) has failed. Nagios is infrastructure monitoring software.

That said, on the one hand, it seems trivial to point out that digitization is *not* a transcendence of matter. Does anyone really doubt that any human practices are not always and necessarily materially embedded? And yet, on the other hand, my work at the Archive, which often involved interrupting engineers trying to trouble shoot and fix things, seems an opportunity not simply to point out that digitization is a *re-materialization* but also to demonstrate *how*, at least in the case of the mass digitization of book, digitization *re-materializes* and what that rematerialization comprises.

Materiality has multiple registers. Bill Brown has suggested that an “ideal” materialism would account for the phenomenological (the interface between a user and a technology); the archaeological (the physical infrastructure of the medium); and the sociological (the cultural and economic forces that shape a technology) (Brown 2010). This ideal portrait, however, presumes some stability in the very “medium” itself. If the digitized book, or any electronic book, is a “medium,” it is a fragile, emergent, or possibly even a provisional or experimental one. Indeed, it is the task of this chapter to do what I can to assemble the digitized book into an object for analysis. Although literary scholars and historians have commented in a variety of ways on the digitized book as a representation (Mak 2011, Duguid 2007, Deegan and Sutherland 2009, Townsend 2007, Musto 2009), no scholar has presented an accounting of a digitized book as a digital object. So, rather than consolidate our understanding of the “digitized book” as a medium in its “ideal” materiality, in this chapter I seek to *constitute* it as a complex analytical object. In so doing, this chapter contributes to a nest of concerns Gabriella Coleman has termed the “prosaics of digital media” (Coleman 2010): the conditions in which digital media are made and made use of; the technical protocols, infrastructure, and platforms that enable and constrain the circulation and use of digital media; and their actual and material day-to-day operations. My work at the Archive reveals an as yet emergent, unstable, imperfect and vulnerable digital knowledge system.

The Digitized Book

From the perspective of the machine, the outcome of book digitization is not a book but a collection of bits stored in a computer. As Negroponete writes: “A bit has no color, size, or weight” (Negroponete 1995). However, if all we had were the bits, we would have nothing. Digital information is nothing without some constraining matter. But how do we get from a printed book to “bits” and back to a book again, and what else happens along the way? The design of digital objects—whether documents, websites, geographical information systems, JPEG images, video games, and so forth—is specific, contextual, and highly variable. It depends, of course, on the choices those who create them make. As I describe here in detail the choices the Internet Archive has made in the design of its

digitized books, I will be guided by digital preservationist Kenneth Thibodeau's useful set of distinctions for thinking about digital objects. He describes digital objects as three different, mutually constituting types of objects: as physical objects, logical objects, and conceptual objects: "A *physical* object is simply an inscription of signs on some physical medium. A *logical* object is an object that is recognized and processed by software. The *conceptual* object is the object as it is recognized and understood by a person (Thibodeau 2002, 6).

a. *The Digitized Book as Physical Object*

As a physical object, the digitized book is a collection of bits, a digital inscription stored on magnetic disks in the Internet Archive's data centers. It is of no consequence to the machines if those bits form a "book," a "movie," or a "photograph." Throughout the time I was conducting my fieldwork, the Archive ran on a data center of approximately two thousand computers distributed across two data centers (in San Francisco and in Redwood City, CA). Sun Microsystems hosted a third data center on its Santa Clara, CA, campus.⁷ Engineers refer to the computers simply as "machines," and the totality of the Archive's distributed storage system is called the "Petabox cluster" or "the cluster," for short. From one important perspective, the Archive is just this: a collection of server racks (as seen in Figure 1), each of which holds a number of stacked machines. In other words, it is a distributed data center. The cluster is a custom-engineered modular data storage and processing system using commodity hardware and open source software (Ubuntu/Debian), as shown in Figure 1. Designed for the purposes of an archive, the cluster is optimized for low power consumption, high storage

⁷ Sun made this special modular database from a shipping container as an experiment in data center portability. Just a few months after it was completed and began to run, in March 2009, Sun was bought by Oracle, and, at the time of writing, the Archive was having to move the 3+ petabytes of data from the former Sun campus, a delicate operation that betrayed the intended purpose of portability.
http://www.computerworld.com/s/article/9130499/The_Internet_Archive_s_Wayback_Machine_gets_a_new_data_center



Figure 1: The “red boxes” that used to form the Archive’s machine cluster. This photograph was taken in its former data center in downtown San Francisco. Photograph courtesy of the Internet Archive.

density, and low-cost disk storage.⁸ This means it is *not* optimized for data processing, though every machine serves both as a processing unit and as storage. The favoring of archive storage over computational processing proves a challenge to Archive engineers who might want or need to work over large data sets (such as indexing content for full-text search, to take one example). Any process that is “computationally expensive” requires special handling.

During the course of my time working at the Archive, the machines in the cluster were being redesigned. The former cluster, implemented from 2004 to 2006, was composed of storage racks (known in-house as “red boxes”), seen in figure 1, which each held forty machines with four hard drives each. The new system, which was in the works throughout the time of my research, was finally implemented in 2010 and 2011. Its new racks house ten machines each with thirty-six hard drives each. (Total storage capacity depends on the hard drives,

⁸ At the time of writing, the Archive was buying most of its 3TB hard drives at Costco because they had the lowest price.

which depends on what's available commercially. As of writing, 3 terabyte drives (3TB) are the largest available; 4TB are anticipated soon.) The new system enables greater amounts of data storage while reducing the number of machines and racks nearly tenfold. This new cluster is the Archive's fourth generation of machine, part of the perpetual practice of migration. Machines have a relatively short life span, while storage capacities continue to expand.

But, beyond their greater storage capacity and efficiency, the new machines also satisfied Brewster Kahle's aesthetic wishes. The red boxes of Figure 1, he felt, were not just loud but ugly. In redesigning them, he wanted to create a "beautiful" machine for his digital library; he wanted "books that glow." Taking the Jedi Library in *Star Wars: Episode Two* as his point of reference, Kahle sought, with the new racks, a "machine you'd want to be in the same room with," in preference to a machine that you put in a basement, a closet, in a shipping,



Figure 2: Screenshot of the Jedi Library from *Star Wars: Episode 2: Attack of the Clones*. The ideal for Brewster Kahle's redesigned machine cluster.

behind glass—anywhere so as not to have to look at it or suffer its heat and noise. Kahle wanted people to be able to—or, more, want to—occupy the same space as the machines, to work alongside them, as one might read or study among books in the open stacks of a library. At a meeting in the offices of the Archive, where he was pitching the idea to some visiting hardware designers, he scribbled some notes on a whiteboard. Conflating books with machines he referred to server racks as "bookshelves" that would each hold ten rows of thirty-two drives (or "books") each. After two years of planning, the end result was something that wasn't quite as silent as Kahle had hoped for, but the new



Figure 3: “Books that glow.” Part of the Archive’s new data cluster, situated in a niche in the back of the former sanctuary of the Christian Science church that now serves as the Archive’s headquarters in San Francisco. Photograph courtesy of the Internet Archive.

Machines do glow. The new servers, shown in figure 3, now feature small blue lights that illuminated each time an item is uploaded to or downloaded from the Archive. They are also scattered throughout the Archive’s headquarters. Figure 3 shows one group of machines situated meaningfully in a niche in the former sanctuary of the Christian Science church that now serves as the Archive’s headquarters. Kahle finds the Church a fitting home for his digital library as libraries are “cathedrals of learning.”

The cluster is in many ways the heart and lungs of the Archive. They are busy machines, and fragile, subject to any number of unpredictable factors that can cause them to fail. Keeping the machines going is a 24/7 task, requiring devoted vigilance, clever trouble-shooting, and intimate familiarity with the machines in the cluster and their sensitivity to a variety of predictable and unpredictable circumstances. And, perhaps because of the constant challenge of machine maintenance or perhaps because of an affective desire to be closer to the machines, the Archive has slowly been moving away from having its data cluster kept in locations not under its control—an Internet consortium in Redwood City, rented space in an office building, or the campus of a technology company—to

housing the machines on its own property, including the Archive headquarters itself, where some of the machines have been arranged such that the heat they generate is repurposed to heat the Archive's offices.

b. The Digitized Book as Logical Object

According to Thibodeau's schema, a logical object is an object that software recognizes and processes. As such, the digitized book is an organized group of content files and their corresponding metadata—i.e., data that describe other data⁹—combined into an object or “item,” with a unique identifier. Let's take an example. Nietzsche's *Genealogy of Morals*—or, to be precise, the 1923 Modern Library edition of Nietzsche's *Genealogy of Morals*, translated by Horace Samuels and held in the University of Toronto Library—is known to the Archive's cluster as "genealogyofmoral00nietuoft" and is stored on machine ia600404 with a mirror (a copy) on machine ia700404. Figure 4 is a representation of the digital object “genealogyofmoral00nietuoft,” a grouping of content files and metadata, in the Archive's storage system. These eighteen files

Index of /3/items/genealogyofmoral00nietuoft/

../			
genealogyofmoral00nietuoft.djvu	22-Dec-2010	23:16	4356523
genealogyofmoral00nietuoft.epub	22-Dec-2010	23:17	463026
genealogyofmoral00nietuoft.gif	21-Dec-2007	04:31	263818
genealogyofmoral00nietuoft.pdf	16-Dec-2011	21:26	8233006
genealogyofmoral00nietuoft.abbyy.gz	21-Dec-2007	05:41	5356289
genealogyofmoral00nietuoft.bw.pdf	16-Dec-2011	22:24	7054299
genealogyofmoral00nietuoft.dc.xml	20-Dec-2007	15:57	633
genealogyofmoral00nietuoft.djvu.txt	22-Dec-2010	23:17	360244
genealogyofmoral00nietuoft.djvu.xml	21-Dec-2007	05:44	3004785
genealogyofmoral00nietuoft.files.xml	16-Dec-2011	22:24	5956
genealogyofmoral00nietuoft.jp2.zip	21-Dec-2007	04:20	65093255
genealogyofmoral00nietuoft.marc.xml	20-Dec-2007	15:57	2772
genealogyofmoral00nietuoft.meta.mrc	20-Dec-2007	15:57	951
genealogyofmoral00nietuoft.meta.xml	16-Dec-2011	21:07	2318
genealogyofmoral00nietuoft.metasource.xml	20-Dec-2007	15:57	412
genealogyofmoral00nietuoft.raw.jpg.tar	20-Dec-2007	19:14	1165783040
genealogyofmoral00nietuoft.scandata.xml	20-Dec-2007	19:12	137864
genealogyofmoral00nietuoft.scanfactors.xml	17-Nov-2008	07:02	2496

Figure 4. The list of files that makes up an Internet Archive digitized book as digital object. All are viewable or downloadable at <http://ia600404.us.archive.org/3/items/genealogyofmoral00nietuoft/>

are each logical objects—units recognized by some application software—that make up the digitized book “genealogyofmoral00nietuoft”: itself a larger,

⁹ Metadata is “structured, encoded data that describe characteristic of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities.” ALA, Task for on Metadata, Summary Report, June 1999. <http://www.libraries.psu.edu/tas/jca/ccda/tf-meta3.html>.

composite logical object. Roughly half are content files (containing images and/or text) and half are metadata files (containing coded information describing the book or other metadata files).

To understand the file structure of “genealogyofmoral00nietuoft,” we need to back up and review the process of digitization, namely, the process that creates the digital files to begin with. Digitization begins with image capture (or “scanning”). The Archive digitizes books using “non-destructive” scanning, which means that it does not unbind or “guillotine” the book and then feed the pages through a sheet fed scanner, but rather it keeps the book whole, and humans (“scanners”) turning the pages one by one as they are photographed or “captured.” The Archive uses a custom-designed image capture machine called a Scribe (figure 5), designed by Tom McCarty in 2004. Kahle found McCarty through Jim Mason, a Burning man veteran and founder of the Shipyard, a storied collaborative industrial arts space in Berkeley, California, built around recycled shipping containers.¹⁰



Figure 5: The Internet Archive’s Scribe image capture machine. Photograph courtesy of Internet Archive.

Before developing his specifications for the Scribe, Kahle had experimented with a Kirtas robotic scanner in the first Archive scanning center at the University of Toronto, but it proved too expensive and impractical. Besides costing \$130,000, the Kirtas machine sometimes grabbed two pages at a time or tore pages such that it ended up requiring human supervision, which defeated the purpose of the labor-saving robot. Based on the experience he had acquired both in Toronto

¹⁰ McCarty would also later design the Internet Archive’s long-term storage repository as well. <http://blog.archive.org/2011/06/06/why-preserve-books-the-new-physical-archive-of-the-internet-archive/>

and, previously, working with the Million Books Project, Kahle knew what he wanted. He instructed McCarty to develop a manual page-flipping system that would shoot the books through glass (to flatten the book's pages) at high speed for less than \$.10 cents per page. In 2004, McCarty and several subcontractors designed the mechanical side; Mark Johnson, a friend of McCarty, designed and wrote the software. McCarty would make a prototype, take it to the Archive, let them work on it, and then exchange it for a new one, which they would continue to test while refurbishing the first—until they got it right. After the second prototype was completed, in December 2004, Google announced its scanning project, which immediately introduced a competitor to its efforts. This news changed Kahle's and McCarty's plans a bit because Google was on record as offering free scanning to libraries. Obviously, the Archive wasn't going to be able to compete on price, so they decided they would have to compete *on quality*, which meant achieving archival quality images. To that end, they replaced lower quality Sony cameras with much more expensive Canon cameras, and settled on a goal of capturing images at 600 dpi. After three more prototypes, the Scribe went into production at a machine shop, with the first ten shipped in April 2005. They cost \$10,000 each, plus the cost of the cameras.

Though often called “scanners,” the machine in Figure 5 is, to McCarty, more properly termed an “imager.” A scanner moves over and across an image, line by line, like a photocopier, but an image capture machine snaps a photograph of each page. A “scanner” instead indicates the person who operates the imager, as shown in Figure 6.



Figure 6. A scanner operating the Archive's Scribe image capture machine. Photograph courtesy of the Internet Archive.

The scanner places a book on a cradle, a v-shaped surface that supports the book, as shown in figure 6, and uses a foot pedal to raise and lower a glass platen onto the book. This glass weight flattens the book's pages, which are curved when a book is open, so as to eliminate any distortion created by that curvature. Once the page is flat, two cameras above take photographs of the

book. The camera on the right photographs the verso page, and the camera on the left the recto.

Google is very secretive about its scanning operation, but some details of it have either been shared or can be inferred. For one, they do not use any sort of platen to flatten the book's pages. Instead of the platen, they have patented an infrared dewarping system that detects page curvature and corrects for it, illustrated in Figure 7 below.¹¹

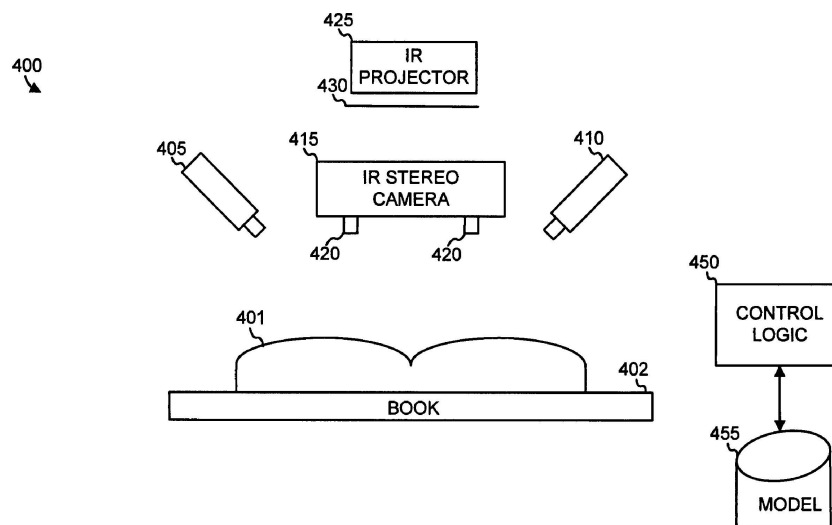


FIG. 4

U.S. Patent

Dec. 9, 2008

Sheet 4 of 13

US 7,463,772 B1

Figure 7. Drawing in Google Patent depicting their infrared page de-warping system.

Whether or not this exact system is actually in use at Google, the elimination of the platen and the foot pedal would significantly speed up the image capture process because the scanner need not stop after each page turn to raise and lower the glass. She need only turn the pages. One downside to this quicker method is that occasionally fingers are caught in Google's book scans. This sudden eruption of a hand in a page image book has earned the label "finger

¹¹ See Patent No. 7463772. "De-warping of scanned images."
<http://www.google.com/patents/US7463772>

spam.”¹² One rarely finds this particular problem in the books digitized by the Archive, because the scanner has to remove her hand in order to lower the glass platen. What is lost in speed, in this instance, is gained in quality. Google, however, anticipates and mitigates this problem by scanning each book twice.¹³

But Archive scanners do more than turn the pages one by one. They make judgments about camera settings depending on the size of a book, and they perform QA (quality assurance) during scanning. Because every book is a different size and different thickness, the Archive scanner, before he begins, needs to adjust the cameras (in one of three positions, 400, 500, and 600 dpi) for the best image capture. He also has to maneuver the book on the cradle so as to get a proper photograph of each page. With books that are tightly bound or books that have very small “gutters” (the white space formed by the inner margins of two facing pages), it can be tricky to get the book to lie well in the cradle. This maneuvering requires jerry-rigging the book with whatever happens to be on hand, such as pencils or folded paper wedges, to get the proper fit (note the pencils in figure 10 below). Similarly, scanners also perform QA (quality assurance) either during or after the page turning is done, depending on which generation of Scribe software they are using. When I scanned books in the Archive’s San Francisco scanning center, I struggled with one particular book, an illustrated book of poems entitled *Black Gibraltar*. The drawings in the book bled off the page, which meant that I had to turn off the automatic page cropping in order to capture the outermost part of the page’s edges. Usually, these would be cropped off but, in this case, such cropping would lose desirable “content.” Accommodating the images required multiple attempts to capture the page correctly and yet, at the end, during the QA review, I was dismayed to see that I had missed a number of pages in what was a fairly short book. With another book, Henry Fielding’s *Tom Jones*, I struggled with its thickness, its tiny margins, and very tight binding, jamming pencils under here and there, worrying each time that I would not capture all of the text, cutting off even a sliver of the text, thus ruining the scan.¹⁴ It was a long book and this anxiety accompanied my every page turn. These brief descriptions of my own scanning are to suggest that capturing a book image well is not simply a matter of turning pages because

¹² Matthew Moore, “Google Book ‘Finger Condoms’ Cause Mirth,” *Telegraph* October 21, 2009. <http://www.telegraph.co.uk/technology/google/6396896/Google-Book-finger-condoms-cause-mirth.html> in “Are Google Books On Demand Books Ready?” <http://fonerbooks.blogspot.com/09/are-google-books-on-demand-books-ready.html>

¹³ I learned this from Brandon Badger, Product Manager for Google Books, during an online seminar on the Google Book Search Settlement in January 2010.

¹⁴ Indeed, as neither of these books has appeared on the Archive website, I assume that I failed to scan them acceptably.

books themselves vary greatly and often violate the standards assumed in industrialized workflows.

The labor of scanning is curious. Supervisors at the Archive's oldest and largest scanning centers told me that when they started they thought they would need to organize the scanners into short shifts because no one would be able to tolerate sitting at a Scribe turning pages for a full-time shift. To their surprise, people did not balk at longer shifts and, indeed, the staff was fairly stable. Those who find the work amenable, I was told, like "working with books." Published requirements for the position include "strong technical skills," "strong verbal and written communication skills," an ability to work independently, to lift heavy books, to gently handle special library materials, to sit for 6 to 7.5 hour periods, and to be "comfortable" with repetitive motion.¹⁵ Combining skilled and "unskilled" labor, scanners are asked both to manage software systems, make judgments about image capture, but also perform repetitive mechanical labor. Scanners, or "digitization specialists" are relatively low paid, averaging \$17/hour (including benefits).

As a point of comparison, Google's scanning depends on automated processing for its digitization workflow, leaving the humans essentially to turn pages, the quickest and most efficient way to digitize a book when it is not feasible to cut off its spine and feed it through a scanner ("destructive scanning"). Little is known about Google's army of scanners, and Google employees are constrained by non-disclosure agreements they sign as a condition of employment. Recently, however, a former contractor at Google has provided some details about the book scanners at Google's Mountain View campus (Wilson 2011). In 2007, Andrew Norman Wilson worked for a video company called Transvideo, which contracted with Google to produce videos on its Mountain View campus. While working there, Nelson became interested in the workers digitizing books at Google, who were in the building next door, Building 3.14159~ (which, by the way, happens to the number π , be a geek joke). He described a color-coded caste system among employees: Google employees ("Googlers") had white badges, contractors red badges, interns green badges, and, the book scanners yellow badges. The scanners are their own category of worker—known locally as ScanOps. They have a different shift, arriving at 4 a.m. and leaving around 2:15 p.m. They are not given the same privileges as other workers, including contractors like himself, such as riding the Google bikes, taking the free limo shuttles home, eating free meals in the cafeteria, and so forth. Wilson filmed the book scanners leaving the building as their shift ended, noting that most were people of color. He attempted to talk to a few of them, but was soon confronted by a Google security employee who told him he was in an "extremely

¹⁵ An example job ad:

<http://www.archive.org/about/archivejobs.php#Digitization%20Specialist%20at%20Sacramento%20California%20State%20Library>

confidential area” where “extremely confidential work” was being done. He was soon thereafter dismissed.

None of these details is particularly surprising—page-turning is low paying, low status work—but it does confirm, as many have recounted to me, how jealous Google guards its book scanning. It is also ironic that what most of what is known about those working in Google’ book scanning centers comes from the scanners inadvertently photographing themselves at work, as seen in the rather arresting images in Figure 8 below.



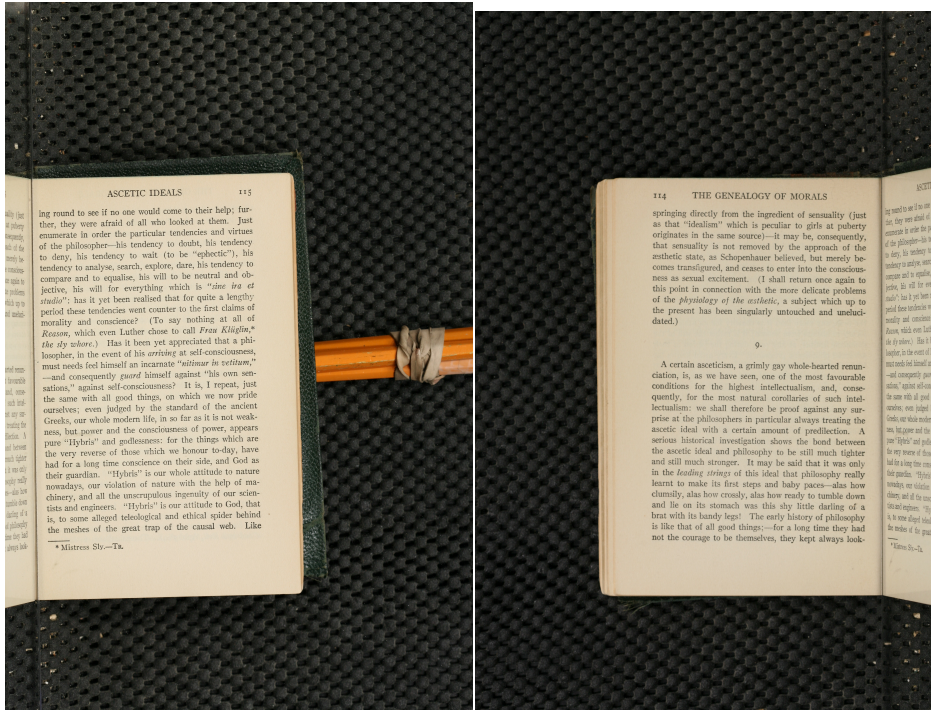
Figure 8: An inadvertent image of a Google worker's hand. "The Inland Printer – 164." Courtesy of Andrew Norman Wilson.

After the book has been captured, the scanner's job is done and the images are sent from the Archive scanning center—of which there are about thirty in six countries—to the Archive's cluster for "processing," ultimately resulting in the digital object represented above in Figure 4. That file list included two types of files: content or data files and metadata files. I want to briefly specify what these files represent. First, the data or content files.

During image capture, the Canon cameras on the Scribe have already engaged in a small bit of processing in the capture. They are capable of outputting their images in just two forms: either "raw" camera images or JPEGs, a "lossy" compressed format. To output in any other format, a camera would have to be custom engineered, which is not possible with the Canon cameras the Archive uses. (Google uses open hardware and open source Elphel cameras, which they custom engineer.¹⁶) At an early stage in the book scanning operation, the Archive used the raw camera image output option. These files are extremely large—up to 25 megabytes (MB) per page image—and hard for anyone to use due to the size, and very expensive to store. The decision then was made to use the other option: image output in the compressed JPEG format, which reduces the size to 4-5MB.

Figures 9 and 10 below show the book pages as they come out of the Canon camera, before they are further processed.

¹⁶ See: <http://video.google.com/videoplay?docid=-3616515426451811910>



Figures 9 and 10. Stored, “raw” images of pp. 114 (top) and 115 (bottom) of *Genealogy of Morals*. Figure 12 below will show these images after rotating, de-skewing, cropping, and reassembly into a facsimile of the book. I have rotated them 90 degrees here.

The page images are sequentially numbered and stored individually (Figure 4: raw_jpg.tar), and from there processing (or what the Archive calls “deriving”) begins. Deriving is a twelve-hour process when “digitization” is actualized. The photographing of the book, or its “scanning,” has prepared the “input” from which these successive stages of machine processing can be accomplished. It begins with image processing. This camera output is first converted into JPEG2000s (“JP2s”) image files, which is the format used for storage (Figure 4: jp2.zip). The JP2 format is a “lossy” image compression format that allows a great deal of end user functionality (like zooming), while at the same time achieving small, manageable file sizes and tolerable loss despite a great deal of compression. JP2 has become accepted as archival quality by libraries. These compressed JP2 images are the most crucial content files of the digitized book, its very backbone. It is from them the other content files derive, in a chain reaction.

Optical character recognition software (OCR), which is one of the processes performed on the JP2 images, is the most “computationally expensive” part of the process, producing encoded text (Figure 4: abby.gz.). Because the OCR text, as will be further examined in chapter 3 (“Books as Data”) is a crucially important element of the digitized book, I want to spend some time on it here. OCR software is a program that “looks” at a picture of a word and converts it into corresponding

digital text. The output of this software is also called “OCR,” as shorthand. OCR pre-dates the computer. It emerged out of electronic data processing techniques in the first few decades of the twentieth century when early patented examples include the conversion of printed text into Morse code and, in the case of a device called the Optophone, the conversion of text into tone “characters” that would allow a blind person to comprehend a text through sounds (see Schantz 1982). But OCR gained commercial viability only after the computer, in the 1950s, as it enabled corporations and government agencies to automate the handling of information, whether Readers Digest processing magazine subscriptions or the US Post Office sorting the mail. OCR then, as now, automated data entry and thereby functioned as a gateway practice of computerization, moving information from paper to machine.

Much of the twelve-hour processing of a scanning book is taken up by the OCR process. There are various types of OCR software (both proprietary and open source), but the “best in class” is the commercial program Abbyy Fine Reader, which both the Archive and Google use.¹⁷ It claims to be able to recognize 189 languages, which makes it suitable for mass digitization.

The OCR process begins by analyzing the structure of the book page. After identifying the blocks of text, the software separates the “blob” into lines, those lines into words, and then into characters (Smith 2007). Once the characters have been singled out, the OCR program compares each individual character against a set of pattern images. It considers numerous hypotheses about what character the image represents and then, after processing a large number of probabilistic hypotheses, the program finally “decides” on which characters it recognizes. In illustration, I include an example of OCR. Below is word as photographed (from the same book I took as my sample in chapter 2):

bribery

Now, below, you’ll find the encoded output from the Abbyy OCR software. This output records the machine’s decision as to what the image of the word represents.

```
<charParams l="470" t="1554" r="496" b="1594" wordStart="true"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="100" serifProbability="100"
wordPenalty="0" meanStrokeWidth="40">b</charParams>
```

¹⁷ Google supplements Abbyy with its own custom programming (conversation with Tom Breuel) and seeks to move away from Abbyy. The company has devoted itself to extensive OCR research, taking over development of the open source Tesseract OCR software (Smith 2007).

```
<charParams l="500" t="1568" r="517" b="1593" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="54" serifProbability="100" wordPenalty="0"
meanStrokeWidth="40">r</charParams>
```

```
<charParams l="520" t="1555" r="531" b="1593" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="43" serifProbability="255" wordPenalty="0"
meanStrokeWidth="40">i</charParams>
```

```
<charParams l="533" t="1555" r="559" b="1594" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="100" serifProbability="100"
wordPenalty="0" meanStrokeWidth="40">b</charParams>
```

```
<charParams l="563" t="1568" r="583" b="1594" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="100" serifProbability="83" wordPenalty="0"
meanStrokeWidth="40">e</charParams>
```

```
<charParams l="586" t="1568" r="604" b="1593" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="42" serifProbability="100" wordPenalty="0"
meanStrokeWidth="40">r</charParams>
```

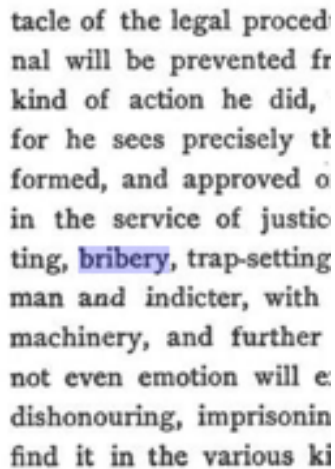
```
<charParams l="606" t="1569" r="633" b="1605" wordStart="false"
wordFromDictionary="true" wordNormal="true" wordNumeric="false"
wordIdentifier="false" charConfidence="100" serifProbability="49" wordPenalty="0"
meanStrokeWidth="40">y</charParams>
```

Since this output was not intended for human reading, I have added boldface to each result, separating it out from the coded information, to assist the human eye, and I have made paragraphs for each block of code. One can make out the letters b-r-i-b-e-r-y amid a series of position coordinates and probability measures that have assisted the algorithms pouring over the text. We can conclude here that the result is correct.

Next, this “character-level” OCR output is processed once again to become “word-level” OCR, which looks like this:

```
<WORD coords="470,1605,644,1554">bribery,</WORD>
```

This word-level XML-encoded OCR includes coordinates explaining where on the page image the word is located. This encoding of words with “bounding boxes” is what enables later highlighted searching within the image of a book page (whether in a pdf format or in the browser-based Book Reader), as seen here:



tacle of the legal proced
nal will be prevented fr
kind of action he did,
for he sees precisely th
formed, and approved o
in the service of justic
ting, bribery, trap-setting
man and indicter, with
machinery, and further
not even emotion will e
dishonouring, imprisonin
find it in the various ki

Figure 11: Page image highlighting the search term—“bribery”—located through a search carried out on encoded OCR text.

OCR enables computers to read in the limited sense of enabling them to “see” and “recognize” images of characters (numbers, letters, ideograms, etc.) and then to create digital equivalents of those images. Without the images, we have no OCR, but without the OCR we no “computerized” text, and thus little functionality and little “digital” advantage. OCR’s primary purpose is to enable one function: searching through a book’s text. When one performs a search query on a digitized book, the machine searches through not the photographs of the book’s pages but the OCR text, its unseen dimension.¹⁸ OCR is the spirit or animator of the digitized book.

As the JP2 images form the basis of that which makes the digitized book readable by humans, the OCR output forms the basis of that which makes the book’s *machine* readability: the topic of chapter 3. The remaining content files (cf. Figure 4) are additional, derivative file formats that provide different methods of viewing the book: pdf (both color and black and white, and both searchable); djvu, an “open,” searchable, and more compressed alternative to pdf; and epub, an open textual format that has become the industry standard for re-flowable,

¹⁸ The Archive provides this data for download in three forms. Two we’ve seen above (the character and word level XML encoded OCR results). The third is the “plain text,” or the ASCII text (words without the XML coding). See: <http://blog.openlibrary.org/2008/11/24/bulk-access-to-ocr-for-1-million-books/>. Commercial firms rarely provide access to the OCR text.

non-image based e-books (for use in such e-readers such the Nook, the Sony Reader, I-Pad, I-Phone, and so forth).¹⁹

The other type of files in Figure 4 are metadata files, and they are of two types—descriptive (data about content) and structural (data about the logical object itself). Metadata as a term was coined in the late 1960s in the context of information retrieval and information science, but as a concept its use extends back to library management practices of the latter nineteenth century, whether classification systems like the Dewey Decimal System or bibliographical cataloging practices like the card catalog. Defined by the OED as “a set of data that describes and gives information about other data,” metadata would also include the library card catalog and many other means of recording and storing information about books before the computer—which could also include records that publishers or booksellers keep and, in some cases, share. The latter half of the twentieth century saw a dramatic increase and complexification in available book metadata, as interlocking computerized systems of bibliographic and inventory control were implemented to standardize and coordinate, both nationally and internationally, the output from the exploding book publishing industry). A good example of this is the various numerical encodings, such as the ISBN and the EAN barcodes, which were developed from the 1970s onward (Striphas 2009, ch. 3).

Digitizers face the task of absorbing the former metadata practices of the traditional book apparatus in addition to their Web-oriented metadata requirements, and it is perhaps no surprise that the complex and often maddening task of managing book metadata has become the Achilles heel of mass digitization projects. Google, in particular, has been taken to task for its messy commingling of metadata from 48 libraries and 21 commercial metadata providers,²⁰ leading to widespread errors in publication dates, classification, and misidentified books that have elicited truculent reactions from scholars, who judge Google Book Search a severely compromised research tool. The Archive struggles with book metadata, too, one extended example of which I will treat at the end of this chapter.

Objects are constructed through metadata. All metadata is local in the sense that “it is not the object itself that determines the metadata but the needs and purposes of the people who create it and those who it will serve” (Coyle 2005). The Archive’s metadata thus describes the content of

¹⁹ One remaining content file (.gif) is an image drawn from the book for display of the book on the Archive’s website.

²⁰ Details about Google’s metadata are from the Declaration of Daniel Clancy in Support of Motion for Final Approval of Amended Settlement Agreement at 1, *Authors Guild, Inc. v. Google Inc.*, No. 05-CV-8136-DC (S.D.N.Y. Feb. 11, 2010).

“genealogyofmoral00nietuoft” in such a way that grows from its purposes as a digitizer and a library. Its “descriptive” metadata provides, unsurprisingly, basic cataloging information in the MARC records (Figure 4: meta.mrc; marc.xml). MARC stands for Machine Readable Cataloging Records, and a standard developed by the Library of Congress in the 1960s that is now used for bibliographic recordkeeping in much of the world. The MARC record provides the mechanism by which *computers* exchange, use, and interpret bibliographic information. Its elements make up the foundation of most library catalogs used today. The MARC record for “genealogyofmoral00nietuoft” provides not only basic bibliographic information (author, title, place of publication, and other information one is used to see in a library’s catalog), but also “authority” records although the MARC record itself: where it originated (this one from the University of Washington library) and what other authorized organizations have since modified it. (It seems worth noting here that the MARC record is not “human readable,” I was able to decode it after a great deal of back and forth with the Library of Congress MARC website (<http://www.loc.gov/marc/>)). The remaining descriptive metadata files combine cataloging information with various specifics about the Archive’s scanning (who, where, when), copyright status, the sponsor of the scanning, and so forth—all the information expressed on the item’s page on www.archive.org webpage (Figure 4: meta.xml), as well as the source of the book’s metadata (metasource.xml). Indeed, some metadata is data about metadata.

To conclude this description of the logical object, “structural” metadata includes information about data or content structure, such as the XML markup above that indicates position page, column, and paragraph structure for every word. Yet another file contains a list of the specifications for the cropping, rotation, and skew angles that were performed on each raw page image (Figure 4: scandata.xml). Finally, there is one file with data about all eighteen files that make up “genealogyofmoral00nietuoft”—the logical object that figure 4 (above) represents.

The digitized book is thus the combination of the physical object (bits stored in a machine) and the logical object (the files that the machine acts upon). A text does not reside in the machine as such. Rather, it is the result of actions taken upon the digital object, which result in an interface between the human and the machine or what I am calling here the conceptual object.

c. The Digitized Book as Conceptual Object

When one views “genealogyofmoral00nietuoft” in the Archive’s browser-based Book Reader application—aptly called “Book Reader”—one finally encounters a visual representation of the conceptual object—a book:

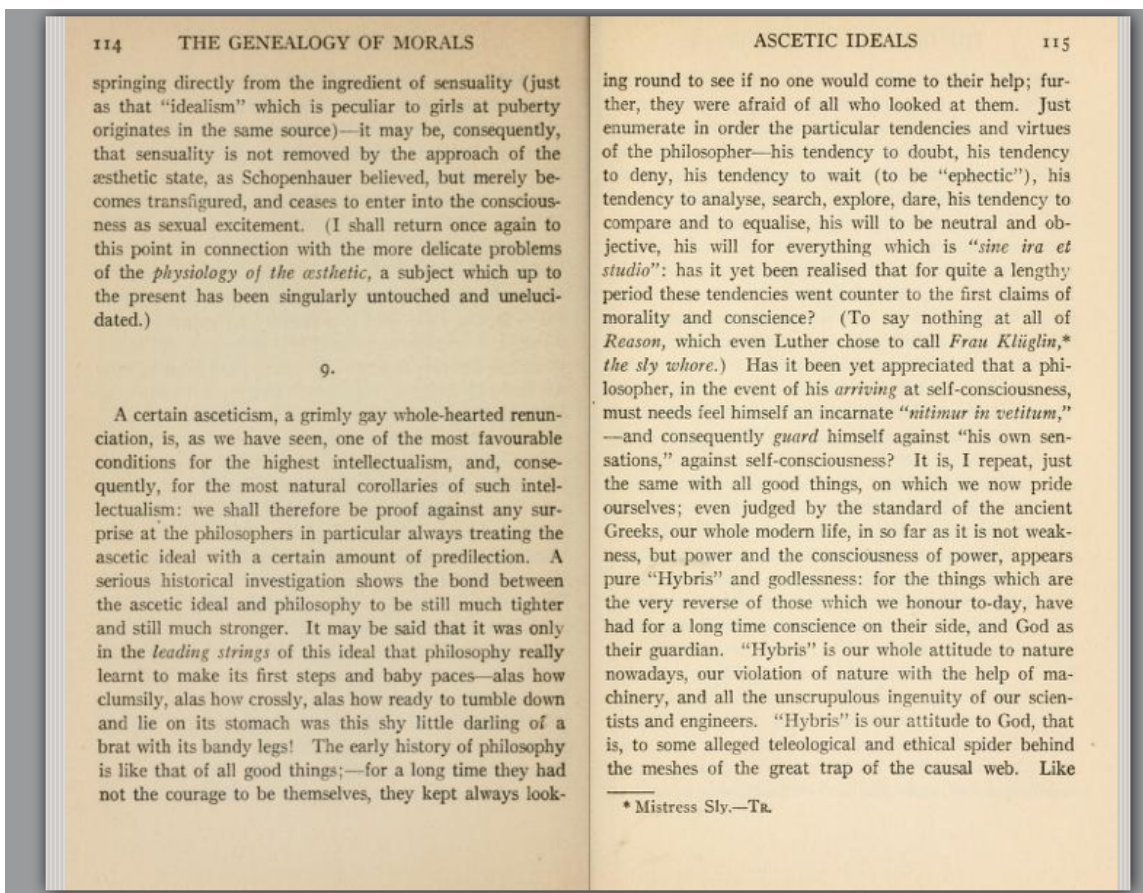


Figure 12. A view of pages 114-15 of *genealogyofmoral00nietuoft* in the Archive's Book Reader, derived from Figures 10 and 11 (above). This page can be viewed "live" here: <http://www.archive.org/stream/genealogyofmoral00nietuoft#page/114/mode/2up>

How does the Archive understand this conceptual object? What choices did they make in their design of the digitized book as digital object, and what difference have these choices made or might they make? Here I'd like to compare the IA and Google's digitized facsimiles in order to draw out these choices and to uncover the thought and judgment that shape them, whether regarding presentation, quality, preservation, accessibility, or openness.

The Archive staff is very proud of their scans: their "superior" quality is a point of pride. One of the things I was told on my first day of research at the Archive was that their scans are superior to Google's. The Archive's scans resemble "real books" but Google's look "like Xeroxes." Indeed, the quality of a book's scan trumps other considerations. One of my first tasks when I began my work at the Archive was to assemble two collections from among their growing collection of public domain books: one a collection of "banned books" on the occasion of Banned Book week; and the other a collection of "great books" based on a list assembled by St. Johns College. I found both tasks vexing: the former because

most banned books are newer and not yet in the public domain, so I could add only a few books to the collection; the latter because there were often multiple editions of these canonical texts and I didn't know how to decide which of the many versions to include in the collection. Also, most of the "great books" were translations but how could I, without close knowledge of the textual histories of these works, identify the superior, preferable editions? I also had the sneaking suspicion that translations published before 1923—that is, in the public domain—would be antiquated and not particularly desirable.

But soon I would realize that I was thinking like a scholar and not a mass digitizer. When I expressed my hesitation about choosing one book over another to my appointed supervisor, I was told to choose on the basis of which book had the best scan. My anxious concerns about scholarly value or the quality of the translation turned out to have been misguided. What mattered was the quality of *the scan itself*. This realization brought to mind something another Archive employee had told me earlier that first week: that the Archive is "agnostic about content." Questions of "content" are up to the library partners. The quality of the scan, the storage, the accessibility—that was the Archive's concern. A few months later, this perspective on priorities was driven home in yet a different way. Some scholars from a local university, human rights center, came had to the Archive in hopes of having their library digitized and put online as part of their project, in a word, to save the world from self-destruction. As they were describing the importance of their aspirations and how digitization was part of it, Brewster interrupted their pitch, half impatiently and half to save them the trouble, and said: "That's all fine and good, but we don't know anything about anything here." In other words, the Archive scans books but not *particular* books.

After understanding the criterion for choosing my "great books" was the quality of the scan, I was faced with another difficulty. What was a good scan? All of the scanned books were of a certain quality already: there were no blurred pages, no photographed hands or "finger spam." But, beyond that, what else mattered? Did a good scan mean there was no handwriting or underlining on the page? Did it mean it was "more readable" in the sense of having more white space, bigger margins—that is, more pleasant to my quite possibly idiosyncratic eye? After finding five possible versions of Shakespeare's *Sonnets*, I could not choose between them. The edition with the larger print attracts my eye, but it also seems *too* yellow. Another is a more pleasant color, a paler yellow, but the page seems too jammed full of type and the close cropping makes it seem all the more jammed. The other editions are grayish in color, with dark gutters. Unable to decide, I appeal to yet another employee who looks them all over and makes a decision easily. "That one, of course," pointing to the one I thought was too pale. Why that one?, I ask. "Because it has an even appearance from left to right, across the two-page spread." I was learning to see a digitized book as a digitizer does.

Both Google and the Archive capture their books in color, but Google represents its books (in the Google Book Search interface) as “bitonal” images as seen in Figure 13.

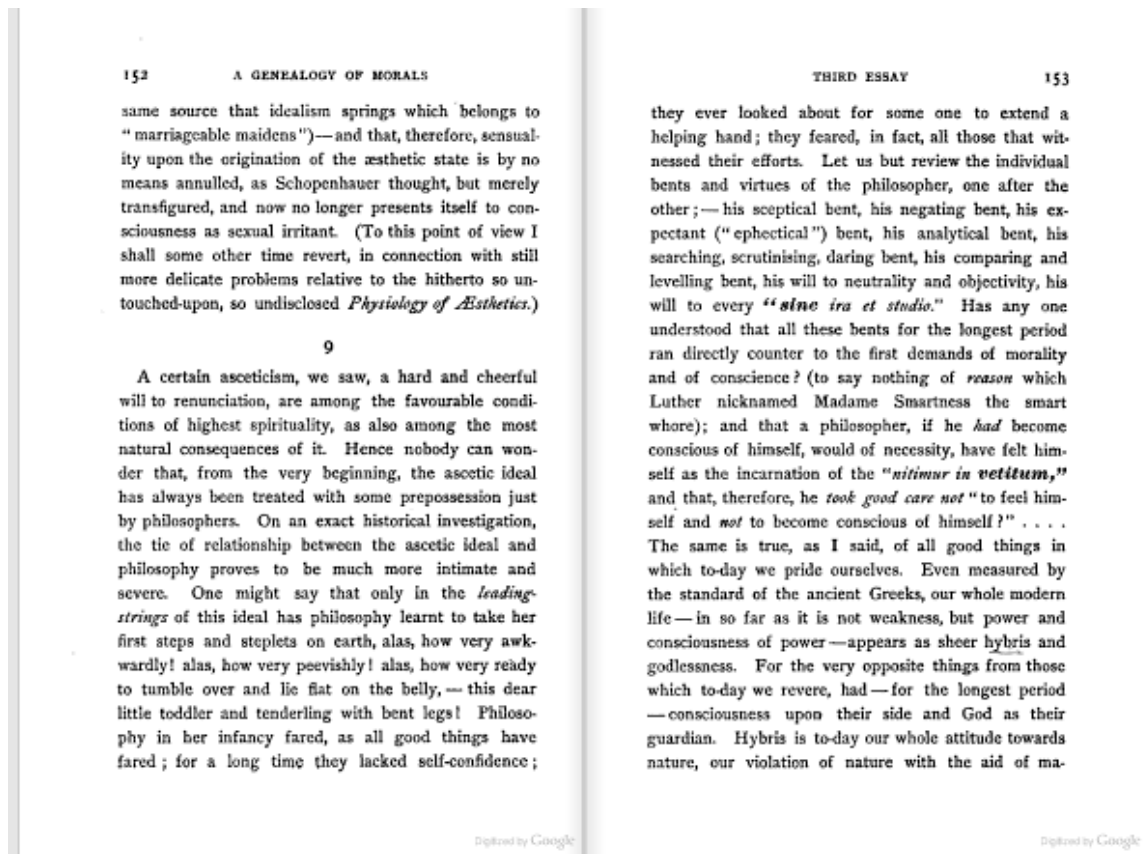


Figure 13: Roughly the same two pages from Google’s *Genealogy of Morals* as in figure 12 (above), though this is a different English translation. This can be viewed “live” here: <http://books.google.com/books?id=n4INAQAAIAAJ&dq=genealogy%20of%20morals&pg=PA152#v=twopage&q&f=false>

This digital facsimile was made from a book in Stanford University’s library, published in 1897. Over 100 years old, it, its “true” state would likely appear yellowed and otherwise aged (as those from “genealogyofmoral00nietuoft” do, in Figure 12 above). But Google has chosen to process the color images in order to create a black-and-white “bitonalized” image for viewing. The contrast of black on white is sharper than a continuous tone image with a yellow background. Such “bitonalization” can also decrease the size of the file, allowing for a faster response time serving the book, which leads, presumably, to greater user satisfaction. Overall, Google’s emphasis is on the *book as text*.

The Internet Archive, on the other hand, displays its book scans in color in order to present a more accurate representation of the actual book that was scanned. One books engineer told me that the Archive's digitization was "about preservation not re-creation." The effect is that the books at first glance appear old because they are often yellowed. The idea is to "preserve" *the book*, to capture it so well that it could, when printed, serve as a facsimile of that very book as found on the library shelf. For the same reason, the Archive leaves visible any marginalia left by previous library patrons, even the garish highlighting from fluorescent markers. Whether annoying or not, this is how we experience library books, warts and all, and the Archive seeks to replicate the experience, to the extent it can. In contrast to Google's emphasis on the book as text, the Archive's is on the *book as artifact*.

Further demonstration of this is the treatment of "foldouts," a special problem that the two digitizers have handled differently. A foldout is an illustration tipped into a book that requires unfolding to view. They can be maps or artwork, and they are often in color. Google does not scan them because foldouts require a different capture process, with different equipment. To photograph foldouts significantly slows down the scanning. Presumably, Google considered it too costly or inefficient to accommodate such exceptions into their workflow. In order to "achieve scale," mass digitization necessarily entails tradeoffs—as we saw above with the design of the image capture machine. In this case, Google chooses to leave out this sort of illustration but they at least lose no text. The handling of foldouts further demonstrates that for Google the capture of text is more important than the capture of the artifact as a whole.

This choice to present books as artifacts rather than as texts stems from the Archive's relationship to its library partners. It primarily digitizes inside libraries. Most of its thirty scanning centers are small—a couple Scribes in a corner of a partner library. It digitizes books *for libraries*. Its partner libraries pay for the scanning, they choose which books to scan, and they are the "customer." In relationship to Google, libraries might also be seen as customers—libraries loan their books to Google in return for a copy of the scanned book—but they are not clients quite as they are for the Archive. The Archive digitizes, hosts, and serves books, but it sees itself as a "backend" service operation, one node in a distributed network. Although the Archive in some limited sense serves "end users" or the "general public," its book scanning was designed as more of a "back end" operation than a public-facing service like Google Book Search.²¹ Google, on the other hand, is deeply committed to the "end user," an imaginary

²¹ This changed somewhat when the Archive's project, Open Library, was redesigned and relaunched in 2010, and when the Archive started its lending and accessible book programs that same year through Open Library, but the scanning system, and all the choices that went into it, was in place well before those efforts.

universal subject who searches for information that might just happen to be in a book. When scholars pen their severe critiques of Google Book Search, they mistake Google as having the intention of trying to please them *as scholars*. At one gathering I attended with regard to the Google Book Search Settlement, a Google Book Search executive memorably remarked, “How many editions of Hamlet does anybody really need?” To a scholar, this remark exhibits an insulting ignorance; to a non-scholar looking for Hamlet’s soliloquy, it might seem perfectly reasonable.

Making a book available in multiple formats is another one of the choices the Archive has made for its books. Figure 12 (above) shows one such format: a two-page spread of *Genealogy of Morals* as it appears in the Archive’s Book Reader, an open-source, browser-based application that draws upon some of the logical objects, or files, listed above, to allow a reader to flip through the book quickly, to approximate skimming. Most commercial e-book sites (iBooks, the Kindle, Google Play) now have an equivalent “book reader” application, as do libraries with digitized book collections (Hathi Trust, Library of Congress, etc.). The Archive’s Book Reader is only for reading image-based digitized books, not e-books. For that, you can send the Archive’s OCR text to a Kindle or you can download that same text in open EPUB format to use with other reading devices such as the Nook, tablet computer, or an Android phone. Those who are blind or visually impaired (and registered as such with the National Library Service) can download a special talking-book format (DAISY), the international standard now for talking books for the visually impaired. The Archive’s books (in English) can also be read aloud with a synthesized machine voice. Books can also be read offline. This same copy of *Genealogy of Morals* can be downloaded as a searchable pdf and printed, if so desired. These various formats are different digital encodings of the same conceptual object made available for a range of possible users. This desire to make its books available in as many formats as possible is fundamental to Kahle’s dogged vision of a distributed library system that doesn’t favor large centralized suppliers or “walled gardens” but open standards and decentralized structure. The Archive sees itself crucially as an free and “open” resource—like a public library—and so also encourages users to bulk download—that is, to take as much of their collections as people want.²²

Finally, this “openness” is another source of pride for the Archive, especially when it compares its book digitization program to Google’s. Although the bulk of Google’s digitized books are in copyright, around three of the twenty million are in the public domain. Until 2010, all of the Archive’s books were in the public domain. Google users to download the public domain books in (an

²² “Bulk Access to OCR for 1 Million Books,” Open Library Blog, November 24, 2008. <http://blog.openlibrary.org/2008/11/24/bulk-access-to-ocr-for-1-million-books/>.

unsearchable) pdf format for free, one at a time. Each time a user clicks on the “download pdf” button on Google Book Search, Google produces a pdf “on the fly,” and, within that pdf, an automatically generated page of “legal text” is inserted as page 1, in both the language of the book and in the language that the user’s browser is using (Langley and Bloomberg 2006). This legal text provides “usage guidelines” that request a user: (1) make only non-commercial use of the files; (2) refrain from automated querying (e.g., text mining)²³; (3) kept the digital watermark that Google puts on each item in tact so that they continue to receive “branding” for the work; and, finally, (4) be aware that although the book is in the public domain in the U.S., it may not be in other parts of the world. Google knew when they wrote this that it would be possible for a user to download all of the public domain books from their site, remove the digital watermarks, and rehost them elsewhere. It simply asks that people not do this. Google is not making claims to have any copyright or property ownership in these public domain works—they know that making copies of a book does not grant you any sort of “sweat of the brow” copyright. In other words, the author of this text knows that these usage guidelines are not legally enforceable (even if the Google engineers above call it a “legal text”). Google is appealing not to any law but to “netiquette,” the convention of “playing nice” on the Internet. The Internet Archive does not embed any sort of instructions or guidelines into its pdfs.

It is understandable that Google would want to protect its investment. As they explain it in their “legal text”: “Public domain books belong to the public and we are merely their custodians. Nevertheless, this work is expensive, so in order to keep providing this resource, we have taken steps to prevent abuse by commercial parties.” But, to public domain advocates such as Brewster Kahle and many Archive supporters, these arguments hold no water. Google had included similar restrictions into their contracts with library partners. To Kahle, these moves to restrict access to works in the public domain is an improper “enclosure,” a privatizing of the public domain, and an affront to the ethos of the Internet, which favors more not less accessibility.

One such public domain advocate is the anonymous “user tpb.” Tpb is someone Kahle has referred to as an “enthusiastic user” of the Internet Archive (email correspondence). Tpb wrote a script that could bulk download Google’s public domain pdfs. Rather than downloading one at a time, s/he was able to grab thousands at a time. Tpb subsequently uploaded them to the Internet Archive, noting there that they were being “uploaded by user tpb.” To date, 907,000 such books have been added to the Archive’s collection, which about half of the number of books the Archive has itself digitized. To much of the Archive staff,

²³ This is a curious request to make of a single downloaded PDF, against which one would not be inclined to automatically query. The legal text reads as if it is usage guidelines for all public domain books on the Google Book Search site in toto.

this “black op” was a coup: how dare Google put restrictions on the public domain!

The problem of how to integrate these pdfs into the Archive’s digitized book collection was a regular topic at the weekly Monday morning meetings of the Books Group. The books were in just one form—unsearchable pdfs—and the Archive had to “derive” them just as they would the page images captured by the Scribe. Processing tens of thousands of books was well above the normal rate and, as always, computationally expensive. To make matters worse, the PDFs lacked metadata, which would, among other things, indicate the language in which a book was written. Without language information, the Archive’s OCR engine is inoperable. The engineers, therefore, had to figure out a way to assemble metadata for all of these Google books. It was a big project, but just the sort of thing the Archive staff relishes: a “hack” in the service of a righteous cause. Archive engineers scraped as much metadata as they could from Google Book Search’s public site, but what they really needed were MARC records, which Google does not make public on their site due to licensing restrictions from OCLC, the licensor. The engineers used a variety of clever methods using available Web resources until they had incorporated most of the books into the Archive site. When the Archive speaks of how many books it has in its collection, few people realize, even though the Archive does little to hide the fact, that nearly a million of those Google has digitized.²⁴

Beyond the initial Herculean labors to reprocess them, numerous problems have dogged the Google pdfs in the Archive collection. First, Google was not happy that someone was violating their usage guidelines or that the Archive was rehosting their scans. The engineering head of Google Books (Dan Clancy) and the Google Books product counsel (Alex McGillivray) met with Kahle to discuss the matter at a San Francisco hotel lobby. A lawyer from the Electronic Frontier Foundation offered mediation, but nothing was accomplished. In fact, it seems to have only hardened the resentments between the two organizations. One engineer’s netiquette is another’s misdeed. All the while, user tpb was playing cat and mouse with Google until finally s/he was bested or lost interest and stopped uploading books. Second, in the bulk download/upload, the Google scans had degraded such that they had to be reprocessed yet again, which is always a big drain on the Archive cluster. And, third, many books regardless of the metadata “hack” ended up with the wrong metadata so that, for instance, when you call up Goethe’s Wilhelm Meister you might be surprised to see before you an American crime novel from 1912 entitled Peter Ruff and the Double-Four.²⁵

²⁴ The Archive’s “Google” books can be found here: <http://archive.org/details/googlebooks>.

²⁵ Some of these were mistakes imported from Google; some were introduced through the methods the Archive employed to harvest metadata from elsewhere.

This last point brings us back full circle to the issue of quality where we started. When I heard about the Google books being added to the Archive's collection at the beginning of my fieldwork, my first reaction was to wonder: if the Archive is so proud of its scans and holds Google's scans in disdain, why would it want to host so many of Google's inferior Xerox-like scans on its site? Wouldn't that taint the whole collection and perhaps diminish it? That proved to be another naïve reaction on my part. Indeed, not only was the little hacker mouse roaring against the corporate bad guy that would "lock up" the public domain, but it was also nearly doubling its book collection. I had learned another quick lesson: in a data-oriented organization, quality is never going to trump quantity. More is always better.

* * * * *

My contribution then to a broader conversation about materiality in the humanities, anthropology, and in media studies is to show not simply that digitized books possess a complex, unfamiliar materiality, such as I have described in depth in this chapter, but also to show also that these differences produce significant matters of concern. I do this in the next two chapters, where I focus on two flashpoints within the re-materialized digitized book: 1) books as data; and 2) books as orphans. Each corresponds to one of the two layers that I have located in the digitized book: the data layer (the logical object); and the image layer (the conceptual object known as a "book"). The data layer is oriented around machine processes that operate on the text data created through digitization; the image layer is oriented around the modern book apparatus. The image layer is more familiar and at least partially assimilable to the print book apparatus, whereas the data layer is unfamiliar and not easily assimilable, if at all. One is centered on the *human* reader, and the other around the *machine* as reader. But, it is important to point out that neither layer operates in isolation from, nor independently of, the other; rather, the two are mutually imbricated. I draw out these distinctions not to demarcate absolute boundaries but as a practical means for objectifying the digitized book and what in it I find to be of significance. That said, because of the need for organization, I separate them out into two chapters: one in chapter 3, and one in chapter 4.

Chapter 3

Books as Data

*“We have a moral imperative to figure out what we have. ...
We are building highways into a forest.”*

-- Jon Orwant, Google Engineer

We are said to be in the era of “big data” and children of the Petabyte Age (Lohr 2012; Boyd and Crawford 2011; Anderson 2008). A meme of increasing scope, Big Data is seen as a source of knowledge and value, however uncertain or latent. Collected both actively and passively from commercial transactions, genetic sequencing, security cameras, social networks, astronomical and other scientific observation, search engine queries, traffic cameras, mobile telephony, email, embedded sensors, and, indeed, “all the world’s books,” Big Data is the offspring of powerful and ubiquitous computers, high bandwidth, low-cost storage, and an apparent desire among organizations, corporations, and governments to collect more and more data to some extent simply because they can. The “big” of big data is not simply the size but the difficulty of it: big data exceeds the conventional capacities of data management and analysis. Big Data is information overload for machines, but in reaction the engineers, entrepreneurs, and researchers involved with it manifest not anxiety but excitement: it is hailed as a new paradigm for scientific discovery, a superior basis for decision-making, a new form of value, a new means of addressing human problems, and more (Hey et al. 2009). The search is on for new tools and “scopes” that can reveal patterns and relationships previously undetectable. Big Data is a man-made frontier that has taken on the character of “nature” (Fayyad 2001). It lures not only economic exploit but also knowledge making from within its excess, its messiness, its illegibility. Big Data is an epistemological gold rush for the computer era.

Beyond the pursuit of profit, actors within this assemblage around massive data are pursuing the means—what Paul Rabinow has termed “equipment” (Rabinow 2003)—by which data can be given force and effect. Equipment to make it legible and comprehensible to non-specialists; to make it “permanent” and archivable; to institutionalize it and give it authority; and to establish it in relation to it something akin to what Foucault called the “author-function” (Foucault 1969) so as to situate data within a discursive regime of truth making. Mass book digitizers participate in this assemblage as producers of “big data” and as proponents of its value.

The digitized book has become significant to a range of actors not only as an electronic version of a familiar textual form—the book—but also as an entirely new inscription, namely, as data. Distinct from both the original, printed book and the “conceptual object” that is the digitized book (chapter 2), the “data” of books is something that digitization produces and creates.¹ Digitization “datafies” books in a double conversion: books become data literally (as ones and zeros) and conceptually (as new sources of latent information and knowledge). Once rematerialized and reconceived as *data*, books are transported from the traditional book apparatus and cast into a new space of knowledge making, of science. Seen as a source or terrain of new knowledge, they are revalued, at least for now, as freshly potent. Book data is a surplus, a bonus, and an opportunity.

For my purposes, examining books as data reveals an unsettling of the relations that exist within a book and within the modern book apparatus. Seeing and acting upon books as *data* is one of the ways that digitization “opens” the book. It cultivates alternative perspectives on the book as form of property and, by inviting the machine into the book, it calls into question the dominion of the author (or copyright owner) as a sovereign figure over the book. In this sense, without forethought, placing importance on books as data has become one of the most effective strategies digitizers (and their supporters) have happened upon to legitimize the activity of digitizing copyright workers. This strategy, such as it is, depends on the mutual reinforcement of pretensions to science for the data generated from books and legal tactics to claim books data as belonging not to a book’s copyright owner but to the rest of us.

In order to explore and substantiate these claims, this chapter takes a complex form that might require some patience to work through. First, I evoke a specific ethos around data through the example of the Internet Archive and the specific career of its director, Brewster Kahle. This evocation will orient the reader—as my fieldwork oriented me—into a milieu in which data is an object of voracious pursuit. Second, I describe how researchers understand the potential for books as a particularly promising form of data. In particular, I look at reactions to and anticipations of the “research corpus” set out in the Google Book Search Settlement, which would have enabled public research access to Google’s digitized book corpus if the Settlement hadn’t been rejected by a Federal judge in March 2011. Third, I look at the actuality of books as data through the specifics of data researcher: “culturomics,” a method developed by two young Harvard bioinformatics researchers using books as data; and the “digital humanities” as practiced by a small group of researchers to whom Google gave research grants to work on their digitized collection. Here I show what exactly book data is and how people interact with it. From there I move

¹ In defense of my use of data as a singular noun, see Carter A. Daniel and Charles C. Smith, “An Argument for *Data* as a Collective Singular.” *The ABCA Bulletin*, September 1982, 31-33.

into the governance of data, and the relationship between data and copyright. Here I track the emergence of terms to describe or refer to the books as data, all of which are negations, indicating an outside to the modern book apparatus that digitizers are seeking to establish so as to legitimize and elevate their activities.

The Ethos of Data

In stark contrast to the librarians and scholars discussed in chapter 1 who, for centuries, have expressed great anxiety at the *multitudo librorum*—i.e., the staggering accumulation of books—Brewster Kahle, a self-described “digital librarian, cannot collect enough data. His Internet Archive is a Big Data organization. At the time of writing, it hosts in its three data centers approximately nine petabytes of data. If the cluster of server racks we discussed in chapter 2 are the heart and lungs of the Archive, the terabytes and petabytes of data—ingested, collected, processed, derived, indexed, “made dark,” saved, archived, and more—are its blood. Data is a need and an object of constant pursuit: the Archive wants as much data as it can get, and, as well as I could determine, there will never be too much. Such is the computational ethos: data is good. At Friday Lunch, a weekly event open to the public at which those gathered tell the others what they’ve been working on that week, Archive employees routinely announce milestones involving large quantities of data: crawls, ingests, and downloads of 5, 25, 45 terabytes; new server racks that increase capacity by so many petabytes; and similar accomplishments. Such achievements are routinely met with “oohs,” applause, or other expressions of affirmation. No one ever mentions weeding, winnowing, editing, selecting, or deleting data. Data goes in one direction at the Archive: up.

Undoubtedly, these are big numbers—a petabyte is a unit of information equal to one quadrillion bytes, or 1000 terabytes (10^{15} bytes). *A quadrillion!* Nonetheless, such pronouncements about terabytes and petabytes were often lost on me. Even though I knew that I should be impressed, I had trouble comprehending the figures in a way that was meaningful to me, even when comparative units of measurement were used: that a book or a daily newspaper is one megabyte; all the books in the Library of Congress are 29 terabytes; or that one petabyte is equivalent to 799 million copies of *Moby Dick* (Hardy 2012; Markoff 2009). Based on my own background, experiences, and predilections—none of which had been deeply informed by computation—my sense of what was valuable or impressive didn’t correlate with such metrics. But I was an outsider, ignorant of “big data”, its uses, its importance. The Archive was my introduction.

Kahle, on the other hand, came of age hungry for data. He began his career at MIT, studying as an undergraduate in the Marvin Minsky’s Artificial Intelligence Laboratory, graduating in 1982. The next year, he joined Thinking Machines Corporation, one of the first parallel supercomputer makers, as lead engineer for the following six years. Thinking Machines was founded to commercialize fellow

MIT student Danny Hillis's Ph.D. research on "massively" parallel processing for artificial intelligence applications. Until parallel processing, machines processed serially—one processor at a time—and so parallel processing promised to greatly enhance computation by allowing multiple processors to work at the same time. The Connections Machine, as the company's one product was called, had more than 65,000+ parallel processors.

Kept afloat by DARPA contracts, Thinking Machines was a glamorous, if short-lived (1983-1994) tech company that once upon a time "had cornered the market on sex appeal in high-performance computing," with its beautiful machines featured in Hollywood films, its showcase offices in Cambridge, MA, and its gourmet cafeteria (Taubes 1995). But it was also a "hacker's paradise" that brought together lots of talent to build and work with its state-of-the-art supercomputers. Two computer science researchers working at Thinking Machines, Craig Stanfill and David Waltz, pioneered what they called memory-based reasoning, which spurred a shift in artificial intelligence research away from rule-based reasoning to memory (Stanfill and Waltz 1986; Waltz and Kasif 1995). Whereas before a computer had to follow a set of programmed rules to come to a conclusion, with memory-based reasoning it would instead sift through its memory to come to a conclusion, solving new problems by analogy with old ones. Such a process required, obviously, a lot of stored information (i.e., data), and the more the better. The problem, however, was that they didn't have a lot of data for the machine to process. As Kahle told me of the Connections Machine: "It didn't know anything.... It needed to read some books."

Pursuing the applications of the Connections Machine was how Kahle explained to me he got "into big data." In addition to his chief role working on the hardware of the supercomputer, Kahle was active "on the side" experimenting with the machine: "Once we had a machine, then I could start to play with it, we could have it read stuff. I found databases from all over the place and collected databases, to feed the machines." Then, compared with today, electronic data was fairly hard to come by and, when it did exist, it usually had to be bought. He found himself less and less interested in building machines and more and more interested in building an ecology or infrastructure for "content" on the Internet. Toward that goal, in 1989, he left the Cambridge offices of Thinking Machines to test the commercial prospects of one of its internal projects: Wide Area Information Servers (WAIS). Anticipating the World Wide Web, WAIS was a full-text retrieval system that made it possible for a remote user to search through databases stored in distant supercomputers. In 1992, Kahle broke away completely from Thinking Machines—which would go bankrupt in 1994 after DARPA ceased its patronage—to start an independent company, WAIS, Inc. His first contract was using WAIS to build a communications network for Ross Perot's presidential campaign, enabling the field offices to connect back to the national office. WAIS Inc. would go on to help a variety of companies and government divisions to establish client-server networks

or, in Kahle's words, to "go online": the Library of Congress, *Encyclopedia Britannica*, the National Archives, the *Wall Street Journal*, among others. In 2012 Kahle was among the inaugural inductees to the Internet Hall of Fame, specifically for the development of WAIS, which the commendation describes as the "first Internet publishing system."²

Throughout this time, Kahle considered himself to be building, incrementally, "the Library," an interconnected ecology of users, content providers, and the open client-server architectures that would connect them. He sold WAIS Inc. to AOL in 1995 with the belief that AOL wanted to use the WAIS technology as a part of building an open infrastructure for "content" but was disappointed to discover otherwise, and left the following year. With the proceeds of the WAIS sale, he started (with Bruce Gilliat) two new enterprises simultaneously in 1996: a for-profit company called Alexa Internet (whose name was meant to invoke the ancient Library of Alexandria³) and a non-profit organization called the Internet Archive. Alexa Internet "crawls" the Web—crawling is a computer program that scans and collects data on the Internet—and analyzes traffic to websites. The Internet Archive began as the archive that would preserve and eventually make publicly accessible the snapshots of the Web that Alexa was generating through its crawls. With the two new enterprises, Kahle had embarked on a truly Big Data project: to capture, analyze, and archive the ever-growing World Wide Web.

After only three years, Amazon bought Alexa Internet for a reported \$300 million—Kahle refers to this windfall as having hit "the dot.com jackpot"—which enabled him to turn his attention more directly to the non-profit Internet Archive. While he was building Alexa, the Archive was little but a "dark archive" of Web data. A dark archive refers to digital storage that is not publicly viewable. The next important steps Kahle undertook were to build a means for public access to the Web data—now known as the "Wayback Machine"—and to find and/or build other collections for the Internet Archive. The Archive has continued to crawl the Web and to archive it. That massive collection has grown today over these sixteen years into "over six petabytes of data," as Kahle will boast during the tours he enjoys giving to visitors. But the Archive stores a wide range of public domain cultural artifacts across four broad categories: moving images, audio, texts, and live music. Among them one can find industrial films, home movies, original manuscripts, every live Grateful Dead concert, U.S. government propaganda films, audiobooks, old commercials, pornography, sermons, vlogs and recorded testimony of many sorts—and much more uploaded every day by individuals around the world. The *New*

² <http://www.internethalloffame.org/inductees>.

³ The company had what Kahle refers to as a "founding document": Luciano Canfora's *The Vanished Library*, a history of the Library of Alexandria (Canfora 1989).

Yorker has described it as “an eclectic electronic bazaar.”⁴ The Archive is also engaged in the continuous 24/7 recording of 20+ television stations around the world since 2000, which is for the most part kept in a “dark archive” for copyright reasons.⁵ Last but not least, of course, among its stores of data are the books—three million of them—to which I now turn.

Books as Data I: Virgin Forest

In late October 2008, several weeks after I had started my fieldwork, Brewster Kahle visited the University of Michigan Information School to deliver the annual John Seeley Brown Symposium Lecture.⁶ The context was a bit awkward, since the University of Michigan was Google’s primary library partner and the head of the library, Paul Courant, would be responding to Kahle’s talk. Michigan had been working with Google as early as 2002 and was the first institution—and, until 2006, the only—to give Google access to their entire book collection for digitization. Other library partners had agreed to provide only public domain (pre-1923) books, but Michigan made all of their over seven million books available for Google’s scanning.⁷ The close affiliation of Michigan’s library leadership with Google and Kahle’s close affiliation with the counter-effort, the Open Content Alliance, had placed them on opposite sides of the pro/anti Google fence. In his lecture, entitled “The Closing of Library Services/The Opening of Library Services,” Kahle argued that digital distribution was, contrary to expectations, not opening up the library system but closing it down. Whereas the traditional (physical) library system has been open and robust, the digital one is in danger of becoming the opposite. He cited gradual shifts: from local control to centralized control; from non-profit to profit; from governance by “law” (copyright) to governance by contract; and from diversity to homogeneity. He suggested a few steps that libraries should take toward building an open, decentralized digital library system: 1) digitize everything in library and archive collections and as quickly as possible; 2) make digitized material accessible: the public domain books without restriction, and the in-copyright books through loaning, following the tradition of the lending library; and 3) allow and facilitate “bulk” or “research” access.

⁴ Anthony Grafton, “Adventures in Wonderland,” *New Yorker* only edition, November 5, 2011. http://www.newyorker.com/online/2007/11/05/071105on_onlineonly_grafton

⁵ The exception is the September 11 archive that was made publicly accessible. See: <http://archive.org/details/911>

⁶ I did not attend the lecture but watched the live webcast. Video is available at: <http://archive.org/details/BrewsterKahlesMichiganTalk>.

⁷ In August 2006, the University of California and Google came to an agreement to digitize both in-copyright and public domain books from its libraries so that California and Michigan became the second Google library partner to allow the scanning of in-copyright books.

This occasion was the first time I had heard of “bulk access” as a possible use of the Archive’s books. Bulk access is a special type of access that allows a programmer, Web developer, or researcher to take large quantities of data from a website. “Bulk access is making [possible] some of the most interesting things that are going on now,” Kahle said. “All sorts of interesting things are coming out, if you go and Hoover up and play with large datasets.” Kahle provided a few examples from work that was being done using the Internet Archive’s books as data: geneticists data-mining the digitized books of the Biodiversity Heritage Library; scientific entrepreneur Steven Wolfram’s “scraping” the digitized historical math books for his Mathematica computer program; and classicist Greg Crane’s work extracting ancient place names from digitized books in order to create a new index structure for Classical literature—and more. “There are possibilities here that we really don’t know,” Kahle said, and “much would be lost” if such research capacities were only to be available to the “big guys” (e.g., Google, Elsevier, Amazon).

Less than a week after he gave this speech, on October 28, the Google Book Search Settlement was announced, severely challenging Kahle’s vision for the future of a digital library system. The agreement laid out an implementation for each of the three aspects of Kahle’s “open” library—mass digitization, public access, and bulk access—but on contrary terms to those Kahle had spelled out. The Settlement, were it approved, would make Google Book Search the epitome of the negative vision that Kahle was warning against in his Michigan talk. It would make Google dangerously powerful as a central portal—or what Kahle calls a “chokepoint”—in the digital book ecosystem, closing it down rather than opening it up. In this chapter I am focusing only on “bulk access,” which the Settlement addressed through the establishment of what it termed a “research corpus.”⁸ I will discuss the Settlement’s research corpus in greater detail later in the chapter, but, for now, it was to be, essentially, a collection of Google’s digitized books made available specifically for computational research at two unspecified university libraries.

The prospect of research access to what would then have been seven million books (and which at the time of writing is more than twenty million) seemed to usher books into the emergent assemblage around Big Data. Publishers, scholars, and others had been producing digital books in one form or another since the 1970s—ranging from the volunteer Project Gutenberg to curated scholarly projects (Perseus

⁸ The “research corpus” provision was one of three prominent public interest features in the Settlement, added to make the deal more palatable to the public. The other two public interest features were: 1) a public access service to be offered free at every public library; and 2) the conversion of all books covered by the Settlement into formats that would be accessible to the blind and visually impaired.

Library and the Whitman, Rossetti, Blake archives) and the commercial microfilm conversion projects *Early English Books Online*, *Eighteenth Century Collections Online* among other publisher databases—but they are dispersed across a patchwork of small digital libraries, some accessible for research, some not. And, even if all were combined, Google’s effort would still dwarf them in size. When the MONK Project, a scalable text analysis website for humanities scholars, ingested a variety of such databases, the total corpus came to an impressive-sounding 151.6 million words (Mueller 2007). But that figure is the equivalent of just two to three thousand books: chump change when it comes to Big Data. The Internet Archive had been making its book data freely available, but, still, in 2008 it had only 1.5 million books to offer. Google had an exponentially larger number of books to offer as “data.” And Google’s collection had the distinction of including not only public domain books but also books in copyright—a rare feature of existing datasets—which made it all that much more attractive to researchers. Google’s digitized book collection was on track to become larger than nearly every university library’s print holdings—the largest, Harvard, has around 17 million volumes—and, if it kept on track, would become as large (or larger) than the largest book collections of in the world, the Library of Congress, which holds around 35 million volumes.

The unprecedented availability of what one researcher described to me as “600 billion data points”—or words—presented an alluring challenge, a knowledge frontier. News of the corpus excited historians, computer scientists, linguists, folklorists, bibliographers, literary scholars, and others with an interest in text mining and “data science.” Universities began to put together plans in a competition to host the facility. They anticipate a remarkable and unique new dataset not only for its sheer size and scale but also its ambition to universality (“all the world’s books”), whatever that might come to mean. Without knowing for sure, they feel certain that knowledge was to be had from millions of digitized books once aggregated into one dataset. Some see the corpus as a record of language use over time; others as a source of historical evidence; an information science testbed for building and refining new retrieval tools; or, simply, as a space of play and experimentation. The director of the digital humanities office of the National Endowment for the Humanities (NEH) characterized the potentials in terms of scale:

If you're a scholar of, say, nineteenth-century British literature, how does your work change when, for the first time, you have every book from your era at your fingertips? Far more books than you could ever read in your lifetime. How does this scale change things? ... How might searching and mining that kind of data set radically change your results? How might well-known assumptions in various disciplines fall once confronted with hard data? Or, perhaps, how might they be altered or re-envisioned? (Gold 2012, 63).

At a public discussion of the proposed Research Corpus, one social scientist anticipated “unknown future application[s] in this space that we can’t fathom ... until researchers have the opportunity to deeply explore this kind of dataset... Maybe it will provide the basis of future understandings about how culture evolves and changes. Who knows, maybe a Darwin-like figure might develop some interesting new theoretical ideas about cultural change.”⁹ Such exploratory hope for reinvigoration and renewal for the human sciences through computational methods has put wind in the sails of the beleaguered humanities. Indeed, the field of digital humanities may now be the only area of the humanities that could be reasonably described as burgeoning, finding support from both public and private funding agencies. In the US alone, the NEH, the National Science Foundation, the Institute for Museum and Library Services; the Mellon Foundation; the Sloan Foundation; the Arcadia Fund; and, as I will discuss, even Google are eager to effect a “digital turn” in the human sciences.

Books have become re-encharmed as data. The books—recomposed, made machine actionable—become a new sort of scientific instrument, enabling the analyst, with the assistance of computation, to see what he could not see before. The eminent literary scholar Franco Moretti, a champion of quantitative methods in the study of literature and brander of the concept “distant reading,” has remarked of Google’s book digitization: “It’s like the invention of the telescope. All of a sudden, an enormous amount of matter becomes visible” (Moretti 2000, 2005; Parry 2010). What was too distant to be perceptible comes nearer and is made accessible to the observer. For Moretti, the availability of “all” digitized books makes it conceivable that researchers could collaboratively discover the structure and laws of literary form. Coming from an entirely different disciplinary background than Moretti, bioinformatics researchers Jean-Baptiste Michel and Erez Lieberman, who were the first to work with all of Google’s massive book data and to publish their results, have branded another analytical method, which they call “culturomics” (pronounced with a long O). Having heard about Google’s book digitization project, the two researchers approached Google and asked for permission to work with the book data for their research on language evolution (Lieberman et al 2007). They began with a specific interest in irregular verbs, but their work at Google grew in scope. Teaming up with a group of fellow Harvard researchers, as well as others from the *American Heritage Dictionary*, *Encyclopedia Britannica*, and Google Books, they developed Google’s book data into a formal, scientifically valid dataset and as a “dry run” of a research corpus made from Google’s digitized books.

⁹ Archaeologist Eric Kansa, speaking at the Conference, Google Book Search Settlement and the Future of Information Access, University of California, Berkeley. August 28, 2009.

In the words of Lieberman, books are only one “shard” of culture but, because of Google’s mass digitization, they are the most amply available “shards” to science. Like Moretti and his colleagues who dubbed their group a literary lab, Michel and Lieberman call theirs a “cultural observatory.” Both are taking books into the spaces of science. Others have analogized Google’s book collection to Big Science/Big Data projects such as the Sloan Sky Survey, but no analogy is as common as that of the Human Genome Project (Crane 2009; Hand 2011). Also like Moretti, Michel and Lieberman offer their own version of Moretti’s “distant reading.” Rather than reading a *few* books *very* carefully, culturomics methods allows a researcher to “read *all* the books *very not* carefully” (Lieberman and Michel 2011).¹⁰ But unlike Moretti, they marry Big Book Data to Big Science: their paper on “culturomics” appeared on the cover of *Science*, one of the most mainstream organs for disseminating science.

The paper, entitled “Quantitative Analysis of Culture Using Millions of Digitized Books,” has three purposes: 1) to introduce, manifesto-like, culturomics; 2) to present the findings of various investigations carried out, algorithmically, on the datasets; and 3) to describe the construction of scientific datasets from within Google’s total corpus. Regarding the first purpose, they define culturomics as “the application of high-throughput data collection and analysis to the study of human culture” (Michel et al 2010b). “High throughput” means that it involves large-scale, distributed computer processing. The choice of the name “culturomics” was meant to evoke genomics and proteomics—fields that “created data resources and computational infrastructures” and, in so doing, energized biology.¹¹ They believe that “culturomics” methods will do the same for the study of culture. Culture here has the general meaning of the creative production of humans. Books are just the first part of this new computational method. Painting, music, sculpture and so forth will come later, after digitization expands to include them. Meanwhile, text data is “easy,” in computational terms, when compared to audio and visual data. Michel and Lieberman, both charismatic young men, have taken culturomics on the road in an entertaining, well-rehearsed stage show that follows a common script. At the one I heard live at the 2011 Digital Humanities conference at Stanford, they concluded by saying: “We are at a remarkable time If you’re an average academic at any average time, no one gives a damn what you’re doing. But we’re at a moment while the world is really kind of interested in where these methods are going to go. So, we’ve been asking: What is it we should do with this moment?” They then presented a series of slides, each of which began “We will...” following by a list of aspirations for the Big Data/Big Science study of culture. These included: digitizing every text written before 1900; making “the entire cultural

¹⁰ “Not reading” has become something of a mantra within digital humanities. See Mueller (2007) and Clement et al. (2006).

¹¹ See the culturomics website FAQ: <http://www.culturomics.org/Resources/faq>.

legacy of the human race” accessible to any child within twenty years; raising \$10 billion and “collider-type” budgets; forging collaborations between computation scientists and humanists; developing *savoir-faire* from the Human Genome Project; teaching humanists to code and to interpret data while also teaching scientists to read carefully and to interpret texts; and even making the humanities “a principle engine of economic growth.” After all, they asked, what are Google and Facebook about but “understanding humans and data” (Lieberman and Michel, 2011a)?

The second aspect of the paper, which presents the findings of their investigations using culturomics methods, does not easily match their aspirations. The questions they put to the datasets include: How many and which English words do not appear in “authoritative” dictionaries?; How many new words are produced each year? How do irregular verbs become regular?; What artists and intellectuals have been purged from mention by repressive regimes?; and At what age does fame peak in different professions (actors, writers, scientists, politicians)? Because of the significant gap between the heat and hype around culturomics and its results, it has come in for a special loathing among researchers in the humanities. Among those with whom I have spoken, the work of Michel and Lieberman with book data was described as “nothing new,” the results “unsurprising,” the methods “only of interest to philologists,” the questions “simplistic” and “yes/no,” the claims “outsized,” the data access “exclusive”, and the notion of culture “absurdly reductive.” One scholar even described culturomics to me as without intellectual content: “It does not involve knowing anything. It’s all data as data.”

The third aspect of the paper—the actual work in preparing the datasets—takes us into the practical details of working with book as data.

Books as Data II: Noise, Dirt, and Other Hazards

At the same time as Michel and Lieberman project expansive visions onto what books enable when they become data, they also provide sobering, detailed descriptions of what is involved in making digitized books into a trustworthy source of knowledge. All data might be good when you’re collecting it, as Brewster Kahle would have it, but it is not all good when you are trying, in the words of the culturomics researchers, to “do science to it.”

I described the digitized book in chapter 2 in terms of two distinct but mutually reinforcing levels: 1) the page images and 2) the text data (OCR and metadata). In nearly every case, the book data that interests computational researchers is the latter: the text data that results from digital processing (as summarized in chapter

2).¹² The text data from books is best described as “found data.” It has not been collected in anticipation of future analysis. Rather, the book data has come about only as a by-product of the processes of mass digitization. It is neither structured nor unstructured data; rather it is somewhere in between; as one researcher put it, it is “relatively unstructured.” This data produced by digitization has two Achilles heels: “dirty” OCR and “noisy” metadata. As I’ll show, they are not the only problems—access and reuse is another major problem—but these two issues are the most evident and the most remarked upon.¹³

Complaints about the metadata in the public interface of Google Book Search are well known. Linguist Geoff Nunberg has staked a claim as a chief critic of Google’s efforts, but especially its metadata, which he has described as a “train wreck,” a “disaster,” and a “mish-mash wrapped in a muddle wrapped in mess” (Nunberg 2009a; 2009b). Adding to that, librarians Ryan James and Andrew Weiss have recently published a study of metadata errors in Google Book Search, which showed that 36 percent of the books in its sample had errors in their basic bibliographic information (James and Weiss 2012). They report that this rate is much higher than what has been found in other studies of metadata errors in library catalogs. In its defense, Google explains that it gets its data from 48 libraries and 21 different commercial metadata providers and many of the errors in Google Book Search come from this information provided to them. It has amassed over three billion book records and uses algorithms to sort through them to create a database of all digitizable books.¹⁴ At such a scale, many errors will occur. They have admitted that they, too, are responsible for some human error in data entry (Orwant 2009). In one instance, an experiment in the adoption of classification codes from retail bookselling resulted in absurd classifications such as Jane Austen’s novels under “antiques and collectibles”; and the medieval studies journal *Speculum* under “health and fitness”—an amusing engineering misfire that they since largely rectified.

Unsurprisingly, nearly all the digital humanities researchers with whom I spoke testified to the difficulties of working with Google’s books metadata. Problems range from the usual missing or inaccurate bibliographic information of who, what,

¹² Only one researcher told me that he only wanted to work with page images. He was a classicist who is developing, with colleagues, his own OCR software. Latin and Greek are not well served by commercial OCR software.

¹³ In a recent report conducted by the Hathi Trust Research Center among the same approximate group of digital humanities researcher I interviewed for this chapter, they concluded that OCR and metadata were the two areas where they could add value to data: by improving OCR and metadata (Varvel and Thorner 2011).

¹⁴ See “Declaration of Daniel Clancy in Support of Motion for Final Approval of Amended Settlement Agreement.” February 10, 2010.

when, where (that is, author, title, date, place of publication). But other metadata problems arise when researchers want some information that does not arise from the typical categories of library cataloging: such as the genre a work represents (e.g., is it a novel?); the gender of the author; or when the work was composed as opposed to when it was published; whether or not it contains multiple languages; among others. Metadata represents a variety of paths into a database. It guides a researcher across a large amount of information and enables them to gauge or calibrate their own methods of inquiry. But data seldom travels with just the right metadata for a given research question. After all, as pointed out in chapter 2, metadata is specific not to the data but to the researcher's needs.

Although one vividly described the Google metadata to me as “craPtastic,” for the most part researchers were more forgiving of the data than critics like Nunberg. They are grateful to have access to it, willing to take it as it came and to find ways to make it useful. In the following remark, two computer scientists working with the Internet Archive's book corpus typify the generally diplomatic disposition of the researchers with whom I spoke: “While we might hope that the size and historical reach of this collection can eventually offer insight into grand questions such as the evolution of a language over both time and space, we must contend as well with the noise inherent in a corpus that has been assembled with minimal human intervention” (Bamman and Smith 2011). As this remark indicates, computational researchers can be quite patient with noisy or dirty data.

In the supplementary materials to their *Science* paper, Michel, Lieberman, and their co-authors describe in considerable detail the series of steps through which they filtered Google's digitized book data (Michel et al. 2010b, 7). In a process they say took over a year, they devised and implemented a series of algorithmic filters that removed the books with the worst OCR and those with the most troublesome metadata. Because they wanted to chart historical word use, the date of publication was especially an especially important metadata category. Date of publication happens to be one of the weakest areas of Google's book data, so the team had to determine, first, the cause of the errors in publication dates and, second, a way of filtering the bad data out.¹⁵ After working to improve the data for over a year, their filters got rid of enough noise and dirt so that finally the data was “good enough to do science to.” By the end, they had reduced the original database by two-thirds, or by ten million books, which indicates the scale of the problem. Nonetheless, they still had five million books—or “500 billion words,” as the *New York Times* reported—with which to work.

¹⁵ The biggest source of publication date errors, they report, were from serial publications, so they removed all serials from the corpus. Gibbs and Cohen (2012) assess that their algorithms were sometimes too zealous: “tossing out—algorithmically and often improperly—many Victorian works that appear not to be books” (72).

Even if OCR is, as I described it in chapter 2, the “animator” of the digitized book, it is also just as notably a great source of frustration. OCR is never perfect; in fact, it is always “errorful.” From the perspective of someone imbued with the standards of print publication, where any typographical error is considered a flaw, all OCR is dirty because *all OCR has errors*. Even if the scanned image accurately captures a page, the OCR will introduce errors. It is never perfect. Even if OCR is 99.9% accurate, which it rarely is, there will be one error for every 1,000 characters, or about two errors per page. The quality of the original copy determines the amount of error, and library books, which serve as the basis of mass digitization, are rarely pristine: they often contain marginal notes or underlining, complex layouts or unusual textual elements, variable fonts, or discolored or damaged paper—all of which are likely to cause OCR errors. They might also contain textual elements in more than one language and many other complexities that confound the parameters of an OCR program. But the single biggest predictor of OCR errors is the age of a book: the older a book, the worse the OCR results. Although they might work to improve OCR processing, no mass digitizer corrects its OCR, though means of OCR improvement are always being entertained, such as “crowdsourcing,” wherein the task would be undertaken by online communities of scholars, readers, and others.

But, in truth, bad OCR is simply not the problem to computer engineers that it is to humanists and to the readers habituated to the standards of print. Information retrieval algorithms, in particular, are “very resistant to OCR” such that good retrieval rates can be had even when up to 50 percent of the words in a text are degraded (Lesk 1997). For this reason, OCR is plenty adequate for search, which is, as I pointed out in chapter 2, its main purpose. But the OCR data that is good enough for search and retrieval was never intended as the basis for the wide range of uses to which it is now being put. The standard practice for linguistic corpora or in scholarly digital textual editions has been to work with texts or databases that have been painstakingly prepared either by scholars or publishing companies—with part-of-speech tagging, text markup, word tokenization, and other forms of labeling.¹⁶ Textual scholars, in essence, curate every word of an important text. Indeed, to a textual scholar of Plato, every character matters. At scale, however, a different perspective develops and the traditional demands among humanists for precision become, as one researcher told me, “unrealistic.” And the bigger the dataset, the less the OCR quality matters. One team I spoke with, who were working with terabytes of data across many hundreds of books, told me that “noisy data is good”; another researcher told me that he actually want his data to be a little “messy” and that his methods are tolerant of orthographic mistakes. Yet another explained to me that noise is part of the job when working with big data: “The humanist’s job is to hear the signal in the noise.” Sacrificed for recall,

¹⁶ Even the non-scholarly Project Gutenberg has had a significant number of its texts proofread through the Distributed Proofreaders network.

precision matters less and the ability to tolerate ambiguity amid the dirt and noise matters more. Such is the ethos of data.

Researchers must “massage,” “scrub,” “filter,” “clean,” “stare at,” “get to know,” and play with data. One researcher told me that working with Google’s data was “very manual ... surprisingly so.” The culturomics researchers describe working with data as a practice that requires habituation, practice, care, attention, and an “intimate familiarity.” Working with data is a hermeneutics, and part of the interpretation is knowing the data deeply: knowing what is wrong with it, what the problems are, what its biases are. As Michel and Lieberman explain: “More than anything else, what makes someone a good scientist is the ability to interpret data effectively in the presence of an array of red herrings and potential confounds. The most crucial thing that makes such interpretation possible is detailed knowledge of how the data was collected and intimate familiarity with the data itself (the kind that results from staring at it, in different ways and through different lenses, for months and years).”¹⁷ Elsewhere, they remark: “Like a scholar who reads very carefully, you have to take a measure of care. Data always lies so you have to figure out how it lies so that you can get it to tell you something that’s helpful” (Michel and Lieberman 2011a). So, whereas quantitative models enable researchers to *not* read carefully, they nonetheless require that the analyst read the *data* as carefully as the scholar once upon a time read the book that has now become data. In other words, machines read the books and humans read the data. This new environment depends on a different literacy, or set of literacies, than the pages images do. Think as way of analogy to the contemporary telescope. Astronomers no longer look through telescopes themselves. Rather, computers gather astronomical data and the astronomers analyze the data collected.

In addition to the complications of noise and dirt, researchers must also deal with the issue of access to the data and restrictions that come with it: those highways into the forest are private roads. One of the key differences between the Harvard researchers who worked inside Google as guest researchers (Michel and Lieberman) and those who didn’t was the access they had to the data. Michel and Lieberman had access to all the book data but the outside researchers I interviewed had access only to the circumscribed data to fit a specific research agenda. Although they each were quick to tell me how much they appreciated the generous support they received from Google employees in receiving data, half the research teams I spoke with told me that, when they initially applied for the grants, they had wanted, and had expected, to get full access to the entirety of Google’s data, or what one researcher called “the keys to the empire,” as Michel and Lieberman had had. After receiving the awards, however, Google informed them that they couldn’t be given full access because of copyright constraints. Knowing this, each team reset their expectations accordingly and many chose to work only on books that were no

¹⁷ Culturomics FAQ. <http://www.culturomics.org/Resources/faq>

longer in copyright, which generally means, in the U.S., books published before 1923. In some cases, however, even after limiting themselves to working on books they understood to be in the public domain, researchers were told that copyright remained a concern. A solution was to either switch one's research to only public domain titles or to use work from keyword "snippets" (i.e., a keyword and a fixed number of adjacent characters to the left and to the right). To the historical researchers, the snippets approach was satisfactory for their researcher purposes; to others (interested in testing and refining computational methods) such restrictions were a bothersome constraint that had to be tolerated and worked around. One researcher actually folded the constraints into his research question: could he achieve what he wanted without full access to the texts? (The answer was yes.) In the end, only one team out of twelve worked with in-copyright works.

This differing in access, however, did not mean that the culturomics researchers did not face substantial constraints. When they asked Google if they could make the book data public so that other researchers could use the data and reproduce their results, Google told them that they could not, again because of copyright constraints. As Lieberman and Michel have elaborated in a series of public appearances, Google told them that five million books means five million authors, and five million authors means five million plaintiffs.¹⁸ Releasing the full texts of books would be copyright infringement (no matter how transformed or mangled by noise and dirt). To get around this problem, Michel and Lieberman devised an "alternative data type." They would release not the textual content of the books but *statistics about* the textual content of the books, in the forms of "n-grams." A term from computational linguistics, an n-gram is, essentially, a sequence of words. A 1-gram is one word, so "United States" is a 2-gram; "the United States" a 3-gram; "United States of America" a 4-gram; and "the United States of America" a 5-gram. From the books data, they built massive statistical tables comprising all one- to five-word sequences (all n-grams) in all the books in the corpus. This time series records how often each sequence appears in the books in any given year from 1800 to 2000.¹⁹ Then they took one final step to make their data "maximally innocuous" from the standpoint of copyright. They removed the lowest frequency n-grams, that is, those that occurred fewer than forty times. Why? Because, as Michel explains, "If [all word sequences] were included, then you could *theoretically* reconstitute the book" (Lieberman and Michel 2011a; emphasis mine). In other words, if Google released comprehensive statistical tables containing *all* n-grams, it would be the legal equivalent of distributing digital copies of the books. But by eliminating the least common n-grams, the company could release the data

¹⁸ Three of their appearances are available online: Lieberman and Michel 2011a, 2011b, and 2011c.

¹⁹ Data before 1800 is too "crummy" and the books published after 2000 are not from university libraries but publishers, factors that combine to make them too different from the rest of the corpus to include.

because, even if one downloaded it all, there would be no way to recompose any book in its entirety, thereby stymying the efforts of lurking copyright infringers.²⁰

The release of the data had been hard won, they write on their FAQ in answer to the question “Why do we only get the n-grams?”:

If the full text corpus could not be released, we thought it was wrong—from a scientific standpoint—to use the full text corpus to do analyses and then publish a paper containing results that could not be evaluated by the scientific community at large. ... We drew ourselves a line in the sand: if we couldn't release the N-grams tables in full, we wouldn't publish anything at all. Of course, we had no guarantee whatsoever that Google would release the N-grams. It was a huge risk and on many occasions.... We thought this work would never see the light of day. Over the course of the four-year project, we were eventually able to convincingly make the case that N-gram tables were an extraordinarily powerful tool and that they should be released. This is to a great extent a testament to our collaborators Dan Clancy, Peter Norvig, and Jon Orwant, who went to bat for the project time and time again.²¹

I infer from this statement a tug-of-war between Google's engineers and Google's lawyers. A copyright lawyer could entertain as plausible the implausibility that someone might spend the inordinate amount of time it would take to re-assemble a book out of its decomposed statistical fragments—or that it would even be possible—rather than buying it or finding it in a library. This only makes sense if the point is not to anticipate plausible human behavior but to defend oneself against the (in fact plausible) *accusation* from rightsholders that their books were being “given away” in their totality, when data derived from them is shared. Any evidence of such malfeasance would undermine Google's fair use defense in *Authors Guild v. Google*. Paranoia becomes a perverse necessity and breeds not just absurdity but confusion and uncertainty. The legal fog around digitization

²⁰ The n-grams are downloadable at <http://books.google.com/ngrams>. In 2006, two machine translation researchers at Google, Thorsten Brants and Alex Franz, released an earlier n-gram corpus of one trillion n-grams culled from Google's web crawl data, from which they also removed the low frequency n-grams. See the announcement at: <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>. The dataset is not freely available but can be licensed from the Linguistic Data Consortium. T. Brants and A. Franz, *Web 1T 5-Gram Version 1*, Linguistic Data Consortium, 2006.

²¹ Clancy and Orwant were, at the time, lead engineers within Google Book Search; Peter Norvig is the Google's Director of Research.

makes the digital books a form of hazardous material that requires special handling.²²

The Google Book Search Settlement was in large part a plan to defuse the hazardous material that copyrighted books have become. Whereas the main thrust of the Settlement was designed to solve the general problem of how to distribute and share copyrighted books—which I will explore further in chapter 4—the Research Corpus was designed to solve the problem of research access to the books as data. The details of the Research Corpus introduce us to the issue of the social legibility of books as data: this new potential created through the process of digitization. One of the key figures in devising the Google Book Search Settlement has described the negotiations about access to their book data as “ridiculously difficult” for its newness: “We were trying to describe something that we still don’t understand as a community. We don’t understand what the potentials are [so] to articulate it in a legal fashion was challenging” (I-School 2009). Although he’s wrong to suggest that “the community” was represented at the table negotiating the Settlement, he is right that the Settlement was charting new territory: What social relations inhere in data or are emerging around data? How do they correspond to those of the image layer (i.e., the book apparatus as we know it)? How does “data” differ from the conceptual object that is the book? Does copyright even apply to “data” produced through automation? Should the author of a book have a relation to the data derived from his book? And if so, what sort of relation? Who owns the data? Google because they possess it? The authors and publishers because they own the copyright in the books? Or the “public” because the facts are in the public domain if the works themselves are not? What interests do rightsholders have in their works once they have been recomposed into data? The rest of this chapter addresses the motion within the data layer of the digitized books.

The Research Corpus, Data, and Copyright

Let’s begin by considering how copyright addresses data. On the one hand, data has a clear status in copyright law: it is not protected by it. Section 102(a) of the 1976 Copyright Statute specifies what is protected: copyright extends to “original works of authorship fixed in any tangible medium of expression, now known or later developed,” and Section 102(b) qualifies that by adding that copyright does *not* extend to “any idea, procedure, process, system, method of operation, concept, principle, or discovery.” Through a history of case law, these two sections have come to be understood as identifying the “idea/expression” or “fact/expression” dichotomy: copyright protection extends to the expression of an idea but not to *the*

²² I do not want mean to single Google out here. Their book project is the subject of an ongoing lawsuit, which, if they were to lose, they would be exposed, given the scale of their digitization, to billions of dollars in statutory damages. Willful copyright infringement has a maximum penalty of \$150,000. Say that Google were to be found having infringed the copyright of just five million in-copyright books, conservatively; that could be \$750 billion in statutory damages.

idea itself. The intention is to entitle the author to protection for what is original to him but also to enable others to take and reuse the facts, ideas, and information that a work might contain. Thus, facts cannot be copyrighted. They reside in the public domain and belong to no one or to everyone. Facts are the flipside of “originality,” marking the point where authorship ends or begins (Jaszi 1994). Following the fact/expression dichotomy, the “data” of books—or any other data—are not eligible for copyright protection and thus fall outside of the purview of copyright. Data has no author.

Books are full of facts and ideas. In chapter 1, we saw how facts formed the “essence” of books for Paul Otlet and others who wanted to free “knowledge” from the encumbrances of books. The value was in the facts not in the authors’ “rhetoric.” Expression—“aspects of the author’s personality”—was expendable, superfluous, unwanted. He wanted to winnow out what he wanted to winnow out expression out to arrive at only the facts. Facts were knowledge and that is what was important to him. Copyright law sees it the other way around: the value that is protected is the expression, the very aspect of a book that Otlet was eager to discard.

On the other hand, however, U.S. copyright law is just as explicit in saying that *compilations* of facts do qualify for protection (Section 103) as long as that compilation involves originality, the fundamental requirement for copyright protection. This creates a tension: facts are not copyrightable but a compilation of facts is. When is a compilation original and when is it not? And when a compilation qualifies for copyright protection, how does this affect the status of the facts that have been compiled with “originality”? This question was at the center of the important 1991 Supreme Court case *Feist Publishing, Inc. v. Rural Telephone Service Co.* It had important findings. First, the court found unanimously that copying the listings of names, towns, and telephone numbers from a competitor’s telephone directory was not copyright infringement because the original directory had not involved the requisite originality to qualify it for protection. The standard for creativity and originality in a compilation is quite low—“The requisite level of creativity is extremely low; even a slight amount will suffice” (U.S. Supreme Court 1991)—but there must be *some*. And second, the court found that, even if a compilation of facts were eligible for copyright protection, that the facts it contained could not be.

A factual compilation is eligible for copyright if it features an original selection or arrangement of facts, but the copyright is limited to the particular selection or arrangement. In no event may copyright extend to the facts themselves” (U.S. Supreme Court 1991).

Because the standard of creativity in a data compilation very low, most databases are likely to be eligible for copyright protection. Copyright thus prevents the

copying of a database, even though taking and reusing the individual facts contained within it would be outside of copyright's reach. *Feist* was also important case for its explicit refutation of early appellate court rulings that granted copyright protection for mere effort without originality, no matter how much effort is expended in doing so. Earlier rulings had afforded copyright protection to facts in exchange for the "sweat of the brow"—that is, the labor and expense involved in creating it. "Copyright rewards originality, not effort." The Court wrote: "Copyright is not a tool by which a compilation author may keep others from using the facts or data he or she has collected."²³ European copyright law is different from the U.S. in this regard. Under its 1996 Database Directive, databases *as well as the* facts they contain are protected.

The Research Corpus, as defined in the Settlement, was to be "a set of all Digital Copies of Books made in connection with the Google Library Project, other than Digital Copies of Books that have been Removed by Rightsholders."²⁴ It would consist in the page images, the OCR files, and the metadata for all the books that Google had permission to include (that is, all but those books rightsholders had directed Google to remove). Google would, at unspecified intervals, update the Corpus as more books were scanned and added to its corpus. Housed at one or two different universities, the Research Corpus was to be independent from Google, which would be the source of the data but not the location of the data.

Here it seems germane to pause here to consider what the difference is between a digital copy—the Research Corpus is a set of digital copies—and a book "as data." To this point, I have been regarding the digital copy, once it becomes an object of computational attention (or "research"), to be data. Once the object of research, words themselves are no longer the vehicle of an author's expression but points of data, and *data has no author*. As such, as we have just discussed, "data" eludes copyright protection whereas, of course, a "digital copy" does not. Indeed, a copy is the very thing regulated by copyright. So, how should something like the Research Corpus be regulated: as a digital library full of copies to which rightsholders have certain exclusive rights; or as a depository of data that *as data* belongs to the public domain? Can it be both? If so, who speaks for it?

The Settlement was one way of answering these questions in the specific case of Google's book data. The terms of the Settlement Agreement would have governed the Research Corpus because it was a contractual agreement between Google and the copyright owner groups. As a contract, it would sidestep copyright law.

²³ Of course, there are a number of other ways that a database publisher can "protect" its data; most common are contracts and technological measures.

²⁴ Section 7.2(d) of the Amended Settlement Agreement. Available online at http://thepublicindex.org/documents/amended_settlement.

According to its terms, the Research Corpus would be available only to “qualified users” for “non-consumptive” research (more on “non-consumptive” below). Researchers would not be able to use the data in a commercial product or service nor could they build a product or service that competed with either Google or the rightsholders. But researchers would be free to develop algorithms from their research using the data, which they could then take and use to build a commercial service.²⁵ The host universities would be responsible for vetting researcher qualifications, approving their intended research and for making sure that the terms of the Settlement were adhered to.

Publishers were wary of the provision in the Settlement for the Research Corpus, partly because they weren’t familiar with computational research. Such uses are, for the most part, not of interest to either publishers or authors—at least they have not been historically—and have never been addressed in standard book contracts. Book publishing has not been a “data-centric” business as are Internet companies like Google. The authors and publishers took a lot of convincing. As Google’s Dan Clancy has described it, it was a sticking point in the negotiation:

When we first started [settlement negotiations], even on our team, when I was saying, “Let’s get this in,” they thought we were crazy [to think] that we would get the folks on the other end to understand what the hell we were talking about. They had lots of trepidation, [thinking]: “Am I giving the keys to something I know nothing about?” The fact that we got this far is amazing.

The Research Corpus also exacerbated the anxieties about security and control that were at the heart of the original litigation.

So why bother? Why was the Research Corpus, orthogonal as it is to the central concerns of the Settlement, included in it? The negotiations were secret, of course, but the motivations can be reasonably inferred. First, Google wanted to be appear publicly as generous and open so as to gain whatever goodwill might come from the donation of their book data for broad research access. After all, the Settlement worked greatly to their advantage and they wanted the deal to appear as broadly beneficial to the public as possible. Second, Google had been hosting in-house researchers (including Lieberman and Michel, above) and was aware of the interest researchers had specifically in their book data. As we’ve seen, those researchers wanted to share their findings and the data they were based on in the name of scientific reproducibility. The university library partners also wanted to provide research access to all of the book data. According to Clancy, outside research

²⁵ Details in this paragraph that go beyond that specified in the Settlement Agreement were stated in public by Google Books Chief Engineer Dan Clancy at events on either June 22, 2009, or August 28, 2009, on the campus of UC Berkeley. I attended both events and either audio-recorded them or took notes.

performed at Google presented logistical challenges, and a separate location for the book data, geared toward the interests and needs of academic researchers, seemed a desirable outcome. In order to make the books available, according to Clancy, Google needed some sort of legal shelter in order to house the book data off-site in order to shelter the company from potential third-party copyright infringement—in the event that a rogue researcher might choose to “liberate” a couple million of the books.²⁶ The provisions of the pertinent to the Research Corpus would have shifted that legal burden from Google to the universities that would host the Research Corpus.

“Non-consumptive” Research

Whatever the reasons, how did Google persuade the authors and publishers, obsessed as they are with the security of digital copies, to agree to two offsite locations overseen by libraries, which they distrust? As with the contorted solution found above to enable the release of the culturomics n-gram tables, lawyers found a compromise solution: the notion of non-consumptive research. A coinage of the Settlement, non-consumptive research denotes the circumscribed activity through which researchers could use the book data: that is, “research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book” (Section 1.93). In other words, non-consumptive research involves the *algorithmic, machine* consumption of a book as opposed to *intellectual, human* consumption (“understanding content”).²⁷ The prohibition on meaningful human interaction with textual content is later qualified thus: “It is permissible, however, for Qualified Users to read Protected material within the Research Corpus as reasonably necessary to carry out Non-Consumptive Research” (Section 7.2(d)(vi)(2)). Some insight into this can be gleaned from my conversations with the recipients of Google’s digital humanities

²⁶ This is not facetious. Paul Aiken of the Authors Guild told me that this was a fear that they had over the security of scans held by libraries.

²⁷ Examples include: image analysis and text-extraction (say, for improved OCR methods); textual analysis and information extraction (such as “concordance development, collocation extraction, citation extraction, automated classification, entity extraction, and natural language processing”); linguistic analysis (investigation of “language, linguistic use, semantics and syntax as they evolve over time and across different genres or other classifications of books”); automated analysis; index and search. Jennifer Urban (2010) provides the following “real world” examples: Franco Moretti’s work to apply techniques such as computational linguistics, data mining, computer modeling, and network theory to understand literary form (McGray 2009); the U.S. Geological Survey’s use of Twitter feeds to map earthquakes and their magnitudes (Dept. of the Interior, U.S. Geological Survey: Twitter Earthquake Detector (TED), <http://recovery.doi.gov/press/us-geological-survey-twitter-earthquake-detector-ted/>); and the now-familiar use of tags to create “word clouds” and other graphical representations of content such as the most-often-used words in U.S. presidential speeches (US Presidential Speeches Tag Cloud, <http://chir.ag/projects/preztags/>).

research grants. Of course, as I've shown above, almost all were working with circumscribed portions and then only one team was working with in-copyright texts, but I do think some insight is possible. Half told me that they engaged "intellectually" with the texts of the books in addition to algorithmic assistance. One team told me their work was 50% non-consumptive; another estimated 85-90%. Had they been working under the auspices of the Research Corpus, would their engagement have been considered "reasonably necessary"? One of the more computational researchers even became intrigued by the anonymous marginalia on the page images and ponders a future research project on just that aspect of the books. How would that square with non-consumptive research? Is there a copyright in anonymous scribbling in the margin of a text? One has to wonder how the host site would demonstrate compliance with this directive, or how researchers would police themselves. Would they stop themselves from reading more than, say, 200 characters at a time? What if the researcher became curious? What worthwhile purpose could possibly be served by regulated that?

Until the Google Book Search Settlement was announced, "non-consumptive," as far as I can determine, was used primarily in environmental discourse, especially around water use, where the meaning is something like "renewable"—that is, it refers to use that does not diminish the original resource. The "consumptive" in this environmental usage is the dominant and somewhat pejorative meaning of *consumptive*: "Tending to consume, use up, or devour, esp. wastefully" (OED). The two meanings—not harming a water resource and not reading books—seem to have nothing in common. But, given the authors' and publishers' belief that some harm is done to them when their books are "consumed" or used without compensation or without explicit permission, the allusion to harm in the borrowing of "non-consumptive" from environmental discourse nevertheless may seem reasonable from their perspective. If they believe that new electronic uses present new markets, then this free use of their copyrighted material by researchers—or anyone—would be seen as a harm to them. We see in "non-consumptive" another compromise contortion that, from the outside, might seem absurd but, from the inside the negotiations, was utterly serious. And, however self-interested or unworkable, it was productive, as we will see below.

Non-display

Non-consumptive had a semantic companion in the Settlement—"non-display"—that served to regulate not outside researchers but Google itself. It is yet another negation that served to delineate the space of non-regulation from that of regulation and to indirectly name the space of the machine inside the book. The Settlement defined non-display uses as: "uses that do not display Expression from Digital Copies of Books ... to the public, with examples that include indexing so that a book can be searched; the display of bibliographic information (metadata), algorithmic listings of key terms; and "internal research and development."

“Display uses” were Google’s uses that involved public display of some portion of a book and the Settlement was centrally about those uses, which were being commoditized in various ways. Non-display uses involved machine processes that happen under the hood, unseen, and they were cordoned off from the regulatory oversight of the Settlement. And, importantly, although the Agreement gave authors and publishers the right to remove their books from all uses described in the Agreement, including the Research Corpus, they expressly *could not* remove their books from Google’s “non-display” uses. Again, to Google go the data spoils. Even though these millions of books were being removed from the Settlement and thus from the Research Corpus, they were not being removed from Google’s servers. Google negotiated the Settlement such that it would not diminish their books as data. Therefore, the Research Corpus was not the same resource that a researcher within Google could make use of. A truth of the digital future being constructed thus far is that the most likely possessors of the “universal knowledge” promised by digitization will be small cadres of engineers at Google and other large companies engaged in many forms of digitization.

As a data engineering company that constantly mines web data in order to improve its search engine (and to sell advertising), Google is fully and fundamentally invested in the collection and analysis of large-scale data. In this respect, Google is both like and unlike the Archive. The Archive, like Google, amasses data with a great zeal, but Google also does things with it. As the *New York Times* has commented, in the context of Google’s collection of personal information from people’s home Wi-Fi: Google is where “engineers rule and data is viewed as a precious asset.... — grab the data and worry about filtering it out later. That’s the engineering mind-set, especially at Google” (Lohr and Streitfeld 2012).

In 2006, science writer George Dyson reported in an essay called “Turing’s Cathedral,” that during a 2005 visit to the Google campus, one of his hosts had told him: ““We are not scanning all those books to be read by people. We are scanning them to be read by an AI [an artificial intelligence engine]” (Dyson 2005). This anonymous remark has often been invoked to suggest that Google has ulterior or at least additional unspecified motives to its book scanning than those of providing online access to books. Given its particular obsession in learning how to use massive amounts of data to improve its business, a number of close observers have speculated to me that Google’s book project was not really about providing access (however restricted) to books but rather about acquiring and making use of the books as *data* for its own internal purposes: Google “didn’t want the books, they wanted the data”; or “they’re doing it for the AI.” I am not convinced of this but enough people believe it that I think it is worth noting. Clancy has dismissed the comments about Google’s interest in a Hal-like AI engine, pointing out that the company doesn’t need the books when it already has so much information from the Web itself, which dwarfs the data to be derived from books (Clancy 2009).

That comment only begs the question of why, if Google already has all the data it would ever need, it would take on the book digitization project, which has already cost the company “hundreds of millions of dollars”²⁸ and which continues to expose it to potential legal liabilities in the many billions of dollars. In 2002, co-founder Larry Page told an interviewer:

The ultimate search engine would understand everything on the Web. It would understand exactly what you wanted and it would give you the right thing. That’s obviously AI, to be able to answer any question, because almost everything is on the Web. So, we’re nowhere near doing that now; however, we can get incrementally closer to that and that’s basically what we work on (emphasis mine).²⁹

In that “almost” resides the impetus to digitize books. Without books, a crucial component of that “everything” would be missing. In the meantime, however, it is justifiable that they would want the books merely to create a superior search engine—no other allows you to search through the text of twenty million books in 478 languages. The remainder of their motivation remains a deferred investment in future possibilities, to which Big Data is seen to hold the keys. At a conference in 2010 on the “future of reading,” Google engineering manager Jon Orwant opened his remarks by saying that, unlike the other speakers at the conference, he would be discussing the future of machines reading as opposed to the future of humans reading (Orwant 2010). And, in this regard, Brewster Kahle agrees: “The stuff we’re digitizing, a lot of it I don’t think people are going to read. No, it’s computers that are going to read it, to make correlations and deductions out of these materials. It’s not a graveyard of human knowledge or detritus that we’re building, no. It’s part of a new world where humans and computers are living symbiotically in a sea of information.” And on another occasion, he put it this way: “Computers are the ones that are going to be the major readers of our books, if not already... That’s our future user community, if you will. ...The next generation that really gets to use these things is computers, and I think that’s really exciting.”³⁰

It is for these reasons that Google has jealously guarded the space of the machine from the copyright owners. Michael Boni, the lead lawyer for the Authors Guild, testified that, during settlement negotiations, plaintiffs sought to receive payment for Google’s so-called “non-display” uses. Boni here once again implies, unsurprisingly, that rightsholders believed they deserved compensation for non-

²⁸ See “Declaration of Daniel Clancy in Support of Motion for Final Approval of Amended Settlement Agreement.” February 10, 2010.

²⁹ August 28, 2002. http://www.youtube.com/watch?v=iD_L5CgmeEo&feature=related

³⁰ <http://archive.org/details/brewsterkahlelongnowfoundation>, November 20, 2011.

display/non-consumptive uses. Copyright owners frequently claim that novel uses of their work constitute an additional derivative market for which a license can be negotiated. By such a view, digitization in general represents numerous new opportunities for licenses. Indeed, the Settlement Agreement was one large license agreement between Google and the book publishing industry. “But,” Boni continued, “Google would not agree under any circumstance.”³¹ Indeed, had Google agreed to compensate the authors and publishers for the uses they were making of the books as data, then they would have compromised their consistent “fair use” defense—should the Settlement fail and litigation be renewed—that what they had been doing all along was a fair use that required no permission nor compensation. But the arcane detail, that Google refused to allow rightsholders to remove their books from Google’s own internal uses, is a crucial detail. It indicates a stark refusal to let the authors and publishers (and their claims of copyright) into the space of data, to protect the space of the machine from the clutches of the copyright owners.

Each of these coinages—non-display uses, non-consumptive researcher—marks off through their negation a space of non-regulation. Each seeks to demarcate the part of a book that is restricted or encumbered from that which is free, so as to make room for the digitizer to maneuver. Each negation marks what they perceive, in the words of James Boyle, as “reserved spaces of freedom inside intellectual property” (Boyle 2008). A book might be copyrighted but within it there are porous recesses insulated from copyright, where humans don’t consume, read, or see the books. Into this de-peopled space of machine reading, copyright only foggily extends.

Non-expressive

Yet a third negation has been introduced into the legal fog of books when they become data. Legal scholar Matthew Sag has argued that “non-expressive” be substituted for “non-consumptive” because it is more directly conversant with the terms of copyright law and is not beholden to the context of the Google Book Search Settlement (Sag 2009; Sag 2012). Sag argues that current case law supports a principle of “non-expressive use” that lies outside the purview of copyright protection. The purpose of copyright is to protect an author’s “expression,” which requires the communication of an original work to the public. However, if a work is copied for purposes other than communicating to the public, that copying does not infringe the owner’s copyright. In short, a copyright owner’s exclusive rights are limited to *expressive communication to the public*. Sag builds his argument upon a series of recent opinions that involved what he calls “copy-reliant technologies.” The cases found it a fair use for search engines to copy and archive (“cache”) web pages (*Field v. Google*) or to display copyrighted works in search results (*Perfect 10*

³¹ Declaration of Michael Boni in Support of Motion for Final Approval of Amended Settlement Agreement.” February 10, 2010.

v. Amazon); and for a plagiarism detection software company to copy and archive student term papers (*A.V. vs. iParadigms*). These cases make clear, Sag argues, that what copyright disallows is “expressive substitution” of the copyright work. Copying that occurs for *machine processing* cannot be an expressive substitution because it is not communicated to the public and thus can’t substitute for anything. In short, he concludes: “The rights of copyright owners do not typically encompass non-expressive uses of their works” (Sag 2009). From this perspective, books when rematerialized as data are a “breathing space” within the exclusive rights that copyright grants to an author: “Just as authors possess no copyright in the facts and ideas contained within their works, the rights of authors to control the copying of their works should not generally include copying that offers no possibility of expressive substitution because it is entirely non-expressive in nature” (Sag 2009). Sag’s discussion of the “non-expressive” has made explicit an implicit assertion of mass book digitization: it produces additional dimensions to a book that do not pertain to an author. The issue is new: before digitization there was no such copying that was directed only toward non-expressive uses. Sag argues that it is “the most significant issue in copyright law today” (Sag 2009).

Legal critic Pamela Samuelson, one of the closest observers of legal controversies around mass book digitization, has expanded upon Sag’s re-articulation of “non-consumptive” to propose it as an element of copyright reform (Samuelson 2011). She has argued for the establishment of general privilege to be carved out for making non-expressive uses of copyrighted works. In fact, she argues that non-expressive uses themselves may well advance the overall goals of copyright law by “promoting innovation.” Large companies with an interest in digitization would gain clarity from the current fog that shrouds digitization’s legality.

Conclusion

When the Settlement was rejected in March 2011, the Research Corpus went with it. The notion of non-consumptive research, however, has traveled beyond it, doing further work. In the wake of the Settlement’s rejection, the Hathi Trust, a repository created by Google’s partner libraries for the storage and maintenance of their digitized books, spun off a new entity called the Hathi Trust Research Center (HTRC). The HTRC would be a post-Settlement version of the Research Corpus made up of the digital copies that libraries received from Google in exchange for having loaned them to Google for digitization. This means that it wouldn’t include any digital copies that Google made for libraries that are not members of the Hathi Trust and that it cannot be a complete replica of Google’s internal corpus. Nonetheless, it now contains nearly nine million books. The HTRC had a celebratory launch at the Digital Humanities conference in Stanford in June 2011, just a few months after the Settlement was rejected, with Google sponsoring the reception. Google’s Jon Orwant welcomed the crowd and urged them to work on

the book data: “We have a moral imperative to figure out what we have. ... We are building highways into a forest.”

The HTRC has had to wrestle with the term “non-consumptive” as it begins to formalize its own research procedures without the shelter that would have been provided by the Settlement. Advised by its lawyers to stick to the exact term used in the Settlement,³² the HTRC has nonetheless reinterpreted it. They now define it thus: “Non-consumptive research involves computational analysis of one or more books *without the researcher having the ability to reassemble the collection.*”³³ This gets much closer to where the harm would be: not in reading (“consuming”) the work but in recomposing the data into “true” copies that could be circulated as such. Because such recomposition could only be done algorithmically, as the culturomics researchers acknowledged above, this redefinition locates the potential problem not in the reading human but in the programming human. To enable computational research while also protecting itself from the accusation of copyright infringement, the HTRC received a \$600,000 grant from the Alfred P. Sloan Foundation to figure out how to establish a “secure” environment for research. As Beth Plale, a co-principal investigator on the Sloan grant explains, the goal is cyberinfrastructure that guards against “conditions of unintended malicious user algorithms.”³⁴ It would establish “trust and verify” mechanisms in place to confirm compliance with non-consumptive research policy. In contrast to the vague directive not to read the books, the HTRC identifies the rogue researcher not by his excess curiosity but by his malignant code.

The notion of the “non-consumptive” has circulates beyond these disciplinary practices growing out of the dispute among litigants into a variety of broader efforts that are, among other purposes, aiding digitizers to “open” the modern book apparatus. A variety of efforts are now underway to establish a legal privilege for digitization undertaken for non-consumptive uses. In the UK, the Hargreaves Report included among its final recommendations to the British government a limitation on copyright that would enable uses that do not “directly trade on the underlying creative and expressive purpose of the work (this has been referred to as

³² According to John Unsworth, speaking at the Orphan Works and Mass Digitization conference, April 2012, UC Berkeley. Audio available: <http://www.law.berkeley.edu/11731.htm>

³³ “HathiTrust Research Center Receives Grant to Investigate Non-consumptive Research.” *D-Lib Magazine*. September/October 2011. <http://www.dlib.org/dlib/september11/09inbrief.html>.

³⁴ Press Release, “IU Data To Insight Center to lead Sloan-funded investigation into non-consumptive research,” August 9, 2011. <http://newsinfo.iu.edu/news/page/normal/19252.html>.

“non-consumptive” use)” (Hargreaves 2011, ch. 5).³⁵ The Report explains that the machine, as a proxy for a human reader, does not engage in the “normal exploitation” that copyright was created to regulate.

In data mining or search engine indexing... the technology provides a substitute for someone reading all the documents. This is not about overriding the aim of copyright – these uses do not compete with the normal exploitation of the work itself – indeed, they may facilitate it. ... That these new uses happen to fall within the scope of copyright regulation is essentially a side effect of how copyright has been defined [i.e., as the regulation of copies] rather than being directly relevant to what copyright is supposed to protect.

Copyright, the Report recommends, needs thus to be directed away from copying per se to the specifically *human* uses to which a work is put. (Many do not share this position, clearly, as the Authors Guild has recently filed a new lawsuit against Google’s university partner libraries for their role in Google’s project even though they themselves are not displaying their digital copies to the public.)³⁶

In the U.S., too, a similar momentum behind Big Data is helping digitizers. Efforts such as the Copyright Reform Act—a project of the think tank Public Knowledge—calls for a Congress to amend the fair use provision of the Copyright Statue to include “non-consumptive” uses (Urban 2010) and not just for, as in the case of the Google Book Search Settlement, for “qualified users” but for anyone. The Association of Research Libraries (ARL) too, has included non-consumptive research in its code of best practices regarding fair use (ARL 2012). The Code sets out eight principles to guide libraries. Principle Seven pertains to non-consumptive research and reads: “It is fair use for libraries to develop and facilitate the development of digital databases of collection items to enable non-consumptive analysis across the collection for both scholarly and reference purposes.” The ARL has written elsewhere that Principle Seven “may provide the most powerful justification for mass digitization of library collections” (Butler 2012).

The strategic concoction of “non-consumptive research,” invented by the litigating parties involved in the Google Book Search Settlement, has escaped the particular confines of that usage to circulate at the center of other, broader strategizing: to expand and encourage the creation of Big Data for the “benefit of society”; to

³⁵ In November 2010 British Prime Minister David Cameron announced an independent review, chaired by journalism professor Ian Hargreaves, of how intellectual property should support economic growth and innovation.

³⁶ *Authors Guild vs. Hathi Trust*, in New York Southern District Court. Case 1:2011cv06351.

promote economic growth; and, for my purposes, to achieve legitimacy for the act of digitizing a copyrighted book.³⁷

³⁷ On July 6, 2012, a group of copyright lawyers and digital humanities scholars submitted an amicus brief in *Authors Guild v. Hathi Trust* in defense of non-consumptive research as fair use. It appeared too late to be incorporated into this chapter. “Brief of Digital Humanities and Law Scholars as *Amici Curiae* in Partial Support of Defendants’ Motion for Summary Judgment.” Available at: <http://dockets.justia.com/docket/new-york/nysdce/1:2011cv06351/384619/ Doc. 123>.

Chapter 4

Books as Orphans

For the net generation, a work does not exist if it can't be found online. Even those who prefer to use materials in print prefer to find them online. Digital libraries are essential to meet these needs, essential to democracy and the cultivating of culture in today's world. Libraries are prepared to fund the digitization of these materials and provide equitable access to them. Their copyright owners, who see no market for [orphan] works, are not. They should not be allowed to deny access to them.

-- Frustrated librarian (2005)

In the mid 1990s, computer scientist and digital library specialist Michael Lesk wrote a treatise on the practicalities of digital libraries.¹ In it, he explained why, in his opinion, entire libraries had not yet been digitized despite the desire and expectation for digital libraries and the fact that all the necessary technology was available and in place: “The biggest reason is that we cannot easily find \$3 billion to fund the mechanical conversion of 100 million books to electronic form, plus the additional and probably larger sum to compensate the copyright owners for most of those books” (Lesk 1997, 3). The problems would seem to be merely financial; if \$6 billion or so could be found, the mass digitization of books would proceed uneventfully. But within Lesk’s statement lurk a number of more complicated themes, which this chapter will pursue.

Lesk perceives the project of creating digital libraries as about *books* even though libraries have always collected material other than books. Further, Lesk sees digitization as a collective project—how will “we” pay for it, he ask—but what collective does he assume or imply? Who are the “we” of digitization: the entirety of the modern book apparatus?; the library “community”?; an elite group of computer scientists?; scholars and researchers?; or all of the above? Are the copyright owners (authors/heirs/publishers) part of the digitizing “we”? If not, why aren’t they? Finally, Lesk anticipates digitization proceeding, first, with mass digitizers pursuing the permission of copyright owners and, second, with the copyright owners granting it. As it has played out, however, an entity—Google—stepping forward to finance mass digitization has not led to comprehensive digital

¹ Michael Lesk figured in the introduction as the program officer at the National Science Foundation who approved funding for the Million Book Project in 2001. Brewster Kahle refers to Lesk as “the father of digital libraries.”

libraries as Lesk anticipated. Indeed, fifteen years after Lesk's *Practical Digital Libraries* book and a decade after Google began its collaboration with the University of Michigan, digitization proceeds precariously, as some "stakeholders" seek to control, modify, or even stop it.

The last chapter focused on what I called the data layer of the digitized book: the largely invisible machinic aspect that harbors "digitality" and animates the digitized book. "Data" provides a speculative new space wherein digitizers and their allies summon the glimpsed potentials of the unregulated. This chapter, in turn, finds its anchor in the image layer of the digitized book, which is its conceptually more familiar aspect in that it represents or models the book and testifies to its origin as a printed book. It is a copy of a copy. As such, the image layer of the digitized book appears to fit within the established regulatory field of the printed book—namely, copyright. Mass digitizers seek to digitize every book that they can get their hands on and they want to do so in legitimately, in conformity with the law. But doing so is not a straightforward task, as this chapter will show in some detail. Can mass digitization and copyright be made compatible?

The problem of squaring copyright with mass digitization invites both creative problem solving and accusations of wrongdoing. Digitizers are tacticians. The problems they are solving in this chapter are not technical problems. Rather, the problems they are solving are regulatory: that is, legal, political, and cultural. The law is a constraint and they are routing around it. I focus on these efforts in order to investigate the conflicts and blockages within mass digitization—none of which is technological in nature. Rather, they are problems of property, power, and relation. In particular, I focus on the problem of the "orphaned book," a curious figure that has arisen at the center of mass digitization. The orphaned book is most often old, obscure, forgotten, and without commercial value, but digitization revalues it by converting its value into historical value. The orphan, like data in chapter 3, is an opening into the modern book apparatus that digitizers seek to leverage in order to wrest power over "culture" from copyright holders. A bastard born from the expansion of copyright, variation among publishing practices, the idiosyncrasies of individual books (and their authors), and the passage of time, the orphaned book comes into view through the digitizer's particular encyclopedism: a desire to map, standardize, aggregate, and to see "all the world's books" as one masterable data set.

Mass Digitization Confronts Copyright

As discussed in chapter 1, digitization is a reprographic practice in a long history of prior methods of reproduction. Until the turn of the twentieth century, printed books were not mechanically reproducible without intensive capital investment (such as a printing press) but over the course of the past century or so, it has become both easier and cheaper for someone with the desire to produce a new

copy of a book from an existing one. You can “destroy” that copy to make a digitized one by cutting off its spine and feeding it through a scanner. Such a photographed version of *Harry Potter and the Deathly Hallows* was posted on the Internet five days before the book’s original, tightly controlled official release date (Striphas 2009, 152). Or, as in the library digitization I am discussing, you can photograph the book so as to preserve the original. Daniel Reetz, the organizer of a DIY scanner community, marks his entry into the world of book scanning from the moment when, as a student at the beginning of a semester, he balked at paying \$450 for his textbooks when he knew he could buy two digital cameras for much less. He did just that, then built a simple frame, mounted the cameras, photographed the expensive textbooks, and returned them to the bookstore (Reetz 2010). He later won an Instructables prize for his DIY scanner and built a website and online forum that has attracted scanning enthusiasts from around the world. The Internet Archive even briefly hired Reetz to help spread DIY scanning while encouraging the DIYers to upload their scanned materials to the Archive. Digitization is not trivial but it is easy enough that people around the world are doing it and creating shadow libraries on the Internet.

Over this same period and, in part, because of the proliferation of cheap reprographic techniques, copyright regulation has steadily expanded to address new types of copying and forms of use. Many, including Reetz himself, would characterize what he did as copyright infringement (Reetz 2010b); and yet, when at the meeting where I first met him, he openly declared his transgression in front of what seemed a fairly appreciative crowd of legal professionals gathered to talk about the Google Book Search Settlement. (I even overheard a lawyer afterward from a public interest organization offer to represent him should he be sued.) The mass digitizers who are the subject of this dissertation, however, want not simply to capture books in digital form but also to make them available over the Internet—their new Library—and to do so they must confront and overcome the obstacle of copyright law. How, they ask, can book digitization be undertaken legitimately now that it is affordable, possible, and so vitally important that “we” do so? There is no easy answer to this question, and the difficulties of it have turned digitizers into inadvertent activists. They have also attracted a variety of actual activists to their cause.

For digitizers, books are vital and necessary cultural objects hopelessly caught up in the vagaries of copyright law. They see books, fundamentally and necessarily, as occupying one of three fundamental categories: the in print; the out-of-print (but in copyright); and the out-of-copyright (or “public domain”).² These “native” categories structure both perception and action. In contrast, when a library patron—or some other “ordinary” user—scans a shelf of books, she mostly likely is

² These categories commingle copyright status and commercial status (or availability), which don’t necessarily align. A book can be out of print and yet still “protected” by copyright.

unaware (and uninterested) in the copyright status of those books. To the reader of a printed book, copyright is quiet, but to the digitizer it is a thunderous roar. This hyper-consciousness about copyright surprised and challenged me when I began my research because, although I had worked as an acquisitions editor for a book publisher for many years, it seemed I had, in comparison to my new colleagues, a relatively naïve understanding of copyright law, despite it being the governing legal framework of my former profession. The experience of laboring to catch up on the complexities of the U.S. copyright law provided an “ethnographic” insight. I realized that as I worked within the stable book apparatus, the specifics of copyright had not proved particularly salient in the everyday performance of my job. Others confirm a general “ignorance” about copyright among publishing employees (e.g., Covey 2005a, 51). Despite its undeniably central importance, copyright, I came to realize, had been part of what Michael Polanyi called “personal” or “tacit” knowledge, what one knows in order to do one’s work but what one likely does not explicitly consider. Through this small example, we can see how in a stable apparatus, as opposed to an emergent assemblage, “expert” knowledge may well be tacit because routinized methods and procedures reproduce and maintain it. But to the mass digitizer—an agent in an emergent assemblage—copyright is utterly salient, never taken for granted, never *not* an issue. It is a tactical field of opportunity and risk, of thought and action.

Before I continue, I need to back up and explain how digitizers understand copyright to be misguided. Mass digitizers articulate three main points of criticism: copyright lasts too long; its scope is too broad; and it is too easy to acquire copyright protection. These factors have resulted in a situation that digitizers believe contradicts the original purpose of copyright. The Constitution specifies that it is the power of Congress “to promote the progress of science and useful arts, by securing for limited times to authors ... the exclusive right to their respective writings” (Article 1, section 8, clause 8).³ In line with a robust community of “copyleft” scholars, digitizers interpret this clause to mean that the purpose of copyright is to promote knowledge and learning by providing the public with access to new work (Patterson 1991; Lessig 2001; Litman 2001; Patry 2009, 2012; Vaidhyanathan 2004; Boyle 2008). It is an economic right, but one beholden to a public purpose. Mass digitization brings together, in a fractious collision, the recent tendency toward copyright expansionism and the generally recognized desirability for greater accessibility to published materials on the Internet.

One of Brewster Kahle’s many memorable turns of phrase is that he longs for “Nixon’s copyright.” What he means is that the problems with copyright, for him, began with the 1976 Copyright Act, passed two years after the end of Richard Nixon’s presidency. An omnibus revision of the 1909 copyright statute, the 1976 law (which went into effect on January 1, 1978) introduced a host of significant

³ I have eliminated the language here that refers to patents (“inventors” and “discoveries”).

changes. For digitizers, the most prominent (and problematic) among them are: the extension of the copyright term; the relaxation of procedural limitations to copyright (also called “formalities”); and the expansion of the scope of copyright. I will explain these three changes separately and then discuss how they have combined with subsequent developments to create what is now called the “orphan work” problem:

The term of copyright. The first U.S. copyright act (1790) created a federal copyright, which lasted fourteen years from a book’s publication. If the author was alive at the end of those fourteen years, then he could opt to renew the copyright for another fourteen years. If he did not renew the copyright, his work “entered” the public domain.⁴ The 1976 Act switched the basis of a copyright term from a fixed number of years after publication (with one renewal) to a variable term based on the author’s life plus 50 years (with no renewals).⁵ The effect of this change from the 1976 Act—as well as subsequent extensions that I will specify below—is that the average duration of copyright has increased dramatically. In 1973, under “Nixon’s copyright,” the average term of copyright was just 32.2 years. Today, because renewals are no longer required, the average term of copyright is the *maximum* term (which can surpass 100 years) (Lessig 2004, 135). Critics of copyright term extensions argue that the “limited time” of copyright specified in the Constitution has effectively been eliminated.

The nature of copyright protection. Before the 1976 Act, U.S. copyright law had extended only to published works: federal copyright was granted in exchange for making a work public, and common (or state) law regulated unpublished works. The 1976 Act eliminated this distinction by abolishing common law copyright and folding both published and unpublished works under federal law. This change meant that copyright protection now begins not when a work is published but at the very moment it is “fixed” in “any tangible medium of expression” (Section 102)—even if that medium is a soiled cocktail napkin. The effect of this change is that one must presume every cultural scrap is copyrighted.

Formalities. One goal of the 1976 Copyright Statute was to harmonize U.S. copyright law with that of Europe, as codified in the 1886 Berne Convention, so that the U.S. could take a more prominent role in international copyright.⁶ Before

⁴ The initial term was lengthened to twenty-eight years in 1831, and the renewal term to twenty-eight years in 1909, then to forty-seven years beginning in 1962, and to sixty-seven years beginning in 1998.

⁵ Lawrence Lessig has tallied that the U.S. Congress extended the term of existing copyrights eleven times alone between 1962 and 2001 (Lessig 2001, 107).

⁶ The vast majority of countries are now signatories to Berne.

the 1976 Act, one major difference between the Berne Convention and U.S. copyright law was that the U.S. required “formalities,” while the Berne convention prohibits them. Formalities are specific requirements that must be met to initially gain a copyright, such as registering it with the Copyright Office, including a copyright notice such as the copyright symbol—the letter “c” in a circle (©)—or deposit with the Library of Congress. Later, to keep and maintain that copyright, renewal was a further formality that guaranteed continued copyright protection. The 1976 Copyright Act did away with all such formalities, considering them too great a burden for the copyright owner and a “trap for the unwary,” thus making it much easier to obtain and to sustain one’s copyright. Indeed, under the 1976 Act, the copyright owner need never do anything to protect her copyright. The main effect of the removal of formalities is that it has become very difficult to figure out whether a work is under copyright and, if so, who owns that particular copyright.

Broadly, one might think of these changes together as the U.S. copyright system being changed from a formal system to an informal system, and developments since the 1976 Act have continued to complicate matters. In 1992, the renewal requirement for copyrighted works published before 1976 was dropped, keeping older books that would likely have entered the public domain from doing so; and, most notoriously, Congress passed the Sonny Bono Copyright Extension Act in 1998, which, in order to prevent Mickey Mouse from entering the public domain, extended all existing and future copyrights by an additional twenty years. These changes to copyright protection over time have drastically muddled the relations in and around copyrighted work by making the limited property rights that copyrights were supposed to be into something approaching perpetual and absolute rights over intellectual creation. To critics of these expansionary tendencies, the switch from formal to informal has moved the U.S. copyright system away from one that “balances” the interest of owners and users to one that overwhelmingly favors the owners. In the words of one of the most prominent (and sharp-tongued) copyright scholars, the changes to copyright in the past thirty or so years amount to a “social disaster” (Patry 2012, 200).⁷

Beyond the recent changes in copyright regulation, the relations in and around books have always been complex. Books necessarily have “authors” but authors are often not individuals or even “creators.” They are often corporations, publishers, estates, agents, and other representatives. A book’s publisher will change over time; it may be bought, sold, or go out of business entirely. Furthermore, the publisher and/or authors often contract with yet further individuals or businesses for “third-party copyrights” within the same book: illustrations, artwork, maps, graphs,

⁷ Patry is a well-known copyright scholar and author of the seven-volume treatise on U.S. copyright law, *Patry on Copyright*. He is also Senior Product Counsel at Google. In his published work (including Patry 2009 and 2012), Patry includes a disclaimer that he does not speak for his employer. I respect his right to speak only for himself, but I also feel obligated, in a dissertation that discusses Google’s book digitization activities, to acknowledge his employment.

forewords, afterwords, prefaces, contributory essays, and so forth. Books are full of people in varying relation to it and to one another, and the older the book, the harder it is to determine anything about such relations for sure. The changes in the regulation of the intangible rights in books outlined above have made them all the more complex—and, to the digitizer, impossibly so. If one needs the permission of potentially all of these people in order to digitize a book, how does one proceed in the circumstances henceforth described? The “gaze” of the digitizer is normalizing: it seeks modularity, manageable systems, standardization. Such a gaze magnifies the complexity—the “mess”—of books, but it also establishes that mess newly as a problem.

In terms of copyright, mass digitizers have had to develop “total” strategies designed to accommodate books of any copyright status. Digitizers think and act systematically: they are not digitizing this book or that book but books as an aggregation—that is, *all books*. Every newly digitized book is one book closer to *all books*. The goal is universality and totality, whether or not it is reached, which returns us to the three “native” categories that structure the digitizer’s behavior: the in-print; the out-of-print (but in copyright); and the out-of-copyright (or “public domain”). The digitizer must develop a strategy for each category. “In-print”⁸ books are those currently for sale from a publisher and copyright protected, and they account for, approximately, 10 percent of books.⁹ “Out-of-print but in-copyright” books are those that are no longer available for sale from a publisher—though they may well appear in secondary markets as used or rare books—but that remain copyright-protected. These books account, roughly, for 70 percent. “Out of copyright” (or public domain) books are those whose copyright has expired, and they account for, again roughly, 20 percent of books. The “orphan problem” arises out of the second category: books that are no longer commercially available from a publisher but that are still eligible for copyright protection.

Orphans arise at the confluence of two problems: the difficulty of determining the copyright status of a book and the difficulty of securing permission from the copyright owner to digitize a book. I will take these two problems separately. First, how does one know if a book is eligible for copyright protection? Books published in the U.S. before 1923 are in the public domain.¹⁰ Books published by the federal

⁸ In the context of digitization, the term “in-print” is of course a holdover from print publishing and is being slowly replaced by other terms such as “commercially available.”

⁹ The percentages presented in this paragraph are from Google Book Search spokespeople. Lavoie et al 2005 backs them up. See also Wilkin 2011 and Lavoie and Dempsey 2009.

¹⁰ This year is arrived at by the following calculation: Under the 1909 Copyright Act, [the fixed term of copyright] was 28 years with the possibility of a 28-year renewal term. Extensions by Congress lengthened the renewal period to 47 years, meaning that published works could have at most a 75-year copyright term (47 + 28 = 75). In 1998, with the Sonny Bono Copyright Term Extension Act, all

government are also in the public domain.¹¹ Books published between 1923 and 1964 are *likely* to be in the public domain because, during this period, both copyright notice and renewal after the first copyright term were required. Historical data show that copyright renewal rates have always been low—3 percent per year at the lowest and 22 percent at the highest (Landes and Posner 2003, 500; see also Ringer 1961). It follows that the vast majority of books published from 1923 to 1964 entered the public domain once their original term expired. Because renewal was done through the Copyright Office, its records can to some extent help determine if a book’s copyright was or was not renewed.¹² For the period between 1964 and 1978, some formalities remained and so it is possible that some work during this period may have entered the public domain, though it will be many fewer than those between 1923 and 1964. For the period after 1978, however, one must assume that all works are in copyright and will remain thus now for 70 years after the author dies, or, if a work of corporate authorship, 95 years from publication or 120 years from creation, whichever expires first. That is, if Congress does not once again extend the term.

Now, if these guidelines are hard to follow, it will be no consolation to be told that they are a simplification; there are exceptions that I have not listed.¹³ Furthermore, these guidelines pertain only to books published in the U.S. and not to books published abroad, which is especially pertinent because a significant majority of the books held in U.S. research libraries (and thus the books being digitized) were

copyright terms were increased by another 20 years, including the term for preexisting copyrighted works. Those works, however, whose 75-year term had expired before 1998 remained in the public domain, which means that all works published before 1923 in the U.S. have no copyright protection (Hirtle et al 2009).

¹¹ This principle is not as straightforward as it may appear. The work of contract employees working for a government agency but not government employees would not be included here. Government publications also often contain licensed material, which would remain protected by copyright. In addition, the provision pertains only to the federal government (not state or local governments), and it does not apply outside the U.S.

¹² A community effort has resulted in an easily searchable database—often called the “Lesk database”—of the copyright renewal records received by the US Copyright Office between 1950 and 1992 for books published in the US between 1923 and 1963. Available at: <http://collections.stanford.edu/copyrightrenewals>. Renewals received by the Copyright Office after 1977 are searchable on the Library of Congress website, but renewals received between 1950 and 1977 are not included. Public access to those records is only available through the card catalog in their DC offices.

¹³ Copyright scholars have distilled the patchwork of regulations into tidy charts to help people with the task. The best is Peter Hirtle’s “Copyright Term and the Public Domain in the United States,” which he first published in 2009 and keeps updated. Available at <http://copyright.cornell.edu/resources/publicdomain.cfm>.

published outside the U.S.¹⁴ The safest strategy for those who wish to avoid being sued for copyright infringement is to assume that all books published after 1923 are in copyright, unless evidence to the contrary is in hand. The public domain is fair game; everything else will require permission from the copyright owner.

Identifying, locating, and contacting the copyright owner is the would-be digitizer's second big problem. How does one determine who owns the copyright of a copyrighted work? No collective registry exists that one can consult. Nor is there a publicly available repository of book contracts available. The requisite information is not printed inside a book. The longer the term of copyright, the more difficult it is to track down who owns the rights. Authors move and, eventually, die; the companies to which they transferred their copyrights move, are sold to other companies, or just go out of business; they also sell rights to other companies, which also move, close, or sell them anew. Given these complications, identifying the copyright status of a work, who owns the work, and then locating them is a time-consuming, expensive undertaking. But, if the owner cannot be found, permission obviously cannot be granted, and, in our case, the would-be digitizer cannot move forward. All of these costs would be incurred without knowing the outcome of the effort—that is, whether or not permission will be granted. And, if permission were indeed granted, there might well be a licensing fee to pay on top of all the previous costs. Multiple this expense by millions to see it from a digitizer's point of view. The work required to determine copyright status and to seek permission would likely cost more than the digitization itself, and, at the same time, those efforts would be undertaken amid great uncertainty. The result has been that until recently, as Michael Lesk noted, books that are in copyright—specifically, in copyright but out of print—remained not just undigitized but *undigitizable*.

Orphans as Problem / Orphans as Solution

The name given to this impasse is the “orphan works” problem. The orphan work is the work caught in this conundrum: it is detached one way or another from copyright ownership but nonetheless its “owner” is entitled to privileges gained from that copyright. The broadest definition of an “orphan work” is a creative work that is ostensibly “protected” by copyright but lacks an apparent owner of that copyright. Beyond that, more precise definitions depend, of course, on who is defining it and for what purpose (see Covey 2005b; Hansen 2011). The “official” definition today is that of the Copyright Office: “An orphan work is a term to describe the situation where the owner of a copyrighted work cannot be identified and located by someone who wishes to make use of the work in a manner that

¹⁴ Wilkin (2011) notes that about 72 percent of the 5 million books in the Hathi Trust book corpus (at the time of his essay) were published outside the U.S.

requires permission of the copyright owner.”¹⁵ The meaning here, focused on the inability to locate the owner, measurably differs from how Brewster Kahle understood the term when he embraced it a decade ago. The best way to convey the nuances of the term and its significance over time is to chronicle its emergence through the experiences and activities of Kahle.

Kahle has repeatedly said, in public and private, that he (or the “we” that is the Internet Archive) coined the term “orphan book” (e.g., Kahle 2009a, 2009b). That may (or may not) be technically true, but it is true that he was part of conversations that took place in California in 2002 and 2003 that resulted in “orphan works” or “orphaned works” emerging as a national problem. In 1999 Kahle met and befriended Rick Prelinger, a film collector and “orphanista.” Orphanista is a term used to describe those devoted to rescuing, studying, and creatively reusing older non-commercial films (Streible 2007, 126; Cohen 2004). Such films have also been called “foundling” films or, in Europe, “non-fiction films” (Frick 2011, 153). Orphanistas galvanized the film preservation movement of the late 1980s and 1990s, and, from their activities, the orphan metaphor emerged (Streible 2007; Frick 2011, 120ff). As early as the 1950s, motion picture industry professionals had used the term “orphan film” to denote a feature film that had been judged unlikely to be profitable and thus unworthy of promotion. The authors of an important 1993 Library of Congress report employed the term “orphan” to differentiate those films that lacked either copyright owners or commercial potential to pay for their continued preservation from those films that had evident market value and/or owners available to pay for their preservation (Melville and Simmon 1993). Orphan films included newsreels, industrial films, outtake material, old silent films—almost anything that wasn’t a feature film. They had no one to care for their preservation because, in the words of one Hollywood studio executive, they “did not naturally fall under someone’s ownership” (Frick 2011, 143). In the 1990s, orphaned had come to mean not having a commercial benefactor.

Film archivists sought federal monies to protect the nation’s “visual heritage.” The politics of federal funding required the issue be sorted out on the basis of property ownership. Congress did not want to step on Hollywood’s toes by granting certain privileges to use copyrighted works (preservation copy and public display) nor did it want to pay for something that should be the duty of the copyright “owner.” The “orphan” metaphor allowed Congress to find a way to support film preservation, differentiating between those films that Hollywood should preserve and those that public archives should preserve (Lukow 1999). Responsibility for the preservation of *commercial* films was left to copyright owners (Hollywood film studios). Moving image archivists over time won gradual federal recognition and funding for orphans—in the various National Film Preservation Acts (of 1992, 1996, and 2005)

¹⁵ Register of Copyrights, Report on Orphan Works, 2006. Available at <http://www.copyright.gov/orphan/orphan-report-full.pdf>.

and the Preservation of Orphan Works Acts of 2005—and, in so doing, they became their cultural guardians.

But the problem was more complicated: the films were “orphans” not only because they had no commercial benefactors but also because they had no institutional *archivists* who considered them worthy of preservation (Frick 2011). The orphanistas sought to reform the profession of film archiving by expanding the canon of what should be considered worthy of preservation. “Orphan films” named a previously unrecognized and unappreciated part of the country’s “visual heritage,” which required care in the form of preservation. “Orphan” came to name not only the neglect resulting from having “no mother and father to take care of them” but also the neglect from not having been considered worthy of preservation by traditional film archives (Melville and Simmon 1993, 79). The term now refers to “all manner of films outside the mainstream”: from ethnographic films to medical films, found footage, surveillance footage, advertisements, home movies, and much more (Streible 2006). The main concern surrounding orphans films was not copyright law per se but with finding a means to preserve *all* film as critical cultural “heritage.”

Rick Prelinger introduced the “orphan” metaphor to Brewster Kahle. Prelinger had begun collecting films in the 1980s after working as a research editor on a documentary called *Heavy Petting* (1989), which the IMDb database calls a “a hilarious and salacious exploration of the sexual mores of the 1950s.” The experience got him interested in “ephemeral films”—educational, industrial, training, ethnographic, government, and social guidance films—and the 1980s was an advantageous time to start collecting because film was moving to video and lots of film was being sold off. When Kahle and Prelinger met in 1999, Kahle had just sold Alexa Internet to Amazon and was beginning to turn his attention to building collections for the Internet Archive beyond its archived Web pages. He asked Prelinger to consider “donating” his film collection for free public use at the Internet Archive in exchange for its digitization. At the time Prelinger earned his livelihood from his stock footage film business and his first thought was that to give away his collection for free on the Internet was “crazy.” But Kahle ultimately persuaded him, and the Prelinger Archive became the Internet Archive’s first freely accessible collection.¹⁶ It went live in 2001. In September 1999 the first Orphan Film Symposium was held at the University of South Carolina, and Prelinger was among the invited scholars, collectors, and archivists. Prelinger’s involvement with the “orphanistas” introduced Kahle to the term, and they discussed the utility of extending the term to books and the other materials that they sought to make available in their respective Archives.

¹⁶ The television archive was the Internet Archive’s first collection (after the Web archive) but it isn’t made freely accessible to the public as the Prelinger Archive has been. For more on the Archive’s television archiving, see Lessig 2004, 110ff.

Shortly thereafter Kahle developed another important friendship, this one with the legal scholar and “free culture” activist Lawrence Lessig. After having moved to California to take up a position at Stanford Law School, Lessig sought Kahle out and shared with him the manuscript of his forthcoming book, *The Future of Ideas*, which Kahle read and greatly appreciated.¹⁷ Lessig, with some Harvard colleagues, was then in the midst of the appeal of a lawsuit they had instigated, *Eldred v. Reno* (later, *Eldred v. Ashcroft*).¹⁸ The suit sought to overturn the Sonny Bono Copyright Extension Act of 1998 (mentioned above) as unconstitutional. *Eldred*, first filed in January 1999, had been unsuccessful in the lower court and, when Kahle and Lessig met in 2000, Lessig was in the midst of its appeal to the Ninth Circuit Court of Appeals.

Although Kahle was friendly with Richard Stallman during the early 1980s at MIT, and although he was ahead of the curve in giving away the WAIS software that he was otherwise commercializing, Kahle had not directed his efforts to specific copyright activism. In the early 1990s he had thought the federal government was doing “a good job” with regard to Internet policy, and he regularly expresses his admiration for Al Gore’s leadership in that area. Later, as he started to build collections for the Archive and to deal more specifically with “content” in the late 1990s, his perspective began to change and he became a supporter of a variety of what I am calling “copyleft” causes. In 2002, in his enthusiasm for Lessig’s *Eldred* case, he orchestrated a sort of publicity stunt bringing attention to access to public domain materials. Calculated to coincide with the October 9 Supreme Court oral arguments in *Eldred v. Ashcroft*, Kahle converted a minivan into a satellite-powered bookmobile outfitted to download, print and bind public domain books from the Internet, wherever the bookmobile might be. He drove across the country, stopping at schools, libraries, and museums—such as the Carnegie Library of Pittsburgh which, he admiringly points out, has the words “Free to the People” carved above its entrance—printing public domain books and giving them away.¹⁹ Kahle wrote on the Internet Archive website that, after the Copyright Extension Act, the public domain was “on trial” and that in deciding *Eldred v. Ashcroft*, the Supreme Court would be deciding “how many books are part of the digital library the Bookmobile brings.” Tugging at heartstrings that respond to bookmobiles as a sort of civic-inflected ice cream truck or lemonade stand, he concluded, “Without [public

¹⁷ When I first met Kahle in 2007, he gave me a copy *The Future of Ideas* as a gift. It is one of those important books that Kahle had on hand in multiple copies.

¹⁸ The named plaintiff in the suit, Eric Eldred, was a publisher who prepared e-book editions of books whose copyright had expired.

¹⁹ Kahle later traveled to Egypt and, with World Bank funding, Uganda, where he helped assemble similar print-on-demand bookmobiles.

domain] books, there can be no digital bookmobile.”²⁰ In addition to his own personal activities, the foundation he runs with his wife, the Kahle-Austin Foundation, has financially supported a variety of what I would characterize as “copyleft” organizations: the Free Software Foundation, the Electronic Frontier Foundation, Public Knowledge, the Samuelson Law, Technology, and Public Policy Clinic, the Public Library of Science, and QuestionCopyright.org.

Kahle’s collaboration with Lessig would deepen and, in as it did, “orphan works” would become a national problem. When the Supreme Court handed down its decision in January 2003, it ruled once last time against Eldred and Lessig, upholding the constitutionality of Congress extending the copyright term. Lessig and colleagues turned elsewhere, doggedly seeking other means to continue their crusade to limit the increasing scope of copyright protection. Lessig initiated a legislative effort, called first the Eric Eldred Act and, later, the Public Domain Enhancement Act. It was an attempt to reinstate formalities for copyrighted work by requiring copyright holders to pay \$1 and register their copyright after 50 years, or their work would automatically become part of the public domain. Zoe Lofgren, a Silicon Valley congresswoman, drafted the bill and introduced it in the House in 2003 and again in 2005. Although the legislation received some prominent support, it faced significant opposition, as expected, from the copyright industries and foundered (Lessig 2004, 248ff). In March 2003, the first newspaper mention of “orphan works” appeared in an editorial championing the Public Domain Enhancement Act.²¹ It described the proposed law as intending to “save orphaned works.” There, “orphaned work” meant work “in the dark zone”: out-of-print, without commercial worth and unclear copyright ownership.

Simultaneous to this legislative effort, Lessig and his staff at the Stanford Center for Internet and Society devised another lawsuit against the federal government, this one challenging not the copyright term extension per se but the constitutionality of the removal of formalities. Taking its legal principles from the ruling in *Eldred v. Ashcroft*, the suit argued that the wholesale change of U.S. copyright from a formal system to an informal system was a change to the “traditional contours of copyright protection” and thus required additional “First Amendment scrutiny.” Whereas the public domain had been the emblem for *Eldred v. Ashcroft*, orphans were the emblem of this new suit.

²⁰ “Internet Archive Bookmobile,” <http://archive.org/texts/bookmobile.php>

²¹ “A Simple 50-year Renewal to Save Orphaned Works; Long-forgotten Gems Could be Revived.” Editorial, *San Jose Mercury News*, July 7, 2003. Despite this early appearance, the term does not appear in the *New York Times* until September 28, 2005, Tim O’Reilly, “Search and Rescue.” <http://www.nytimes.com/2005/09/28/opinion/28oreilly.html>

In early 2004, Lessig asked Kahle and Prelinger to serve as the named plaintiffs for the new suit. In their early discussions, they debated the wisdom of naming “orphans” as the problem. Lessig appears favored the idea. He had used the term, if rather curiously, in his then forthcoming book *Free Culture*. Referring to the changes in copyright law that I laid out above, Lessig wrote:

Th[ese] change[s] meant that American law no longer had an automatic way to assure that works that were no longer exploited passed into the public domain. And indeed, after these changes, it is unclear whether it is even possible to put works into the public domain. *The public domain is orphaned by these changes in copyright law* (Lessig 2004, 135; emphasis mine).

I take Lessig to be saying that Congress had abandoned its obligation to enable copyright to nurture or contribute to the public domain, both by increasing terms, by removing formalities, and by making copyright protection automatic at the moment of any fixing in tangible form. Those works that *should* be the public domain had instead become “orphaned.”

This use of the orphan metaphor differs significantly from the orphanistas’ use, which, remember, was not solely about copyright but more significantly about the need for an organization or patron who would serve as a commercial benefactor with regard to a film’s preservation and its being made accessible to the public. As Lessig employs the metaphor, however, copyright law is the parent and the public domain is its child. Whereas the film archivists had sought new institutional attachments and relation, Lessig is more focused on righting copyright law. And the problem he describes seems less one of relation than one of production: the copyright system is failing to produce its rightful and eventual progeny, the public domain. When Lessig asked him, Kahle agreed to lend his name to the suit (as did Prelinger). It became known in shorthand as *Kahle vs. Ashcroft* (renamed in 2005 *Kahle vs. Gonzales*, which I will use hereafter).²² On the day he filed the suit in March 2004, Lessig announced it on his blog with the title “Save the Orphans.”²³ The suit defined orphan works as “books, films, music, and other creative works which are out of print and no longer commercially available, but which are still

²² *Kahle vs. Ashcroft* 3:2004cv01127; *Kahle v. Gonzales*, 474 F.3d 665 (9th Cir. 2007). The case was filed while John Ashcroft was Attorney General and concluded after Alberto Gonzales replaced Ashcroft in February 2005. More information available at: <http://cyberlaw.stanford.edu/our-work/cases/kahle-v-gonzales>

²³ Lawrence Lessig, “Save the Orphans,” March 22, 2004. http://lessig.org/blog/2004/03/save_the_orphans.html

regulated by copyright.”²⁴ Orphan works had arrived on the national stage as a specific problem *with copyright*.

During the formulation of these varying legal tactics to correct overreaching copyright legislation, Lessig and his colleagues, including Kahle and Prelinger, had recast “orphan works” into a much more wide-ranging problem than that of film preservation. Kahle refers to their adoption and extension of the orphan metaphor as “channeling Jack Valenti,” referring to the former long-time head of the Motion Picture Association of America, the leading Hollywood trade group. Reviled by copyleft activists, Valenti was an effective if inflammatory rhetorician (Patry 2009; Lessig 2004). One oft-noted example is his statement to Congress in 1982 that “the VCR is to the American film producer and the American public as the Boston strangler is to the woman home alone” (Patry 2009, 145). He led the MPAA for nearly four decades and during that time was successful in pushing a copyright expansionist legislative agenda that insisted on copyright being understood as an absolute property right, culminating in the 1998 double whammy of the Sonny Bono Copyright Extension Act and the Digital Millennium Copyright Act. In adopting a term from the film industry itself, Kahle felt that they were exacting a clever revenge against Valenti’s infamous rhetoric in support of “copyright maximalism” (Samuelson 1996).

Kahle v. Gonzales

It need be noted that *Kahle v. Gonzales* was not a significant legal case and now is something of a footnote to a series of efforts to stem the expansion of copyright. The district court dismissed it, as did the appellate court, and, finally, the Supreme Court declined to review it. As much as Lessig and his colleagues protested otherwise, the case was seen as treading the same legal ground as *Eldred*. In the context of this dissertation, I find it significant for two reasons: 1) as an important early articulation of “orphan works” and the first focus on the “orphan *book*”; and 2) as an full expression of the critique of copyright that mass digitization embodies. In the initial complaint (March 2004), Lessig and his colleagues at Stanford’s Center for Internet and Society defined orphans further as: “work that the author has no continuing interest to control, but which, because of the burdens of the law, no one else can effectively archive, preserve, or build upon in the digital environment.”²⁵ Copyright regulation had become an undue burden that “blocks the cultivation of our culture and the spread of knowledge” (2). Although the changes to copyright law had been driven by the “legitimate and valuable objective” of benefiting authors, they had introduced unintended consequences to the continuing use of

²⁴ Center for Internet and Society, “Kahle vs. Gonzales,” <http://cyberlaw.stanford.edu/our-work/cases/kahle-v-gonzales>

²⁵ “Civil Complaint for Declaratory Judgment,” p. 1-2. *Kahle vs. Ashcroft* 3:2004cv01127.

“knowledge and creative work” once its commercial life has passed. This “orphaning effect,” the lawyers wrote, was the “sole focus” of *Kahle v. Gonzales* (8).

Note that none of these definitions sees the orphan works as one for which it is impossible to locate the author—the core of the now official Copyright Office definition.²⁶ The crucial factor rather was that the work was no longer commercially available. The complaint cites the Internet Archive’s involvement in the Million Book Project, which, as I explained in the Introduction, was how the Archive became involved with mass book digitization. It explained to the court that:

Plaintiffs Kahle and the Internet Archive *do not intend to offer, free of charge, digitized versions of copyrighted works that are commercially available. They instead intend to provide access to “orphaned” works, while providing the author or copyright holder the right to request that is work not be made available. But because of copyright regulation, these “orphan” books cannot be made generally available. The project’s scope has thus necessarily been restricted. The result is that a vast number of copyrighted yet no longer commercially valuable works sit idle rather than enriching public knowledge* (18; emphasis mine).

The complaint makes the first reference I have found to the “orphan book,” which may confirm Kahle’s assertion (above) that the Internet Archive was the first to use the term “orphan books.” What is more interesting here, however, is this early meaning of orphan works. It’s an archivist’s meaning. We’ll see below that this original meaning will change considerably as the issue grows as a national issue, but it remains how Kahle understands the problem. The meaning given in the complaint is close to that of the orphanistas: films that were not being preserved (or made available) because no one cared about them. Like the film archivists, the Internet Archive was stepping forward as a custodian/caretaker for work that the hypothetical “owners” no longer valued.

The problem of orphan *books* has at least one fundamental difference from orphan *films*: in that non-commercial, obscure, specialized books haven’t lack for caretakers, whether commercial or not. The commercial status of a book has never determined whether a library acquired, stored or cared for a book. Indeed, it seems the opposite: the books that dominate the collections of Google partner libraries—all major research libraries—are largely low-circulation, out-of-print books. And, whereas orphanistas sought to legitimize and elevate the non-feature film as a part of a shared “film heritage” (Frick 2011), books are already well established as

²⁶ This definition is repeated in the appellate brief: “The Archive intends to provide access to a large number of “orphan” works, meaning work that remains under copyright, but that is currently out of print, and generally unavailable” (9). *Kahle v. Gonzales*, 474 F.3d 665 (9th Cir. 2007). Available at: <http://cyberlaw.stanford.edu/our-work/cases/kahle-v-gonzales>

“cultural heritage” (probably more than they should be). The problem that motivates the need to name “orphans” is the critique of libraries that is deeply embedded, if rarely explicit, in this specific critique of copyright. The traditional library of print materials, like copyright, is seen as an antiquated access system. Consider the comments of Bertelsmann’s Richard Sarnoff, who was the lead negotiator for the publishers in the Google Book Search Settlement: “What we were establishing was a renewed access to a huge corpus of material that was *essentially lost* in the bowels of a few great libraries” (Helft 2009; emphasis mine). Given the long-standing antagonism between publishers and libraries, such comments might not be surprising, but such comments are endemic in and around mass book digitization. It is common among them to hear, as I did countless times, that, because of *copyright*, books are “locked up” and “inaccessible.” But when I countered that, strictly speaking, the books aren’t locked up but rather they are available in libraries, the answer inevitably was: “But, to kids, if it’s not online, it doesn’t exist.” Or: “In twenty years, if it’s not online, it won’t exist.” Like the crusade to microfilm in order to forestall the consequences of paper perishability, mass digitization, in its copyleft, activist mode, seeks to preserve—indeed, to save—the twentieth century by finding a way to make it all digitizable. Locked up by libraries and locked up by legal regulation, books—at least those published between 1923 and, say, 2000—are in danger of being left behind, forgotten, left off the bus to the future. What digitizers seek foremost is a *new library*, and for the new library they need *new copyright laws*.

Strategizing for Orphans

As can be seen in the quotation from Michael Lesk with which I opened this chapter, the prevailing assumption had been that the digitization of a book required the permission of its copyright owner. In 1998, legal scholar Pamela Samuelson—herself deeply involved in mass digitization—wrote a piece on copyright and digital libraries, in which she made the point as clearly as one can: “Everyone who is developing a digital library knows it would be a big mistake to include digital copies of copyrighted works in the library without obtaining permission” (Samuelson 1998, 14). In 2005, the plaintiff lawyers in the *Kahle v. Gonzales* appellate brief repeated this belief: “If a digital archive such as plaintiff Internet Archive sought to make those out-of-print books available on the Internet, it would need the permission of the copyright owners of each work. (Electronic access, even if copies were not distributed, would infringe an exclusive right of copyright.)”²⁷

Such strictures meant that any book digitization (by anyone other than the copyright owner) could be done only with permission or would need to be confined to public domain materials. Project Gutenberg, which some consider not

²⁷ Appellates’ Amended Opening Brief, *Kahle v. Gonzales*, 474 F.3d 665 (9th Cir. 2007. January 31, 2005.

only the first e-books but also the first “mass digitization” project, only digitized public domain books. Libraries as well had largely confined their digitization efforts to what such understandings of the law allowed: that is, they only digitized public domain books or those books that fit within the special allowances of Section 108 of the 1976 Copyright Act. (Section 108 allows libraries to make one copy of a book, with notice, if one cannot be found to buy at a fair price, if the copy has no commercial purpose, and if the library provides unrestricted access to the public.) Focused projects, such as the Making of America project—an early and important collaborative digitization project between the University of Michigan and Cornell University libraries (and supported by the Andrew W. Mellon Foundation) focused safely on books in the public domain. Just as the film archivists carved out their domain as “orphan films,” libraries chose to confine efforts to the pre-1923 era. This limitation dovetailed with the urgent concern for “brittle books,” which I discussed in chapter 1, such that libraries attracted federal dollars to digitize “endangered” nineteenth-century (public domain) books. At the snail’s pace they were able to move, they had plenty to keep them busy in the centuries before the twentieth.

The Million Book Project had painstakingly pursued permission for the books it sought to digitize. Although the Project had planned from the outset to digitize *primarily* public domain books, the planners also targeted 10 percent of their total goal to be works in copyright (Covey 2005a). To maximize the number of books they could digitize, the American team chose to focus their efforts not on individual titles but on publishers, writing and asking permission to digitize as many of their *out-of-print* books as the publisher was willing to allow. Between August 2003 and March 2005, a team led by Carnegie Mellon librarian Denise Covey attempted to contact copyright owners to secure their permission to digitize these works and to make them available “open access” online (Covey 2005a). During that time, although they were able to identify and locate all 365 publishers of the books, they received responses from only about half of them; of that half, about 23 percent of them granted permission. After a conservative cost estimate of \$36,501 and a year and a half, they had permission to digitize just 50,000 books (Covey 2005a, 55). Although this result is no small accomplishment, it indicates that, no matter the effort and expense, it would prove impossible to secure permission to digitize an entire library.

Covey’s detailed report shows that the problem is not simply the “burdens” of copyright law but as significantly what she calls the “nineteenth-century record-keeping methods” of publishers. Even if they had wanted to grant permission, some publishers themselves often could not easily determine whether out-of-print books had had their rights reverted to the author or not. Answering those questions required that they search through each file, one by one, to find a contract or a letter indicating whether the rights had reverted to the author. Some told her that, although they wanted to help the Project, they didn’t have the time or the staffing

to research the rights issues. Others knew that the rights to their books had automatically reverted to the author when a book went out of print, in which case Covey and her staff had to start over in pursuit of permission from the author. A further complication was that publishers could not know for sure whether they even had the electronic rights to give. Indeed, the issue of who owns electronics rights to a book published before the 1990s—author or publisher?—was and remains so volatile that neither publishers nor authors want to push it for fear of coming out on the wrong end. The Million Book Project’s well-documented experience allows us to see how complex a business rights clearance is for all sides: the effort to *give* permission has high “transaction costs” as well.

What was a mass digitizer to do? The transaction costs of securing permission would likely amount to an even greater sum than Michael Lesk’s \$3 billion projection at the beginning of this chapter.²⁸ In addition, it would also take a very long time. Although it was evident that most of the older (pre-1964) books were likely to be in the public domain, it was just as evident that there was no easy way to know *which* were in definitively in the public domain and which not. Even if the expense and time were manageable factors, permission-seeking would not advance a mass digitization project significantly, as permission would only be found for a fraction of the books. Waiting for copyrights to expire wouldn’t be wise, as Congress seems intent on making them perpetual in effect. Confining themselves to pre-1923 books would never create a universal digital library. Mass digitizers didn’t want only public domain books—themselves only and 20 percent of what is on research library shelves. They wanted *all the books*. They had grown impatient with publishers, who appeared disinclined to make books available electronically despite the constant anticipation of an “e-book” revolution throughout the 1990s. Publishers also had no incentive to digitize their own backlists of out of print books; they were “out-of-print for a reason,” that is, they had outlived their commercial viability. Finally, digitizers wanted to automate as many of their processes as possible, so no painstaking labor-intensive rights clearance would be tolerable. Give these general dispositions and circumstances, the options were: 1) to do nothing; 2) to reform or change the laws; 3) to do the previously unthinkable: to digitize without permission.

As early as 2001, Kahle was contemplating a large-scale digitization strategy (Kahle et al. 2001; Hardy 2009). To overcome the orphan problem, his strategy is to digitize them and make them available through loaning. Loaning refers to a system whereby a library patron could have access to a digitized book subject to various restrictions analogous to print circulation (e.g., only one person can check out a book at a time, for a limited period of time; etc.). It is a compromise that, by mimicking a familiar and established practice, Kahle hopes will not threaten

²⁸ Band (2006, 8) throws out the figure of \$25 billion (\$1,000 per book x 25,000,000 books) without indicating where he got the \$1,000/book figure.

corporate legal departments. When, and if, an “orphan” copyright expires, then greater access could be given, but in the meantime they would at least have been digitized and be available to readers. Unlike Google, whom we’ll see “lawyered up” with a full-blown legal strategy, Kahle purports to merely rely on tradition. Libraries have always loaned material: “Since the roles and responsibilities of libraries for providing public access have not changed simply because the format of material has moved from analog to digital, it is reasonable to assume that the rights to perform these societal tasks have not significantly changed either” (Kahle et al. 2001). Why, he asks, should anything be different? “Where printed objects must be lent out physically, the digitized and digital library material can be lent out digitally. Using this model, libraries can continue to serve the public good in the future without a major change to the institutional structure as we digitize the collections” (Kahle et al. 2001). It would take Kahle nearly a decade to implement a loaning system, but he has stubbornly maintained his loaning idea as the best path forward (Fowler 2010). “If you own it, you can loan it,” is his catchy way of summing up his strategy.

Google chose a different path forward. First, following the lead of Amazon, who started a mass digitization effort on its website in 2003 (called “Search Inside”), Google created a program with publishers whereby they licensed the right to digitize, index, and display portions of in-print books through books.google.com. They shared some advertising revenue with the publisher—usually a pittance, as ads next to book search results have not proved lucrative—but the main benefit to them was exposure for their products. In May 2007, Google folded the books into main “universal” engine at google.com. The site provided links both for buying a book online or finding it in a library. This program, called Google Print, was announced in October 2004. It addressed only the in-print books, the easiest to deal with because they have an obvious rightsholder—the publisher who is offering it for sale.

But its much grander project was the Library Project, since it would address the remaining 90 percent of books. No digital library of “all the world’s books” could be made simply through partnering with publishers. For that, any mass digitizer would need the cooperation and collaboration of large libraries that possess millions of books conveniently cataloged and housed together. For two years or more before the project was made public in 2004, Google had been discussing the Library Project with the University of Michigan and Stanford University. Among the first libraries to join as partners—called the “Google Five”—Harvard, Oxford, the New York Public Library, and Stanford agreed to provide Google with only public domain books, whereas the University of Michigan agreed to allow them to digitize every book in their collection (or, at least, all that would fit on their scanning equipment). Michigan had the advantage of being a public university and, as such, the Constitution’s Eleventh Amendment grants them “sovereign immunity” from monetary damages should they be found guilty of copyright infringement. It also

was among the most expert of university libraries in terms of digitization and digital publication, as it had been an originator or major partner in JSTOR, the Text Encoding Initiative, the TULIP journal publishing platform, the Making of America project, and more. Much of Google Books' early history with Michigan is hard to discover due to non-disclosure agreements, but when Google announced its "Google Library" project in December 2004, it made its digitization strategy clear: it would digitize copyrighted books without permission and the public domain without the need for permission. This way, they could digitize "efficiently" and make the task of digitizing tens of millions of books, which had previously been thought practically impossible, possible.

The issue of copyright would be dealt with, after digitization, when determining how much of a book to display to the public. The copyright status of a book would be determined based on what type of book it was, and when and where it was published. Books determined to be in the public domain were made available in "full view" and books determined to be in copyright, even if just presumptively so, would only be made available through what Google calls "snippet view." Snippet view means that the only part of a book's text that can be seen by a reader is the area immediately around a keyword. Although the presentation of the Google Book Search page has changed over time, today a snippet appears thus:

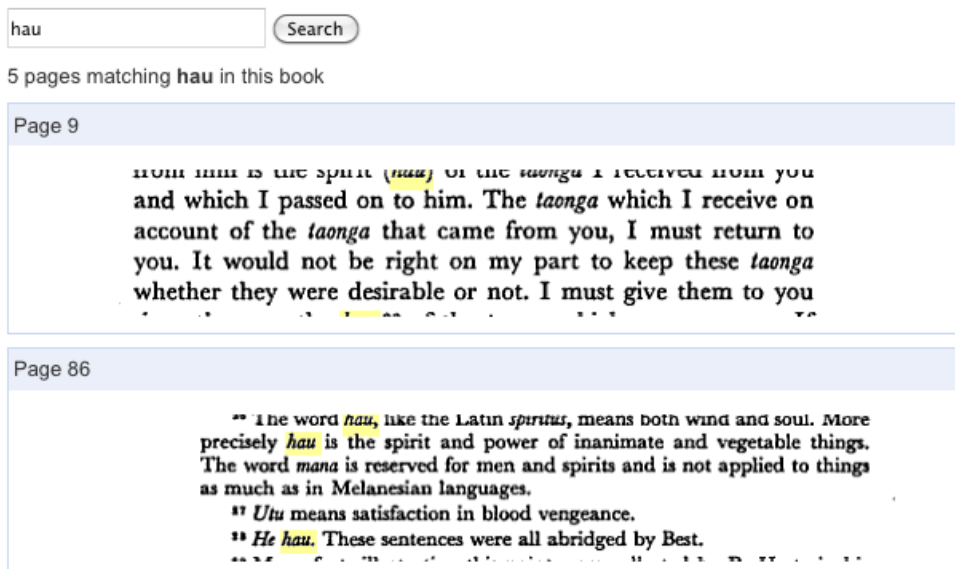


Figure 1: A snippet from the 1954 English translation of Marcel Mauss's *The Gift*. The search keyword was "hau."

Google's strategy, by eliminating the request for permission, also eliminated the orphan problem. Orphans, technically, are only a problem if you need to find the owners. The trade-off, of course, is that for (roughly) 70 percent of the books,

access is severely limited. It is for this reason that many have described Google Book Search—at least the Library Project part of it—as an index rather than as a book collection or a library (e.g., Lessig 2006). It is an index to millions of books, with differential access to those books based on copyright status.

The company defends its actions as a “fair use” under Section 107 of the Copyright Act. Fair use operates as an exception to the rights of copyright owners so as to allow for unauthorized but socially beneficial uses of copyrighted works. Fair use is “fact specific”; it is an anticipated *defense*, meaning that whether the use is actually “fair” can only be decided in retrospect, by a judge, taking into consideration the “four factors” specified in Section 107 of the Copyright Act as well as other circumstances specific to the individual case. The four factors are 1) the purpose and character of the use; 2) the nature of the copyrighted work; 3) the amount and substantiality of the portion taken; 4) and the effect of the use upon the potential market. Google sees its use as fair because it is only displaying a tiny portion of the work (see figure 1 above); its use is “transformative” (it is a searchable index of books); and its use does no harm to the potential market because there is, by definition, no existing market for the out-of-print/orphan books. Many legal analysts have assessed the validity of Google’s fair use argument (Hanratty 2005; Lessig 2006; Band 2009; Sag 2010; Samuelson 2011a), and, although opinions vary, there is certainly a strong case to be made for Google’s use being found fair. Rehearsing the strengths and weaknesses of Google’s fair use arguments, however, is not germane to my argument here. What is germane, however, is the importance of “books as orphans” within the dispute.

With the orphan metaphor, digitizers (and not just Google) have made a distinct problem actionable. It is impossible to ask the phantom copyright owners for permission and thus also impossible to receive a license that would authorize their digital copying. As such, the orphans represent a pure “market failure” in that one half of the necessary market transaction is not present. When new copying techniques are deemed to be socially beneficial and yet their use cannot be managed through a market mechanism such as licensing, fair use has tended to prevail. In the famous *Betamax* case, the Supreme Court agreed that it was a fair use for individuals to make what Hollywood had characterized as “unauthorized” copies (*Sony Corp. v. Universal City Studio*, 464 U.S. 417 (1984)). Legal scholar Wendy Gordon has termed this judicial tendency “fair use as market failure,” and her theory suggests that the higher the transaction costs involved with a new use the stronger a fair use defense will be (Gordon 1982). In short, the “orphan” problem provides the ground for legally justifying the mass digitization of books.

By using a fair use justification for mass digitization, the *display* of the work to the public becomes the copyright-significant act, not the initial copying/digitizing of the books. It shifts the concern for infringement away from the copyright owner’s “exclusive right” to reproduction (as explicitly stated in Section 106 of the

Copyright Act) to the specific uses to which the copied work is put. From Google's perspective, the significant facts are how the books appear to the public. This attempt to move away from a "maximalist" understanding of the reproduction right as the *sine qua non* of copyright is fundamental to Internet activists, digital librarians, and digitizers. In this approach, they follow a line of copyright interpretation advanced before the Internet, perhaps most elegantly articulated by the legal historian and scholar L. Ray Patterson (Patterson and Lindberg 1991). With regard to books (or "literary works"), Section 106 of the U.S. Copyright Act enumerates five exclusive rights: to copy, to adapt (or "to prepare derivative works"), to distribute, to perform publicly, and to display publicly. Patterson argues that the right to copy is here not a independent right, as the other four rights are, but a dependent right, by which he means that it is given only so as to enable the exercise of the other four rights: "The reproduction right embodies the right to copy (duplicate) for public distribution, public performance, or public display—and no more" (Patterson and Lindberg 1991, 153). He goes on to argue that the pervasive belief that a copyright owner has an absolute right to reproduce a copyrighted work is fallacious: "Copyright was intended as an economic incentive to create and distribute works for the purpose of promoting learning. ... Copyright was never intended to give the copyright owner an arbitrary power to control and individual's use of the copyrighted work—and it would be most damaging to the basic copyright policies and principles if copyright ever becomes so corrupted" (Patterson and Lindberg 1991, 159).

Patterson argued that copyright needed to serve "authors, entrepreneurs, and users" equally but that, as the law developed after 1976, it had come to overemphasize the rights of the copyright industries (the "entrepreneurs") to the detriment of the creators and users.²⁹ Patterson's work has significantly influenced the "copyleft" activists that I have identified throughout this chapter as in intellectual alignment with mass digitization, especially his articulation of the need for a shift in copyright law from the rights of the "monopolists" to the rights of the users (i.e., the public), specifically their right "to learn" (cf. Lessig 2001, Litman 2001). In her book *Digital Copyright*—which Brewster Kahle refers to as "required reading" at the Internet Archive—Jessica Litman analyzes the ill fit between current copyright law and the uses people now make of digital technologies (Litman 2001). Litman calls for a copyright reconceived with the "ordinary end user" in mind rather than the industry players around whose interests the current rules have been crafted. Her ultimate recommendation is that the copyright owner's current exclusive right of *reproduction* (i.e., to copy) be replaced with an exclusive right of *commercial exploitation*. In this rearrangement, making money off someone's copyrighted work would be an infringement, whereas merely copying it would not. Patterson,

²⁹ Although U.S. law does not recognize the "moral rights" of authors as European law does, Patterson believed that authors were entitled to rights "by reason of their creation of new work" but that these should be treated through laws other than copyright (Patterson and Lindberg 1991, 122).

however, would have argued the point more aggressively, that, in a proper reading of copyright law, *there never has been* an exclusive right of reproduction.

Orphans Go Mainstream

In January 2005, around the same time that Brewster Kahle, Rick Prelinger and their lawyers were submitting the appellate brief in their orphan works lawsuit, *Kahle vs. Gonzales*, the Copyright Office undertook an official “inquiry” into the orphan works problem. Pressure had been building on the Copyright Office to do something since the passing of the Sonny Bono act and the final outcome of *Eldred vs. Ashcroft* in 2003. The Google Library Project announcement, which occurred one month prior to the Notice of Inquiry, may have been a factor as well. At any rate, in January 2005 the Copyright Office requested opinions from any parties interested in the “issues raised by ‘orphan works,’” which, again, the Office defined as “copyrighted works whose owners are difficult or even impossible to locate.”³⁰ The Copyright Office sought a global solution to *all* types of orphan works and not just books. First it invited “initial comments” and then replies to those comments. By May it had collected over 850 written comments from a wide range of organizations and individuals who have an interest in either protecting copyright interests and/or enhancing the ability to access or reuse copyrighted material: artists, photographers, documentary filmmakers, film producers, Hollywood (MPAA) and other industry trade groups (AAP, Authors Guild), consumers groups, advocacy groups (e.g., Public Knowledge, Creative Commons, Electronic Frontier Foundation), research libraries, academic associations (AHA, CAA), archives of every stripe, museums, law professors, big tech companies (Google, Microsoft), and, of course, the Internet Archive.³¹ On the basis of these comments as well as two roundtable discussions held in Washington and in Berkeley,³² the Copyright Office compiled and published a report in January 2006, which became the beginning point of a process to draft legislation that would solve the “orphan works” problem.³³ The Copyright Office report made two fundamental recommendations. First, in exchange for use of an orphan work, the user needed to have both conducted and documented a “reasonably diligent search.” Second, if the owner of

³⁰ *Federal Register*, Vol. 70, No. 16, Wednesday, January 26, 2005, p. 3739.

³¹ The comments and replies are tallied, grouped, and analyzed in Covey 2005b.

³² Brewster Kahle represented the Internet Archive at the Roundtable in Berkeley on August 5, 2005.

³³ The Report, comments, replies, and text of proposed legislation are all available on the Copyright Office’s orphan works website: www.copyright.gov/orphan.

the copyright showed up after the orphan work had been used, the existing and quite harsh remedies called for in copyright law would be nearly eliminated.

Legislation based on the Report, known as the Shawn Bentley Act, nearly became law in 2008. It was passed in the Senate but ultimately died in the House as consensus support started to fall apart, especially under continued resistance to the legislation from photographers. The act provided for a limit on the fines (known as “remedies”) that a judge could levy for the copyright infringement if the user of that work had met certain requirements, most significantly a reasonably diligent search in good faith to locate and identify the copyright owner before using it. Although the bill failed, it got very close to passage and various observers believe it could be successfully revived. From the perspective of mass book digitization, however, the bill, even if it had passed, would have been of no help because it required a “diligent search” for each copyright owner. Such a solution might work for individual cases involving one or several works, but certainly not millions of works. What a mass digitizer wants is an “opt-out” solution: the right to digitize orphans en masse and then, should someone object, they will deal with the objection, say, by taking the book down. What mass digitizers want is a way around copyright as it is currently understood and practiced.

The Google Book Search Settlement

Shortly after the Senate passed the Shawn Bentley Act in September 2008, but before it had died in the House, Google, the Authors Guild, and the Association of American Publishers (AAP) announced what became known as the Google Book Search Settlement. The Settlement was a proposed resolution of the two lawsuits filed against Google’s book project in 2005, one filed by the Authors Guild and the other by the AAP. Secretly negotiated for over two years, the Settlement was an intricate blueprint for the creation of the absent market that had created the market failure that is orphan books, as described above. The Settlement quickly became christened as having “cut the Gordian knot” of orphan *books*, at least, and perhaps providing a model that might work for other types of copyrighted works.

The case drew special power through the peculiarities of U.S. law, namely the class action procedure. U.S. law allows group litigation where a small number of plaintiffs claim to represent a class of individuals who have suffered a similar harm from the defendant’s alleged wrongful conduct. In the settlement of *Authors Guild et al. v. Google*, the Authors Guild and the AAP presumed to represent all rightsholders who owned a U.S. copyright interest. Although the jurisdiction of the case was limited to the U.S., one can own a U.S. copyright interest without residing or even doing business in the U.S. As long as an author or publisher published a

book in a country that has “copyright relations” with the U.S.³⁴—that is, a country that is party to the Berne Convention)—then the Authors Guild and the AAP came to represent that individual or corporation in the Settlement negotiations. Thus, the parties to the Settlement hoped that the class action procedure—and what legal scholar Pamela Samuelson termed its “legal jujitsu”—would enable Google a license capacious enough to cover as many books as possible that Google will find to digitize. (To date, Google claims to have digitized at least 20 million books in 478 languages.)

The Settlement, it was argued, would solve the orphan book problem by providing an incentive—cash—for rightsholders to come forward and stake a claim in their books, something they would otherwise not be inclined to do. Although the company admitted to no wrongdoing, Google agreed to pay \$125 million to the authors and publishers in return for permission (i.e., a license) to develop a commercial database from their digitized library books that were not otherwise commercially available. Google would sell subscriptions to the database to universities, libraries, and other institutions.³⁵ A portion of the revenue generated from the digitized book corpus (67 percent) would be turned over to a collecting agency called the Books Rights Registry, which would in turn be responsible for distributing the revenues to the authors and publishers who had signed up with the Registry. By registering with the Registry, the rightsholders would be eligible to receive money for the past digitization of their book(s), a portion of ongoing revenues, and payment for the inclusion of their book(s) in the database. The parties to the Settlement argued that having a “pot of money” would bring forward the formerly unfindable owners and that ultimately the orphan problem would be shown to be either non-existent or very small indeed.³⁶ Indeed, the Settlement sought to solve the orphan problem by eliminating orphans altogether. Indeed, the term is not used in the Settlement at all, despite its ghostly presence.³⁷

The AAP and Authors Guild insist that there really is no orphan *book* problem because books, especially when compared to photographs, contain a lot of copyright-relevant information on or inside them, and they argue that most authors

³⁴ For what it’s worth, the U.S. does not have copyright relations with Afghanistan, Bhutan, Eritrea, Ethiopia, Iran, Iraq, Nepal, and San Marino.

³⁵ There were other projected “revenue models” but the institutional subscription database was understood to be the main source of revenue.

³⁶ Revenue generated from books that remained orphaned (or, in the language of the Settlement, “unclaimed”) would go into an escrow account. If unclaimed after a period of time, the money would be given to a relevant charity.

³⁷ There is one small reference to the possibility of Orphan Works legislation.

are quite findable. However, the Authors Guild, AAP, and Google were seeking to solve the same problem as Google: they wanted to be able to hand Google a total license, satisfying Google's desire to digitize "everything." The more comprehensive the database, the more valuable a product they would have to sell. But they knew also knew that all copyright owners would not show up, or at least that it would take a long time for most of them to do so. They would never be available to engage in one-to-one "market" transactions with Google. For these, they chose a different term. In place of orphans, the Settlement speaks of "unclaimed works." With "unclaimed" they opted for vocabulary analogous to "orphan" in that it acknowledges the same problem but solves it more narrowly and in their own self-interests.³⁸

The Settlement envisioned, in essence, a new regulatory system for books—a private one, since it was governed by a contract and not by copyright law—that reinstates a formal system like that in force before the 1976 Copyright Act. It would have created formalities. Rightsholders would have to affirmatively come forward and "opt-out" or otherwise they were subject to the terms of the Agreement. In every other circumstance, the Authors Guild and the publishers oppose any newly instated formalities because a formal system places burdens on copyright owners. The net effect of the Authors Guild and the AAP bargaining was to force their constituency—that is, all owners of any portion of in-copyright books—to re-accept the burden of formalities in order to: 1) create a new market that Google (and only Google) could exploit; 2) create a competitor for their nemesis Amazon; and 3) domesticate the Internet to "traditional" tenets of copyright: requiring permission, compensation, and control over how their work would be used. As one observer put it to me privately, "With the Settlement, the authors and publishers basically bargained for a piece of Google." In turn, Google had little to lose from the acceptance of this compromise because it would have secured some legal cover and a green light to digitize, which is what they had most needed.

³⁸ Unfortunately for the Settlement's prospects, it solved the "legal fog" of digitization for Google alone. As the Internet Archive's lawyer Hadrian Katz argued to the judge in the case, the Settlement gave Google "a right, which no one else in the world would have, ... to digitize works with impunity, without any risk of statutory liability" (Testimony, Fairness Hearing Transcript, page 95). Everyone else remained in the same position as before. It conferred on Google an enormous privilege, something like a copyright exception but for just one company. Google had by this time digitized many millions of books and the Settlement only cemented their advantage, leaving them to be a sole actor in book digitization. They, and only they, had a license to digitize books without limit. The point was made similarly in the Internet Archive's formal submission to the court in opposition to the Settlement: "Google would have the right to make complete copies of orphan works and use them for both display and non-display purposes, with no risk of copyright liability. Competitors that attempted to do the same thing, however, would [continue to] face exposure to statutory damages." Legal filings and other documents related to *Authors Guild v. Google* can be found at www.thepublicindex.org.

However, it was on behalf of the “orphans” or the “unclaimed” that Judge Chin rejected the Settlement. Chin was sympathetic to the argument that the Settlement would allow Google to “expropriate” rights of copyright owners without their consent. He wrote: “The notion that a court-approved settlement agreement can release the copyright interests of individual rights owners who have not voluntarily consented to transfer is a troubling one.” He also suggested that a move from an opt-in system to an opt-out was hard to justify: “Here class members would be giving up certain property rights in their creative works, and they would be deemed—by their silence—to have granted to Google a license to future use of their copyrighted works.” His final recommendation was that the Settlement be changed so that it would apply only to those books for which an owner had come forward to claim it. In other words, he suggested that, for a court to be able to approve the Settlement, the parties would need to change it from an opt-out to an opt-in and thus make it conform to current copyright. That meant that they would have to remove from the Settlement the mechanism that would have roped in the orphans/unclaimed, thus solving the problem of “orphaned” books. Needless to say, such an option is unacceptable to Google as it would foreclose the totality that the mass digitizer requires. Mass digitization requires a solution that “will scale.” The crux of the Settlement came down to the absent presence of the orphans. As Judge Chin ruled, there could be no deal with the orphans, and yet for Google there would be no deal without them.

Failed Adoption?

The orphanistas’ early call to reform film archives by “saving the orphans” and, later, Rick Prelinger, Brewster Kahle, and Larry Lessig’s efforts to reform copyright by “saving” even more orphans helped paved the road to the two foregoing notable proposed solutions to the problem of orphan works: one public (Orphan Works legislation) and one private (the Google Book Search settlement). Both efforts failed and some have asked if it is time to hang up the metaphor. William Patry has complained that the term is inapt and its use is “greatly inhibiting resolution of this critical problem” (Patry 2009, 77). It is inapt because it draws upon the mistaken “natural law” theory of copyright, which holds that certain rights spring naturally from the creative genius of an author, thus granting “moral rights” in her work. Although European copyright law supports moral rights, U.S. copyright, definitively, does not. The orphan work, Patry argues, supports the natural law line of thinking by suggesting a work is the “child” of its author who “naturally” needs to be the person taking care of it (Patry 2009, 75ff). To Patry, the invocation of a parent-child relationship is a moral language misapplied to an economic right and its use can only justify more rent seeking. He prefers the description “willfully abandoned works” since its use would elicit less sympathy for the parent who has willfully abandoned not only his or her “child” but also the special rights granted as a parent. To Patry, the only solution in fact is to restore formalities and reduce the term of copyright (Patry 2012).

However, since the orphan metaphor emerged not from copyright maximalists, as Patry implies, but from copyleft activists—not so different from Patry himself—his critique seems to be misdirected and somewhat off target. The point in choosing and promoting the “orphan” metaphor was not to promote the reuniting of a work with its “natural” parent, though he is correct in pointing out that this was the effect of Lessig, Prelinger, and Kahle’s efforts. The new digital archivists had wanted to “rescue” the orphans from copyright, for creative reuse, and for future generations, but not for a new kinship relation. But Patry is write to criticize it if in retrospect the “orphan metaphor” worked only to inflame the fears of copyright owners to whom the lawful sanctioning of “unauthorized use” is deeply threatening.

Another copyright scholar, Lydia Loren, has recently followed Patry’s lead in criticizing the orphan metaphor further (Loren 2012). Loren argues that, in its Dickensian invocation of mistreated street children, the orphan metaphor has created a “narrative of the potential abuse” and that this has impeded the passage of orphan works legislation (see, e.g., Holland 2010). In its place, she suggests a new metaphor: “hostage works.” Hostage works are works that are wrongly ensnared, trapped, and imprisoned by copyright. Continuing a military connotation, Loren advocates granting limited immunity for entities that act as “special forces” to free them. In return, these special forces would be obligated to provide public access to a digital copy of the work that adheres to the basics of open access principles. When Loren presented this at an orphan works symposium in April 2012 in Berkeley, the talk was enthusiastically received. Kahle told me that she was right: they’d been using the wrong term. Hindsight is 20/20³⁹. In retrospect, the orphan metaphor, which worked well for activist film archivists, did not work so well for Internet copyright activists. Books, at least, didn’t need the sort of rescue that ephemeral films did.

If Brewster Kahle and Rick Prelinger see themselves as activists, they see themselves as cultural activists not legal activists. Indeed, despite his thorough immersion in strategies to deal with copyright (and its failures), Kahle regularly claims to know little about copyright law. If asked to explain how he understands the legal underpinnings of some of the Archive’s activities, he pleads ignorance. For example, at a 2012 conference sponsored by the Berkeley Center for Law and Technology, a member of the audience (made up largely of lawyers) asked Kahle what laws govern the Archive’s digitization efforts in countries other than the United States.⁴⁰ He responded, laughing: “Is there a lawyer in the house? ... I don’t

³⁹ Nevertheless, it is hard to see how a metaphor so aggressively expressed would be any more likely to succeed.

⁴⁰ The Archive has scanning centers in five countries outside the U.S.

know. We just try to do the right thing. I'm not trying to plead ignorance. I am ignorant" (Kahle 2012).

The wisdom of his, shall we say, "unsophisticated" approach is borne out by the success of the Archive in avoiding copyright lawsuits, against the expectation of many legal observers. In 2001 the Archive debuted the Wayback Machine, a tool for accessing the websites it had been amassing since 1996. To a copyright maximalist, the Wayback Machine perpetrates massive copyright infringement because it makes and stores copies of (copyrighted) websites and displays them to the public without the authorization of the copyright holder. As one scholar wrote to the *Chronicle Higher Education* after it ran an article about the Web archive: "The Internet Archive is nothing more than an enormous copyright violation disguised as a library."⁴¹ And yet fourteen years later the Wayback Machine has become a highly valued research tool. In his analysis of informal web standards that eliminate conflict, Jonathan Zittrain has written: "Consider, for example, the Internet Archive. Proprietor Brewster Kahle has thus far avoided what one would think to be an inevitable copyright lawsuit as he archives and makes available historical snapshots of the Web. He has avoided such lawsuits by respecting Web owners' wishes to be excluded as soon as he is notified" (see Zittrain 2008, 322-33n125). Rather than defend itself in legal terms, the Archive employs an informal "exclusion policy" that honors the request of any webmaster to have a site removed, either by direct request to the Archive or by the placement of a "robots.txt file" on the site's Web server, which will signal to the Archive's crawlers that the site should not be copied.⁴² This "opt out" approach is what mass digitizers seek for books, too.

Kahle's pragmatic approach is based on the experience of doing things and seeing what happens: "We've learned a major lesson. Aside from all the laws, if people are pissed, they are going to try to find some way to stop you. But if they are not, they will let it go forward. The key thing is to do things on that side of the line... Things can work if you're receptive and respectful" (Kahle 2012). At other times, he has referred to his pragmatic approach as trying not to do things that "smell bad" and "not upsetting the apple cart too much."

Another example will return us to the "orphaned book." In 2010, the Archive started digitizing books in copyright. In June they announced a program not to *display* the books on the Web but to *loan* them via the Web. Only one reader can

⁴¹ Stephen R. Brown, "Is On-Line Archive Fair Use?" Letter to the Editor. *Chronicle of Higher Education*. May 1, 1998.

⁴² See the Archive's FAQ page: <http://archive.org/about/faqs.php>. For more information on robots.txt as a "code-backed norm" that enabled "harmony" among Web users and search engines without any application of law, see Zittrain 2008, 223-24.

have a book at a time—just as with print lending—and after two weeks the loan expires (the program uses Adobe Digital Editions, a digital rights management program trusted by publishers). When Kahle met with library leaders in advance of announcing the Archive’s loaning program, he was told by at least one lawyer who advises libraries that, by digitizing in-copyright works he has risking—even asking for—a lawsuit. The lawyer found this an exciting prospect; Kahle did not. After the announcement appeared in the national press, various commentators had a similar reaction. One blogger described the program as “a move that seems designed to spur a legal reaction from publishers” (Hellman 2010). One legal analyst inferred that Kahle’s legal defense would have to be based on fair use and that, if so, it would be a very weak case (Grimmelman 2010). Yet another legal specialist saw the program as legally indefensible but nevertheless admirable: “There seems to be little legal basis for what the [Archive] is doing. But unless they start lending books owned by an author or publisher who objects, there is no one to bring legal action. And if no one objects, then making material as easily available as possible has lots of benefits. ...[Y]ou have to admire Kahle’s willingness to act as if copyright were rational and respectful of the interests of both authors and users, rather than what we have now” (Hirtle 2010).

Now over two years out, the Archive has *not* been sued for its digitizing and loaning out of in-copyright books, as Google was. One might argue that this lack of litigation is because the Archive is a small nonprofit without Google’s deep pockets. But, although it is true that commercial actors are more attractive targets, money is not always the key factor in copyright litigation. In 2011, for example, the Authors’ Guild filed an infringement suit against five universities whose libraries have allowed Google to digitize books from their collections and who have received in return digital copies of their books. The Authors Guild, however, is not seeking monetary damages but rather an injunction to stop the libraries’ cooperation with Google and the “impoundment” of all the libraries’ digitized books.⁴³

In response to questions about his strategy, Kahle simply explains: “We’re just trying to do what libraries have always done” (Fowler 2010). At other times, he summarizes his thinking this way: “If you own it, you can loan it.” This statement invokes the first sale doctrine from Section of the 1976 Copyright Act, yet another limitation on copyright, which holds that the distribution rights of a copyright holder end when a copy is lawfully transferred. This doctrine undergirds a library’s ability to lend books out and for readers to sell their books to a used bookstore and for a used bookstore to sell them without having to pay royalties to the copyright owner. Since most electronic books are distributed not as alienable commodities but as license agreements (or rentals), one cannot resell an e-book, and a library

⁴³ “Complaint,” *The Authors Guild et al. v. Hathi Trust et al.* New York Southern District Court 1:2011cv06351, filed September 12, 2011.

can only lend an e-book according to the terms of the license agreement. With regard to a digitized out-of-print book, however, what rules should obtain? To Kahle, it's simple. If a library could do it before in print, then a library should be able to do it now, in digitized form. He has attracted one thousand libraries to donate books and to enable the lending from their catalogs. He has also created a large physical archive of the books that have been digitized for the program, not for circulation but for long-term storage (if you loan them, you must own them).

In his recent reflections on the sobering lessons of his experience as a legal reformer, Kahle evinces a new desire to depend not on legal innovations or top-down solutions but case-by-case community development. Any global solutions, he fears, are likely to make the situation worse, as shown by the narrowing of the definition of "orphan" to a definition that promised only minor and nearly inconsequential change—and even it could not pass.

I was among the first to use the term orphan works, and I was naïve about orphan works. I thought it would be a slam-dunk to go and try to get the courts to say yes. Then it went to Congress. I thought this would be easy. Who would complain? Wow, are people who will complain! ... So this has been a real turnabout for me. I thought we should solve things through these mechanisms but at this point I think we should slow down a bit and let some things evolve... We have to make sure that we can play our traditional roles but in a newly digital world. ... We should continue to do our jobs as cultural institutions: collect, preserve, provide access, and purchase things. If there are things that we need to keep out of a library, then we have lots of experience to deal with that, but that doesn't happen that often. We can handle those things specially (Kahle 2012).

Naïve perhaps but also slyly strategic, Kahle assumes the public role of a humble "digital" librarian who by his actions is asking not for legal combat but for someone to tell him how copyright could ever be a reason to keep a book out of a library.

Conclusion

Ray Bradbury, *Fahrenheit 451*
Vincent Canfora, *The Vanished Library*
Abbie Hoffman, *Soon to be a Major Motion Picture*
Lawrence Lessig, *The Future of Ideas*
David Graeber, *Debt*

-- Books on Brewster Kahle's Desk (many in multiple copies)

As I conclude this dissertation, nearly four years after I began work at the Internet Archive and eight years after Google announced its Library Project, the terrain of mass book digitization has, perhaps unsurprisingly, shifted, as new responses to it continue to emerge. Google and the book publishing industry tried to resolve their differences, in the form of the Google Book Search Settlement, but the authors and publishers were unable to convince a Federal judge that their plan was in the interests of all those they claimed to represent. In response to that ruling, the publishers are trying to come to terms outside the courtroom, and the Authors Guild has resumed its suit against Google. In pursuit of summary judgment in that case, Google has just made its fullest statement yet defending its actions as fair use.¹ In counterpoint, the Authors Guild wants \$750 for each digitized book, which would come to a few billion dollars depending on the number of books that would qualify. In yet a different courtroom, the Authors Guild has initiated a further copyright infringement lawsuit, *Authors Guild v. Hathi Trust*, against five public universities who partnered with Google. They seek to have the libraries' digitized books impounded and all in-copyright book digitization enjoined.²

Many of the central negotiators of the Google Book Search Settlement have moved on: in 2009, the product counsel for Google Book Search and chief architect of the Settlement, Alexander McGillivray, left to become the general counsel for Twitter; in 2010 Daniel Clancy moved to YouTube; in 2011, Richard Sarnoff, the chief publisher negotiator, left Bertelsmann for a private equity firm. Settlement post-mortems are being written (Crawford 2012; Grimmelman 2012). Reports circulate that Google is scaling back its investment in book digitization (Howard 2012). Jon Orwant, an engineer in Google's Boston office and frequent spokesman for Google Book Search, was reassigned within Google to digital humanities, patent search,

¹ See "Defendant Google Inc.'s Memorandum of Points and Authorities in Support of Its Motion for Summary Judgment or in the Alternative Summary Adjudication." Document 1032. July 27, 2012. <http://dockets.justia.com/docket/new-york/nysdce/1:2005cv08136/273913/>.

² The five universities are: the University of Michigan, University of California, University of Wisconsin, Indiana University, and Cornell University.

and data visualization. Jimi Crawford, who had replaced Clancy as engineering director of Google Books, has just left Google to join Moon Express, a privately held “lunar transportation and data services company.” A close observer told me: “No one at Google now is going to make a career by working on Google Books.” Its moment in the sun has passed. Even though the company continues to defend the book project vigorously in court, it is not difficult to imagine the company coming to regret the endeavor. It cost them a great deal of money—hundreds of millions of dollars—and has made them very little in return; it was the first but not the last federal antitrust investigation into the company; it has mired them in expensive and possibly perilous litigation; and that litigation, even the attempt to resolve it, branded Google as “a bad guy” to many and gave the company a black eye in the press.

The Open Content Alliance, without fanfare, slowly evaporated after the Settlement was announced in 2008, though the Internet Archive and continues to digitize books with many partner libraries. Still, the Archive, after years of engineering and other forms of investment in a variety of book-related projects—the OCA, Open Library, opposition to the Google Book Search Settlement, the Book Server, Books in Browser conferences, its long-term “deep storage” physical archive for books, and book loaning—is pulling back from books so as to expand its investment in other digital library horizons, such as its television archive.

Momentum has shifted to the Digital Public Library of America (DPLA), a loosely defined effort to build an alternative to Google Book Search that, like the OCA, would be properly “in the public interest” (Darnton 2010, 2011a; Yi 2012). The DPLA has received \$3.5 million from the Sloan Foundation, \$2.5 million from the Arcadia Fund, and the \$1 million from the National Endowment for the Humanities. When I attended the DPLA West meeting in April 2012, which was held at the Internet Archive,³ I could clearly discern a direct through line, starting from Brewster Kahle’s phone call to the Sloan Foundation’s Jesse Ausubel in 2005—as recounted in the Introduction—to this “big tent” gathering on behalf of the DPLA, with Robert Darnton replacing Brewster Kahle as the motivating personage.

This shift of center from an “Internet library” to the nation’s richest private university, and from a West Coast Internet entrepreneur to an East Coast European historian is telling. In its short life, mass book digitization has moved in jerking motions toward consolidation and stabilization. The first such motion was the dueling efforts of the OCA and Google, short-lived but productive. The second was the Settlement, which would have domesticated Google’s project to the modern book apparatus, securing to authors and publishers firm control over their copyrights and providing them rents that were previously uncollectable. It would

³ You can see a photograph of the anthropologist at a meeting of the DPLA Steering Committee here: <http://blogs.law.harvard.edu/dplaalpha/about/steering-committee/>

have also achieved for Google what Google wanted—a license to scan all the books they chose to—or, in other words, a green light to digitize. The Settlement also created a safe harbor for Google’s partner libraries, enabling them to form the Hathi Trust, a central repository of their digitized books that they could develop separately from Google Book Search, a product that has failed to meet research library metadata or cataloging standards, or scholars’ expectations (Nunberg 2009a; Duguid 2007; Townsend 2007). Part of that safe harbor was an elaborate security protocol to which the libraries had to conform, lest they be subject to fines up to \$2 million. The security protocol had been intended to ally author and publisher anxieties about “digital risk”: that is, digital copies being released, like a virus, to the wilds of the Internet for infinite replication, forever forfeited from legitimate commerce. The Settlement failed for having kicked the beehive of the very book apparatus it had sought to serve, various representatives of which convinced the judge that it strayed too far from the current copyright system. Those who appreciated some of the ideas in the Settlement, however, are pursuing a third motion, a new path in a state-centered politics of legislative reform in the name of making a digital library for the nation. They are busy strategizing how best to turn the desirable parts of the Google Book Search Settlement into federal legislation (Samuelson 2011b; Darnton 2011b). Whether that will lead to new orphan works legislation, an “extended collective license,” or new exceptions for libraries is not clear, but this effort moves forward with a focused, savvy, Washington-minded momentum (Samuelson 2012).

After the DPLA West meeting in April 2012, when I floated by Brewster Kahle my observation that it was his phone call seven years prior that had put a circuitous path to the DPLA in motion, he thought about it for a moment then groaned. As with orphan works, he reflects back on his appeal to national-scale remedies as a misfire that has grown into the opposite of what he had sought: a centralized effort to make one “national library” rather than a distributed open library system made up of many libraries. He fears mostly the legislative remedies to copyright that will be sought from Congress, especially an “extended collective license,” which would establish a collecting society to assess fees for the use of digitized books (making, he says, every use of every conceivably copyrighted work into a “billable event.”) To him, such an outcome would be the death of the Open Web; to others, Kahle is a stubborn purist who doesn’t know what’s good for him.

This dread of Congressional action was one of the reasons Google representatives had said it needed the private Settlement; it was the only hope for mass digitization. Congress was broken; copyright could not be fixed. Lawrence Lessig has even switched career trajectories because of this impasse, from combatting copyright expansionism to combatting money in politics. Fred von Lohmann of the Electronic Frontier Foundation—who was a legal adviser to the Internet Archive during the Settlement period, before he became Google’s senior copyright counsel—expressed to me the same sentiment: the Settlement, if not ideal, was better than a

legislative solution because anything Congress might do would be worse. And yet it is toward Congress that the energy derived from the Settlement debate is primarily moving.

* * * * *

In February 2012, an illicit digital library—i.e., a popular book downloading portal—called library.nu closed itself down in response to an injunction sought and won by an international coalition of large commercial publishers. Among its fans, library.nu was valued for being scrupulously well curated with superb metadata and well stocked, especially with university-level books.⁴ Its closing was a replay of what happen to the digital music library Napster in 2001, though with much less fanfare. Library.nu existed among the “darknets,” the Internet underground where networks of users share various forms of digital “content.” I had learned of the darknets early during my time at the Archive, from one of the engineers as we chatted about digital books. Through my browsing about in IRC channels⁵ and BitTorrent sites,⁶ I realized that mass digitization such as I was studying (that done by the Internet Archive and Google) was only part of the story. The darknets were a mass digitization venue, too, only underground, “from below.” A small number of people administer sites like library.nu, thousands of people contribute to them, and millions more use them. Library.nu, in particular, had excellent cataloging: “Over-the-top good, beautiful metadata,” according to one person I talked to who knows the site well. That’s more than people have been willing to say about Google Book Search.

After library.nu was shutdown, I emailed Kahle a copy of fellow anthropologist Chris Kelly’s essay about the significance of library.nu (Kelly 2012). The Archive had already been involved in preparing its collection for BitTorrent distribution—which Kahle considers “the people’s distribution system”—and so he was curious to look into library.nu, which he hadn’t known about. (Library.nu of course still exists, only now in other reaches of the darknets). Once he had, he came away deeply impressed at the size (around 800,000 books), the sophistication of its metadata, but especially by the apparent commitment to its maintenance. People make and maintain these libraries, he told me, “not because they want to listen to Metallica; it’s because they care. This is obsession. Maybe these underground

⁴ For a range of comments about the shutdown of the library.nu site by its users, see: http://www.reddit.com/r/trackers/comments/ppfvc/librarynu_admin_the_website_is_shutting_down_due

⁵ Internet Relay Chat (IRC) is an Internet protocol for real-time messaging in channels where one can download or transfer data files.

⁶ BitTorrent is a common peer-to-peer file sharing Internet protocol that makes it easy to distribute large amounts of data.

worlds *are* the library. Maybe we're missing the boat. Maybe we are a relic." When he proudly quoted to me a recent tweet about the Archive being "the over ground of the underground world darknets," I sensed that he had been led by events to the conclusion that directing his efforts toward the digitizing multitude, rather than those "fiddling around with conferences," might be time better spent toward his goal of building the Library.

* * * * *

Mass book digitization jerks back and forth within a nest of open questions: Where does the over ground meet the underground? Where does the legitimate shade into the illegitimate? Is there hope for a "digital copyright"? If so, how will it happen—through changes in law or changes in practice? Will copying remain central to what copyright protects? Is "digital risk" a real fear or just a feverish nightmare? Will the twentieth-century, as embodied in books, be lost to future generations who only use digital libraries? Or, are such future behaviors mere phantoms conjured by techno-futurists? Despite their differences, Google and the Internet Archive are legitimating buffers between an extant copyright regime that works through ever-stricter controls and a massive global user base with a ravenous appetite for finding, sharing, and searching. This maneuvering means both organizations walk a razor's edge, risking their own existence with each move, constantly calibrating and strategizing.

Each organization took a very significant interest in books, for a while. As recounted in the Introduction, Kahle had thought that bringing books into our inevitable electronic future was a task for publishers, who would want to sell them, and for libraries, which would want to serve their patrons who will want to find and read their books on screens. He didn't think that Internet-focused companies and organizations would do it, but that's how it happened—at least in the case I have presented here. Still, mass digitization—of the largely copyrighted "twentieth century," when more books were published than all previous centuries combined—may yet fail. It is possible that the momentum will pass back to the publishers, the bulwarks of the modern book apparatus, and that they will selectively digitize themselves, creating aggregated databases that they can sell to libraries—a tried, true method. It is also possible that mass digitization will succeed; that is, Google will win its fair use argument and the green light to digitize will be turned on. In either case, the 25 to 30 million books that have been digitized so far in the projects I've been discussing in this dissertation definitely exist and will continue to exist, if experimentally, underground and/or over ground, in a fluid and uncertain, non-linear series of pathways to a future book.

Works Cited

- Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*. June 23.
- Anderson, Ivy. 2010. "Hurling Toward the Finish Line: Should the Google Books Settlement Be Approved?" Blogpost. February 16. Available at: <http://www.cdlib.org/cdlinfo/2010/02/16/hurling-toward-the-finish-line-should-the-google-books-settlement-be-approved/>
- Arato, Cynthia. 2009. Presentation on "L Is For Lawsuit" panel, D Is for Digitize Conference. October 10. New York Law School, New York, NY.
- Association of Research Libraries (ARL). 2012. *Code of Best Practices in Fair Use for Academic and Research Libraries*. January. <http://www.arl.org/pp/ppcopyright/codefairuse/code/index.shtml>
- Association of Research Libraries (ARL). 1964. "The Preservation of Deteriorating Books: An Examination of the Problem with Recommendations for a Solution," Report of the ARL Committee on the Preservation of Deteriorating Library Materials. Prepared for the Committee by Gordon Williams. September.
- Baker, Nicholson. 2001. *Double Fold: Libraries and the Assault on Paper*. New York: Random House.
- Bamman, David, and David A. Smith. 2011. "Extracting Two Thousand Years of Latin from a Million Book Library." *ACM Journal on Computing and Cultural Heritage*.
- Band, Jonathan. 2009. "The Long and Winding Road to the Google Books Settlement," 8 *The John Marshall Review of Intellectual Property Law* 227.
- Band, Jonathan. 2006. "The Google Library Project: Both Sides of the Story," *Plagiarism: Cross-Disciplinary Studies in Plagiarism, Fabrication and Falsification*.
- Barlow, John Perry. 1996. "A Declaration of the Independence of Cyberspace." Available at: <https://projects.eff.org/~barlow/Declaration-Final.html>
- Balsamo, Luigi. 1984 [1990]. *Bibliography: History of a Tradition*. Translated from the Italian by William A. Pettas. Berkeley, CA: Bernard M. Rosenthal, Inc.
- Battin, Patricia. 1991. "The Silent Books of the Future: Initiatives to Save Yesterday's Literature for Tomorrow." *Logos* 2 (1) 11-17.
- Bello, Susan E. 1986. "Cooperative Preservation Efforts of Academic Libraries." University of Illinois Graduate School of Library and Information Science Occasional Papers, No. 174. October.
- Benjamin, Walter. 1935. "The Work of Art in the Age of Its Technological Reproducibility, and Other Writings on Media: Second Version." In Jennings 2008, pp. 19-55.
- Benjamin, Walter. 1928. "Attested Auditor of Books," in Jennings 2008, pp. 171-72.
- Binkley, Robert C. 1936. *Manual on Methods of Reproducing Research Materials*. Ann Arbor, MI: Edwards Brothers, Inc.

- Binkley, Robert C. 1935. "New Tools for Men of Letters," *The Yale Review*. Spring. Reprinted in Fisch 1948, pp. 179-97. My pages numbers are to the reprinted version.
- Binkley, Robert C. 1931. *Methods of Reproducing Research Materials*. Ann Arbor, MI: Edwards Brothers.
- Binkley, Robert C. 1929. "The Problem of Perishable Paper," in Fisch 1948, pp. 169-78.
- Birkerts, Sven. 1994. *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. London: Faber and Faber.
- Blair, Ann. 2010. *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven: Yale Univ Press.
- Blanchette, Jean-François. 2011. "Material History of Bits." *Journal of the American Society for Information Science and Technology*. 62(6): 1042–57.
- Born, Lester K. 1960. "History of Microform Activity," *Library Trends* 8 (30): 348-58.
- Bowker, Geoffrey. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.
- Boyd, Danah, and Kate Crawford. 2011. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 21. Available at SSRN: <http://ssrn.com/abstract=1926431> or <http://dx.doi.org/10.2139/ssrn.1926431>
- Boyarin, Jonathan, ed. 1993. *The Ethnography of Reading*. Berkeley, CA: Univ of California Press.
- Boyle, James. 2008. *The Public Domain: Enclosing the Commons of the Mind*. New Haven: Yale Univ Press.
- Brown, Bill. 2010. "Materiality," in *Critical Terms for Media Studies*. W. J. T. Mitchell and Mark Hansen, eds., p. 49-63. Chicago, IL: Univ of Chicago Press.
- Buckland, Michael. 1997. "What Is a Document?" *Journal of the American Society for Information Science*, 48 (9) September: 804-09.
- Burke, Colin. 1994. *Information and Secrecy: Vannevar Bush, Ultra, and the Other Memex*. Metuchen, NJ: Scarecrow Press.
- Butler, Brandon. 2012. "The Librarians' Code, Orphan Works, and Mass Digitization." ARL Policy Notes blog. April. <http://policynotes.arl.org/post/20908237209/the-librarians-code-orphan-works-and-mass>
- Bush, Vannevar. 1953. "We Are in Danger of Building a Tower of Babel," *Public Health Reports* Vol 68, No. 2 (February): 149-52.
- Bush, Vannevar. 1945. "As We May Think," *Atlantic Monthly* 176 (July 1945): 101-108. Available at <http://www.theatlantic.com/magazine/archive/1969/12/as-we-may-think/3881/>.
- Carlson, Scott and Jeffrey R. Young. 2005. "Google will Digitize and Search millions of Books from Five Top Research Libraries," *Chronicle of Higher Education*. January 7, p. 37.
- Canfora, Luciano. 1989 [1990]. *The Vanished Library: A Wonder of the Ancient World*. Berkeley: Univ of California Press.

- Chartier, Roger. 2004. "Language, Books, and Reading from the Printed Word to the Digital Text," *Critical Inquiry* 31, 1 (Autumn): 133-52.
- . 1994 [1992]. *The Order of Books: Readers, Authors, and Libraries in Europe between the Fourteenth and Eighteenth Centuries*. Stanford, CA: Stanford Univ Press.
- . 1993. "Libraries without Walls," *Representations* 42 (Spring): 38-52.
- Christen, Kimberly. 2009. "Access and Accountability: The Ecology of Information Sharing in the Digital Age." *Anthropology News*. April.
- Clancy, Daniel J. 2009. Presentation, "I Is for Industry." D Is for Digitize conference, New York Law School, New York, NY, October 9.
- Clement, Tanya, Sara Steger, John Unsworth, Kirsten Uszkalo. 2009. "How Not to Read a Million Books." March. Available at: <http://people.lis.illinois.edu/~unsworth/hownot2read.html>.
- Cohen, Emily. 2004. "The Orphanista Manifesto: Orphan Films and the Politics of Reproduction" *American Anthropologist*. Vol. 106, Issue 4, pp. 719–31.
- Coleman, Gabriella. 2012. *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton, NJ: Princeton Univ Press.
- Coleman, Gabriella. 2010. "Ethnographic Approaches to Digital Media." *Annual Review of Anthropology*. 39:1-19.
- Council on Library Resources (CLR). 1986. *Brittle Books: Reports on the Committee on Preservation and Access*. Washington, DC: Council on Library Resources.
- Covey, Denise Troll. 2005a. *Acquiring Copyright Permission To Digitize and provide Open Access to Books*. Washington, DC : Digital Library Federation Council on Library and Information Resources Available at: <http://www.clir.org/pubs/reports/pub134/pub134col.pdf>.
- Covey, Denise Troll, 2005b. "Rights, Registries, and Remedies: An Analysis of Responses to the Copyright Office Notion of Inquiry Regarding Orphan Works." *Library Research and Publications*. Paper 48. Available at: http://repository.cmu.edu/lib_science/48.
- Coyle, Karen. 2005. "Understanding Metadata and Its Purpose." *The Journal of Academic Librarianship*. 31 (March) 2: 160-63.
- Crane, Gregory, et al. 2011. "What Did We Do with a Million Books? Rediscovering the Greco-Ancient World and Reinventing the Humanities." NEH White Paper. <http://www.perseus.tufts.edu/hopper/about/publications>.
- Crane, Gregory. 2009. Letter to Judge Denny Chin. August 7, 2009. Available at: <http://thepublicindex.org/documents/responses>
- Crane, Gregory. 2006. "What Do You Do with a Million Books?" *D-Lib Magazine*. Vol. 12, No. 3 March.
- Crane, Gregory. 2005. "Reading in the Age of Google." *Humanities*. September/October. Volume 26, no. 5.
- Crawford, Walt. 2012. "It Was Never a Universal Library: Three Years of the Google Book Settlement." *Cites and Insights*. Vol. 12, No. 7. August.

- Csiszar, Alex. 2010. "Broken Pieces of Fact: The Scientific Periodical and the Politics of Search in Nineteenth-Century France and Britain." Ph.D. Dissertation. Harvard Univ, Dept of the History of Science.
- Darnton, Robert. 2011a. "Google's Loss: The Public's Gain." *New York Review of Books*. April 28.
- Darnton, Robert. 2011b. "Jefferson's Taper: A National Digital Library." *New York Review of Books*. November 24.
- Darnton, Robert, 2010. "Can We Create a National Digital Library?" *New York Review of Books*. October 28.
- Darnton, Robert. 2009. *The Case for Books: Past, Present, Future*. New York: Public Affairs.
- Darnton, Robert. 1999. "The New Age of the Book." *New York Review of Books*. March 18.
- Deegan, Marilyn, and Kathryn Sutherland. 2009. *Transferred Illusions: Digital Technology and the Forms of Print*. Surrey, England: Ashgate.
- Dirks, Nicholas. 2002. "Annals of the Archive: Ethnographic Notes on the Sources of History." In *From the Margins: Historical Anthropology and Its Futures*, ed. Brian Axel, pp. 47-65. Durham, N.C.: Duke Univ Press.
- Drucker, Johanna. 2012. "Humanistic Theory and Digital Scholarship," in Gold 2012, pp. 85-95.
- Drucker, Johanna. 1994. *The Century of Artists' Books*. New York: Granary Books.
- Duguid, Paul. 2007. "Inheritance and loss? A brief survey of Google Books." *First Monday*, volume 12, number 8 (August).
Available at: http://firstmonday.org/issues/issue12_8/duguid/index.html
- Dyson, George. 2005. "Turing's Cathedral." *The Edge*. October 24.
http://www.edge.org/3rd_culture/dyson05/dyson05_index.html
- Edwards, Eli. 2004. "Ephemeral to Enduring: The Internet Archive and Its Role in Preserving Digital Media." *Information Technology and Libraries*. March.
- Eisenstein, Elisabeth. 1979. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-modern Europe*. Cambridge: Cambridge Univ Press.
- Escobar, Arturo. 1994. "Welcome to Cyberia: Notes on the Anthropology of Cyberculture." *Current Anthropology* Vol. 35, No. 3 (June): 211-31.
- Farkas-Conn, Irene S. 1990. *From Documentation to Information Science: The Beginnings and Early Development of the American Documentation Institute-American Society for Information Science*. New York: Greenwood Press.
- Fayyad, Usama. 2001. "The Digital Physics of Mining." *Communications of the ACM*. March, Vol. 44, No. 3, pp. 62-65.
- Febvre, Lucien, and Henri-Jean Martin. 1976 [1958]. *The Coming of the Book: The Impact of Printing, 1450-1800*. London: Verso.
- Fessenden, Reginald. 1896. "Use of Photography in Data Collections." *Electrical World*. August 22. Vol 28, 8:222-23.

- Fisch, Max H., ed. 1948. *Selected Papers of Robert C. Binkley*. Cambridge, MA: Harvard Univ Press.
- Foucault, Michel. 2007. *Security, Territory Population: Letters at the College de France 1977-1978*, trans. Graham Burchell. New York: Palgrave.
- Foucault, Michel. 1980. "The Confession of the Flesh," in *Power/Knowledge: Selected Interviews and Other Writings 1972-1977*, pp. 194-98. Trans. Colin Gordon et al. New York: Random House.
- Foucault, Michel. 1978. *The History of Sexuality, Vol 1*, Trans. Robert Hurley. New York: Random House.
- Foucault, Michel. 1969. "What is an Author?", in *Textual Strategies: Perspectives in Post-Structuralist Criticism*, ed. in Josué V. Harari, Ithaca, NY: Cornell Univ Press, 1979.
- Fowler, Geoffrey A. 2010. "Libraries Have a Novel Idea: Lenders Join Forces to Let Patrons Check-Out Digital Scans of Shelved Book Collections." *Wall Street Journal*. June 29.
- Franzen, Jonathan. 2012. "Jonathan Franzen Hates E-Books," Huffington Post, January 30. http://www.huffingtonpost.com/2012/01/30/jonathan-franzen-ebooks-quotes_n_1242151.html.
- Frick, Caroline. 2011. *Saving Cinema: The Politics of Preservation*. New York: Oxford Univ Press.
- Fussler, Herman. 1940. "Microfilm and Libraries," in William Madison Randall, ed., *Acquisition and Cataloging of Books*, Chicago, IL: Univ of Chicago Press, 1940, 321-354. Reprinted in Venear 1979, 5-21. My page numbers refer to the Veaneer volume.
- Gadd, Ian. 2009. "The Use and Misuse of *Early English Books Online*." *Literature Compass* 6/3: 680-92.
- Gass, William. 1999. "In Defense of the Book: On the Enduring Pleasures of Paper, Type, and Ink." *Harpers*. November. <http://harpers.org/archive/1999/11/0060708>.
- Geoghegan, Bernard Dionysius. 2011. "From Information Theory to French Theory: Jakobson, Levi-Strauss, and the Cybernetic Apparatus." *Critical Inquiry* 38 (Autumn): 96-126.
- Gibbs, Frederick W., and Daniel J. Cohen. 2012. "A Conversation with Data: Prospecting Victorian Words and Ideas." *Victorian Studies*. 54:1: 69-77.
- Ginsburg, Faye, Lila Abu-Lughod, and Brian Larkin, eds. 2002. *Media Worlds: Anthropology on New Terrain*. Berkeley: Univ of California Press.
- Gitelman, Lisa. Forthcoming. "The Typescript Book." Manuscript.
- Gitelman, Lisa. 2006. *Always Already New: Media, History, and the Data of Culture*. Cambridge, MA: MIT Press.
- Gold, Matthew, ed. 2012. *Debates in the Digital Humanities*. Minneapolis: Univ of Minnesota Press.
- Goody, Jack, and Ian Watt. 1963. "The Consequences of Literacy." *Comparative Studies in Society and History*, Vol. 5, No. 3 (Apr., 1963), pp. 304-45.

- Gordon, Wendy J. 1982. "Fair Use As Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors." *82 Columbia Law Review* 1600.
- Grimmelman, James. "A Bridge Too Far: Future Conduct and the Limits of Class-Action Settlements." April 12. Berkeley Center for Law and Technology, Orphan Works Symposium, Berkeley, CA.
- Grimmelman, James. 2010. "Internet Archive Starts Lending In-Copyright E-Books." Blog post. June 30.
http://laboratorium.net/archive/2010/06/30/gbs_internet_archive_starts_lending_in-copyright_e
- Hafner, Katie. 2005. "In Challenge to Google, Yahoo Will Scan Books." *New York Times*. October 3.
- Hafner, Katie. 1999. "Between Tech Fans and Naysayers, Scholarly Skeptics," *New York Times*, April 1.
- Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems*. March/April, vol. 24 no. 2.
- Hand, Eric. 2011. "Culturomics: Word Play." *Nature* 474, 436-440. June 17.
- Hanratty, Elisabeth. 2005. "Google Library: Beyond Fair Use?" *Duke Law and Technology Review* 10.
- Hansen, David R. 2011. "Orphan Works: Definitional Issues" (December 19). Berkeley Digital Library Copyright Project White Paper No. 1. Available at: <http://ssrn.com/abstract=1974614> or <http://dx.doi.org/10.2139/ssrn.1974614>
- Hansen, David R. 2012. "Orphan Works: Causes of the Problem." Berkeley Digital Copyright Project White Paper No. 3. Available at SSRN: <http://ssrn.com/abstract=2038068>.
- Hapke, Thomas. 1999. "Wilhelm Ostwald, the 'Brücke' (Bridge), and Connections to Other Bibliographic Activities at the Beginning of the Twentieth Century," In *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, 139-147. Edited by Mary Ellen Bowden, Trudi Bellardo Hahn, Robert V. Williams. Medford, NJ: Information Today.
- Hardy, Quentin. 2009. "Lend Ho! Brewster Kahle Is a Thorn in Google's Side." *Forbes*. November 16.
- Hardy, Quentin. 2011. "The Big Business of 'Big Data'" *New York Times*, Bits Blog. October 24.
- Hargreaves, Ian. 2011. *Digital Opportunity: A Review of Intellectual Property and Growth*. Available online: <http://www.ipo.gov.uk/ipreview.htm>
- Hayes, Robert. 1987. "The Magnitude, Costs, and Benefits of the Preservation of Brittle Books", Report #0 on the Preservation Project. Sherman Oaks, CA. November 30.
- Heath, Shirley. 1980. "Functions and Uses of Literacy." *Journal of Communication*. Volume 30, Issue 1, March, pp. 123-33.
- Helft, Miguel. 2009. "Google's Plan for Out-of-Print Books Is Challenged." *New York Times*, April 3.

- Hellman, Eric. 2010. "Internet Archive Sets Fair-Use Bait With Open Library Lending." Blog post. July 2. <http://go-to-hellman.blogspot.com/2010/07/internet-archive-sets-fair-use-bait.html>
- Hesse, Carla. 1996. "Books in Time," in Geoffrey Nunberg, ed. *The Future of the Book*. Berkeley: Univ of California Press.
- Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Press.
- Hirtle, Peter. 2010. Comments to Grimmelman 2010, above.
- Hirtle, Peter, Emily Hudson, and Andrew T. Kenyon. 2009. *Copyright and Cultural Institutions: Guidelines for Digitization for U.S. Libraries, Archives, and Museums*. Self-published.
- Holland, Brad. 2010. "Trojan Horse: Orphan Works and the War on Authors." *The Journal of Biocommunication*. Vol. 36, Nov. 1.
- Howard, Jennifer. 2012. "Google Begins to Scale Back Its Scanning of Books From University Libraries." *Chronicle of Higher Education*. March 9.
- Howell, Beryl A. 2006. "Proving Web History: How to Use the Internet Archive." *Journal of Internet Law*. February.
- Information School, UC Berkeley, 2009. Conference: The Google Book Search Settlement and the Future of Information Access. August 29. Video at: <http://www.ischool.berkeley.edu/newsandevents/events/20090828googlebooksconference>
- Jacobs, Adam. 2009. "The Pathologies of Big Data." *ACM Queue*.
- James, Ryan, and Andrew Weiss. 2012. "An Assessment of Google Books' Metadata." *Journal of Library Metadata*. 12:1, 15-22
- Jaszi, Peter. 1994. "On the Author Effect," in Jaszi and Martha Woodmansee 1994.
- Jaszi, Peter, and Martha Woodmansee, eds. 1994. *The Construction of Authorship: Textual Appropriation in Law and Literature*. Durham, NC: Duke Univ Press.
- Jennings, Michael, ed., et al. 2008. *The Work of Art in the Age of Its Technological Reproducibility, and Other Writings on Media*. Cambridge, MA: Harvard Univ Press.
- Johns, Adrian. 1998. *The Nature of the Book: Print and Knowledge in the Making*. Chicago, IL: Univ of Chicago Press.
- Kahle, Brewster. 2012. Presentation. Orphan Works and Mass Digitization conference. Berkeley Center for Law and Technology conference. Berkeley, CA. April 21-22. Audio available at: <http://archive.org/details/Orphanworksandmassdigitization20120412>
- Kahle, Brewster. 2009a. "It's All About the Orphans." Blog post. February 23. <http://www.opencontentalliance.org/2009/02/23/its-all-about-the-orphans/>
- Kahle, Brewster. 2009b. "Google Claims to be the Lone Defender of Orphans: Not Lone, Not Defender." Blog post. October 7. <http://www.opencontentalliance.org/2009/10/07/google-claims-to-be-the-lone-defender-of-orphans-not-lone-not-defender/#comments>

- Kahle, Brewster. 2006. Plenary address. Wikimania conference. Harvard Law School, Cambridge MA. August 5. Available at: http://archive.org/details/wm06_kahle_plenary_5aug2006
- Kahle, Brewster. 2005. Presentation in "Managing Knowledge and Creativity in a Digital Context" series. Library of Congress, Washington D.C. December 13. Available at: <http://www.c-spanvideo.org/program/184428-1>
- Kahle, Brewster, Rick Prelinger, and Mary E. Jackson. 2001. "Public Access to Digital Material." *D-Lib Magazine*. Volume 7, Number 10. Available at <http://www.dlib.org/dlib/october01/kahle/10kahle.html>
- Kahle, Brewster. 1997. "Preserving the Internet." *Scientific American*, March, pp. 82-83. Special issue on "The Internet: Fulfilling the Promise."
- Katz, Eli. 2012. "Web freedom faces greatest threat ever, warns Google's Sergey Brin." *The Guardian*. April 15. Available at: <http://www.guardian.co.uk/technology/2012/apr/15/web-freedom-threat-google-brin>
- Kelly, Kevin. 2006. "Scan This Book!" May 16. *New York Times Magazine*.
- Kelty, Chris. 2012. "Our Disappearing Virtual Library." March 1. <http://www.aljazeera.com/indepth/opinion/2012/02/2012227143813304790.html>
- Kelty, Chris. 2008. *Two Bits: The Cultural Significance of Free Software*. Durham, NC: Duke Univ Press.
- Kichuk, Diana. 2007. "Metamorphosis: Remediation in *Early English Books Online* (EEBO)." *Literary and Linguistic Computing* Vol. 22, No. 3: 291-303.
- Kirschenbaum, Matthew. 2008. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press.
- Kittler, Friedrich A. 1999. *Gramophone, Film, Typewriter*. Stanford, CA: Stanford University Press.
- Kittler, Friedrich. 1990 [1985]. *Discourse Networks 1800 / 1900*. Stanford, CA: Stanford Univ Press.
- Krajewski, Markus. 2011. *Paper Machines: About Cards & Catalogs, 1548-1929*. Cambridge: MIT Press.
- Landes, William M. and Richard A. Posner. 2003. *The Economic Structure of Intellectual Property Law*. Cambridge, MA: Harvard Univ Press.
- Landes, William M. and Richard A. Posner. 2003. "Indefinitely Renewable Copyright." *University of Chicago Law Review*, Vol. 70, No. 2 (Spring): 471-518
- Langley, Adam, and Dan S. Bloomberg. 2006. "Google Books: Making the Public Domain Universally Accessible." *Document Recognition and Retrieval XIV*.
- Larkin, Brian. 2008. *Signal and Noise: Media, Infrastructure, and Urban Culture in Nigeria*. Durham, NC: Duke Univ Press.
- Lass, Andrew. 1999. "Portable Worlds: On the Limits of Replication in the Czech and Slovak Republics." In Michael Burawoy and Katherine Verdery, eds. *Uncertain Transition: Ethnographies of Change in the Former Socialist World*. Berkeley: Univ of California Press.

- Lavoie, Brian, Lynn Silipigni Connaway, and Lorcan Dempsey. 2005. "Anatomy of Aggregate Collections: The Example of Google Print for Libraries." *D-Lib Magazine*. Volume 11, No. 9. September.
- Lavoie, Brian, and Lorcan Dempsey. 2009. "Beyond 1923: Characteristics of Potentially In-copyright Print Books in Library Collections." *D-Lib Magazine*. Nov/Dec, No. 11/12.
- Lesk, Michael 1997. *Practical Digital Libraries: Books, Bytes, and Bucks*. San Francisco, CA: Morgan Kaufman Publishers, Inc.
- Lessig, Lawrence. 2006. Video. "Is Google Book Search Fair Use?" http://lessig.org/blog/2006/01/google_book_search_the_argumen.html
- Lessig, Lawrence. 2004. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York: Penguin Press.
- Lessig, Lawrence. 2001. *The Future of Ideas: The Fate of the Commons in a Connected World*. New York: Random House.
- Levi-Strauss, Claude. 1955 [1975]. "A Writing Lesson." In *Tristes Tropiques*. Translated by John and Doreen Weightman. New York: Atheneum.
- Levy, Steven. 2011a. *In the Plex: How Google Thinks, Works, and Shapes Our Lives*. New York: Simon and Schuster.
- Levy, Steven. 2011b. Interview with Andrew Keen. Hillside Club, Berkeley, CA. April 17.
- Lieberman, Erez, and Jean-Baptiste Michel. 2011a. "Culturomics: Quantitative Analysis of Culture Using Millions of Digitized Books." Closing keynote. Digital Humanities 2011 conference, Stanford University. June 22. <http://youtu.be/sqRz3g8aIN4?t=14m37s>.
- Lieberman, Erez, and Jean-Baptiste Michel. 2011b. TED Talk, "What We Learned from Five Million Books." http://www.ted.com/talks/what_we_learned_from_5_million_books.html
- Lieberman, Erez, and Jean-Baptiste Michel. 2011c. "Culturomics: Quantitative Analysis of Culture Using Millions of Digitized Books," Presentation, Berkman Center, Harvard University. Available at: <http://cyber.law.harvard.edu/interactive/events/luncheon/2011/05/culturomics>
- Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang and Martin A. Nowak. 2007. "Quantifying the evolutionary dynamics of language." *Nature*. Vol 449. October 11.
- Litman, Jessica. 2001. *Digital Copyright*. Amherst, NY: Prometheus Books.
- Lohr, Steve. 2012. "The Age of Big Data." *New York Times*, February 11.
- Lohr, Steve, and David Streitfeld. "Data Engineer in Google Case Is Identified" *New York Times*, April 30.
- Loren, Lydia Pallas. 2012. "Abandoning the Orphans: An Open Access Approach to Hostage Works." *Berkeley Technology Law Journal*, Forthcoming; Lewis & Clark Law School Legal Studies Research Paper No. 2012-10. Available at SSRN: <http://ssrn.com/abstract=2049685>

- Lukow, Gregory. 1999. "The Politics of Orphanage: The Rise and Impact of the 'Orphan Film' Metaphor on Contemporary Preservation Practice." Paper delivered at the Orphans of the Storm I Conference, September 23. Available at: <http://www.sc.edu/filmsymposium/archive/orphans2011/lukow.html>
- Luther, Frederic. 1959. *Microfilm: A History 1839-1900*. Annapolis, MD: The National Microfilm Institute.
- Lyman, Peter. 2002. "Archiving the World Wide Web," in *Building a National Strategy for Digital Preservation*. CLIR Publication 106, pp. 38-51. Washington, D.C.: Council on Library and Information Resources.
- Lyman, Peter, and Brewster Kahle. 1998. "Archiving Digital Cultural Artifacts: Organizing an Agenda for Action," *D-Lib Magazine*. July/August.
- McGray, Douglas. 2009. "Print: Applying Quantitative Analysis to Classic Lit," *Wired*. November 12. http://www.wired.com/magazine/2009/11/pl_print/
- McLean, Austin J. 2001. "Early British Printing Meets the Electronic Age: A Large-scale Digitization Case Study." *Microform & Imaging Review*. 30:4, pp.127-34.
- McLuhan, Marshall. 1962. *The Gutenberg Galaxy: The Making of Typographic Man*. Toronto: Univ of Toronto Press.
- Mak, Bonnie. 2011. *How the Page Matters*. Toronto: Univ of Toronto Press.
- Manovich, Lev. 2001. *The Language of New Media*. Cambridge, MA: MIT Press.
- Marcus, George, ed. 1999. *Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas*. Santa Fe, NM: School of American Research Press.
- Markoff, John. 2009. "A Deluge of Data Shapes a New Era in Computing." *New York Times*. December 15.
- Markoff, John, and Edward Wyatt. 2004. "Google Is Adding Major Libraries to Its Database." *New York Times*, December 14.
- Martin, Henri-Jean. 1994 [1988]. *The History and Power of Writing*. Chicago, IL: Univ of Chicago Press.
- Melville, Annette and Scott Simmon. 1993. *Film Preservation 1993: A Study of the Current State of American Film Preservation: Report of the Librarian of Congress*, vol. 1. Washington D.C.: National Film Preservation Board of the Library of Congress. Available at: <http://www.loc.gov/film/study.html>
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden. 2010a. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 14 January: 331 (6014), 176-182.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden. 2010b. "Supporting Online Material," for (Michel et al. 2010). www.sciencemag.org/cgi/content/full/science.1199644/DC1

- Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models For A Literary History*. London: Verso.
- Moretti, Franco. 2000. "Conjectures on World Literature." *New Left Review*.
<http://www.newleftreview.org/A2094>
- Mueller, Martin. 2007. "Notes Towards a Users Manual of MONK." Available at:
<http://monkproject.org/MONK.wiki/Notes%20towards%20a%20user%20manual%20of%20Monk.html>
- Murrell, Mary. 2010.
- Musto, Ronald G. 2009. "Google Books Mutilates the Printed Past," *Chronicle of Higher Education*. June 12.
- Nakashima, Ellen. 2008. "FBI Backs Off from Secret Order for Data after Lawsuit." *Washington Post*. May 8.
- Negroponce, Nicholas. 1995. *Being Digital*. New York: Alfred A. Knopf.
- Nunberg, Geoffrey. 2010. "Counting on Google Books." *The Chronicle Review*. December 16.
- Nunberg, Geoffrey. 2009a. "Google Books: A Metadata Trainwreck." Language Log blog. August 29. <http://languageblog.ldc.upenn.edu/nll/?p=1701>.
- Nunberg, Geoffrey. 2009b. "Google Book Search: A Disaster for Scholars." *The Chronicle Review*. August 31.
- Nunberg, Geoffrey, ed. 1996. *The Future of the Book*. Berkeley, CA: Univ of California Press.
- O'Connor, Carol. 1983. *A Sort of Utopia: Scarsdale, 1891-1981*. Albany: SUNY Press.
- Ong, Aihwa, and Stephen Collier, ed. 2003. *Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems*. Malden, MA: Blackwell.
- Orwant, Jon. 2010. "Google Books," Conference presentation. "Future of Reading." Rochester Institute of Technology. Rochester, NY, June 11.
- Orwant, Jon. 2009. Response to Nunberg 2009a.
<http://languageblog.ldc.upenn.edu/nll/?p=1701>.
- Otlet, Paul. 1918. "Transformations in the Bibliographic Apparatus of the Sciences," in Rayward 1990, pp. 148-56.
- Otlet, Paul. 1903. "The Science of Bibliography and Documentation," in Rayward 1990, pp. 71-86.
- Otlet, Paul. 1892. "Something about Bibliography," in Rayward 1990, pp. 18-24.
- Otlet, Paul, and Robert Goldschmidt. 1925. "The Preservation and International Diffusion of Thought: The Microphotoc Book" in Rayward 1990, pp. 204-10.
- Otlet, Paul, and Robert Goldschmidt. 1906. "On a New Form of the Book: The Microphotographic Book," in Rayward 1990, pp. 87-95.
- Parry, Marc. 2010. "The Humanities Go Google." *Chronicle of Higher Education*.
- Patry, William. 2012. *How to Fix Copyright*. New York: Oxford Univ Press.
- Patry, William. 2009. *Moral Panics and the Copyright Wars*. New York: Oxford Univ Press.
- Patterson, L. Ray, and Stanley Lindberg. 1991. *The Nature of Copyright: A Law of Users' Rights*. Athens: Univ of Georgia Press.

- Power, Eugene. 1990. *Edition of One: The Autobiography of Eugene B. Power, Founder of University Microfilms*. Ann Arbor, MI: University Microfilms International.
- Rabinow, Paul. 2010. Bios-Technika. Concepts: Pathway. See <http://www.bios-technika.net/concepts.php#pathway> Last accessed January 25, 2012.
- . 2003. *Anthropos Today: Reflections on Modern Equipment*. Princeton, NJ: Princeton Univ Press.
- . 1999. *French DNA: Trouble in Purgatory*. Chicago, IL: Univ of Chicago Press.
- . 1989. *French Modern: Norms and Forms of the Social Environment*. Cambridge, MA: MIT Press.
- Rabinow, Paul, and Gaymon Bennett. 2012. *Contemporary Equipment: A Diagnostic*. Available at: <http://www.bios-technika.net/>
- Rabinow, Paul, and George Marcus. 2008. *Designs for an Anthropology of the Contemporary*. Durham, NC: Duke Univ Press.
- Rackley, Marilyn. 2010. "Internet Archive." In, *Encyclopedia of Library and Information Sciences*, pp. 2966-976. Third Edition. Boca Raton, FL : CRC Press.
- Rayward, W. Boyd. 2008. "European Modernism and the Information Society: Introduction," in *European Modernism and the Information Society*, ed. W. Boyd Rayward, pp. 1-26. Hampshire, England: Ashgate.
- Rayward, Boyd, ed. 1990. *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet*. New York: Elsevier.
- Rayward, W. Boyd. 1983. "The International Exposition and the World Documentation Congress, Paris 1937," *The Library Quarterly* Vol 53, No. 3 (July): 254-68.
- Rayward, W. Boyd. 1975. *The Universe of Information: The Work of Paul Otlet for Documentation and International Organisation*. Moscow: International Federation for Documentation, 1975.
- Reetz, Daniel. 2010a. "The Why in DIY Book Scanning." *New York Law School Law Review* 55: 251-69.
- Reetz, Daniel. 2010b. "The Why in DIY Book Scanning." Presentation. D Is for Digitize conference, New York Law School, October 9. Audio available at: http://nyls.mediasite.com/mediasite/FileServer/Podcast/0d1762bc-40a1-4454-a449-de5cf367e3ea/d_is_for_digitize.xml
- Rider, Fremont. 1944. *The Scholar and the Future of the Research Library*. New York: Hadham Press.
- Riles, Annelise, ed. *Documents: Artifacts of Modern Knowledge*. Ann Arbor: Univ of Michigan Press.
- Ringer, Barbara. A. 1961. Study No. 31: Renewal of Copyright. *Copyright Law Revision: Studies Prepared for the Subcommittee on Patents, Trademarks, and Copyrights of the Committee on the Judiciary, United States Senate, Eighty-sixth Congress, first [second] session*. Washington, D.C.: U.S. Government Printing Office.

- Rosenzweig, Roy. 2007. "Scarcity or Abundance? Preserving the Past in a Digital Era." In *Institutions of Reading: The Social Life of Libraries in the United States*, pp. 310-42. Eds. Thomas August and Kenneth E. Carpenter. Amherst: Univ of Massachusetts Press.
- Sag, Matthew. 2012. "Orphan Works as Grist for the Data Mill." (working paper) <http://ssrn.com/abstract+2038889>.
- Sag, Matthew. 2010. "The Google Book Settlement and the Fair Use Counterfactual," *New York Law School Law Review*, 55.
- Sag, Mathew. 2009. "Copyright and Copy-Reliant Technology," *Northwestern University Law Review*. 103.
- Samuelson, Pamela. 2012. "Reforming Copyright Is Possible: And It's the Only Way to Create a National Digital Library." *Chronicle of Higher Education*. July 9.
- Samuelson, Pamela. 2011a. "The Google Book Settlement as Copyright Reform." *Wisconsin Law Review*. Volume 2011, no. 2.
- Samuelson, Pamela. 2011b. "Legislative Alternatives to the Google Book Settlement," *Columbia Journal and Law and Arts*. 34.
- Samuelson, Pamela. 2009. "K is for Keynote," D Is for Digitize Conference, October 8-10, New York Law School, New York, New York.
- Samuelson, Pamela. 1998. "Encoding the Law into Digital Libraries." *Communications of the ACM*. April, Vol. 41. No. 4, pp. 13-18.
- Samuelson, Pamela. 1996. "The Copyright Grab," *Wired*, 4.01. January.
- Schultz, Claire K., and Paul L. Garwig. 1969. "History of the American Documentation Institute—A Sketch", *American Documentation*. April: 152-60.
- Sloan Foundation. 2006. Annual Report. available at: <http://www.sloan.org/pages/15/annual-reports-financial-statements-of-the-alfred-p-sloan-foundation>.
- Smith, Abby. 1999. *The Future of the Past: Preservation in American Research Libraries*. Washington, DC: Council on Library and Information Resources, April.
- Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine." *Ninth International Conference on Document Analysis and Recognition, 2007*, vol.2, no., pp. 629-33, 23-26
- Stanfill, Craig, and David Waltz. 1986. "Toward Memory-Based Reasoning." *Communications of the ACM*. 29 (12): 1213-1228.
- Stefik, Mark. 1996. *Internet Dreams: Archetypes, Myths, Metaphors*. Cambridge, MA: MIT Press.
- Street, Brian. 1993. *Cross-cultural Approaches to Literacy*. Cambridge: Cambridge Univ Press.
- Streible, Dan. 2009. "The State of Orphan Films," *Moving Image*. Issue 1, Vol. 9: vi-xix.
- Streible, Dan. 2007. "The Role of Orphan Films in the 21st Century Archive." *Cinema Journal* 46, No. 3 (Spring 2007): 124-28.

- Streible, Dan. 2006. "What Is an Orphan Film?" Available at:
<http://www.sc.edu/filmsymposium/orphanfilm.html>
- Streitfeld, David. 2012. "In a Flood Tide of Digital Data, an Ark Full of Books." *New York Times* (front page). March 3.
- Striphas, Ted. 2009. *The Late Age of Print: Everyday Book Culture from Consumerism to Control*. New York, NY: Columbia Univ Press.
- Stross, Randall. 2008. *Planet Google: One Company's Audacious Plan To Organize Everything We Know*. New York: Free Press.
- Taubes, Gary A. 1995. "The Rise and Fall of Thinking Machines," *Inc. Magazine*. September 15. <http://www.inc.com/magazine/19950915/2622.html>.
- Thibodeau, Kenneth. 2002. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years." In *The State of Digital Preservation: An International Perspective*. Washington, D.C.: Council on Library and Information Resources. Available at:
<http://www.clir.org/pubs/reports/pub107>.
- Thompson, John B. 2010. *Merchants of Culture: The Publishing Business in the Twenty First Century*. Oxford: Policy Press.
- Toobin, Jeffrey. 2007. "Google's Moon Shot: The Quest for the Universal Library," *New Yorker*, February 5.
- Townsend, Robert B. 2007. "Google Books: What's Not to Like?" *AHA Today*, April 30, <http://blog.historians.org/articles/204/google-books-whats-not-to-like>.
- Turner, Fred. 2006. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago, IL: Univ of Chicago Press.
- Tushnet, Rebecca. 2012. "Worth a Thousand Words: The Images of Copyright." *125 Harvard Law Review* 683 (2012)
- Urban, Jennifer. 2010. *Updating Fair Use for Innovators and Creators in the Digital Age: Two Targeted Reforms*, Report for Public Knowledge. Available at:
<http://www.publicknowledge.org/pdf/fair-use-report-02132010.pdf>
- United States Copyright Office. 2006. "Report on Orphan Works," January. Available at: www.copyright.gov/orphan/orphan-report.pdf.
- United States Supreme Court, 1991. *Feist Publications, Inc. vs. Rural Telephone Service Co.*, 499 U.S. 340 (1991). Available online at:
http://www.law.cornell.edu/copyright/cases/499_US_340.htm
- Usai, Paolo Cherchi. 1999. "What is an Orphan Film?" Definition, Rationale and Controversy." Paper Presentation Orphans of the Storm Conference, University of South Carolina, September 23, 1999. Available at:
www.sc.edu/filmsymposium/archive/orphans2001/usai.html
- Vaidhyanathan, Siva. 2004. *The Anarchist in the Library: How the Clash Between Freedom and Control Is Hacking the Real World and Crashing the System*. New York: Basic Books.

- Van den Boomen, Marianne et al. 2010. *Digital Material: Tracing New Media in Everyday Life and Technology*. Amsterdam: Amsterdam Univ Press.
- Varvel, Virgil E., Jr. and Andrea Thomer. 2011. "Google Digital Humanities Awards Recipient Interviews Report." Report Prepared for the Hathi Trust Research Center. University of Illinois at Urbana-Champaign. December. <https://www.ideals.illinois.edu/handle/2142/29936>
- Veneer, Allen B., ed. 1976. *Studies in Micropublishing, 1853-1976: Documentary Sources*. Westport, CT: Microform Review, Inc.
- Waltz, David, and Simon Kasif. 1995. "On Reasoning from Data." *ACM Computing Surveys*, Vol 27. No 3, September.
- Wells, H. G. 1938. *World Brain*. New York: Doubleday.
- Wilkin, John P. 2011. "Bibliographic Indeterminacy and the Scale of Problems and Opportunities of 'Rights' in Digital Collection Building." February. Council on Library and Information Resources Research Paper. Available at: <http://www.clir.org/pubs/ruminations/01wilkin>
- Williams, Raymond. 1977. *Marxism and Literature*. Oxford: Oxford Univ Press.
- Williams, William Proctor, and William Baker. 2001. "Caveat Lector: English Books 1475-1700 and the Electronic Age." *Analytical and Enumerative Bibliography* 12: 1-29.
- Wilson, Andrew Norman. 2011. "Workers Leaving the Googleplex," Video and transcript. <http://www.andrewnormanwilson.com/portfolios/70411-workers-leaving-the-googleplex>
- Wright, Alex. 2007. *Glut: Mastering Information through the Ages*. Washington, D.C: Joseph Henry Press.
- Wujastyk, Dominick. 2012. "Burning the Library of Alexandria, Again." Blogpost. <http://cikitsa.blogspot.com/2012/02/website-formerly-known-as-httplibrary.html>
- Yi, Esther. 2012. "Inside the Quest to Put the World's Libraries Online." *The Atlantic*. July. Available at: www.theatlantic.com/...quest-to-put-the-worlds-libraries.../259967/
- Zittrain, Jonathan. 2008. *The Future of the Internet and How to Stop It*. New Haven, CT: Yale Univ Press.