Books.Files

Preservation of Digital Assets in the Contemporary Publishing Industry: A Report

Matthew Kirschenbaum

Books.Files

Preservation of Digital Assets in the Contemporary Publishing Industry: A Report

Matthew Kirschenbaum

April 2020
University of Maryland
College Park, MD

Produced by

MITH MARYLAND INSTITUTE FOR TECHNOLOGY IN THE HUMANITIES

BISG BOOK INDUSTRY STUDY GROUP

With support from a grant by

THE ANDREW W. MELLON FOUNDATION

Published April 2020

Original instances of this document are available from the Digital Repository at the University of Maryland (DRUM), the Book Industry Study Group, and Humanities Commons

URI: https://drum.lib.umd.edu/handle/1903/25605

Design by Deborah Rust

Suggested Citations

For the main body of this report and the report in its totality, we suggest the following citation:

Kirschenbaum, Matthew, et al. *Books.Files: Preservation of Digital Assets in the Contemporary Publishing Industry*. College Park, MD, and New York, NY: University of Maryland and the Book Industry Study Group, 2020.

For one of the sidebars:

[Last name, first name of sidebar author], "[title of sidebar]," in Matthew Kirschenbaum, et al., *Books.Files: Preservation of Digital Assets in the Contemporary Publishing Industry*. College Park, MD, and New York, NY: University of Maryland and the Book Industry Study Group, 2020.

Correspondence
Please direct correspondence related to this report to Matthew Kirschenbaum (mgk@umd.edu) and Brian O'Leary (brian@bisg.org)

Table of Contents

Preface

This project started as a conversation about *Track Changes: A Literary History of Word Processing*, Matthew Kirschenbaum's 2016 study of the impact of new tools on the form and content of the works produced by authors using those tools. I'd read the book, and I'd hoped to have Professor Kirschenbaum speak at an annual meeting of the Book Industry Study Group. Schedules were not our friend that year, but a natural curiosity led him to ask what BISG did.

I explained that we were the organization that tries to make book publishing work as efficiently and effectively as it can. We talked about BISG's roles in convening people from across the supply chain, amplifying the good work of others, and solving problems that affect two or more parts of the supply chain.

As an organization founded to conduct research on behalf of the book business, BISG was immediately interested in Professor Kirschenbaum's *Books.Files* project. Our members come from all parts of the book industry: publishers, manufacturers, wholesalers and distributors, retailers, libraries, and the partner firms that serve those segments. Trying to answer questions about how the industry works, and how it might work better, is a core part of what BISG does.

Our members also use BISG as a way to engage in several areas—metadata and identification, rights, subject codes, supply chain, and workflow—that intersect with many of the questions that *Books.Files* has surfaced. Particularly in the past 25 years, as digital forms have become the defining vehicles for creating, managing, and distributing book content, BISG has worked to create standards, document best practices, and raise awareness of the issues and opportunities inherent in these digital workflows.

In this report, Professor Kirschenbaum brings many of these ideas together. That's a significant benefit for our industry, as his research shows the interconnected nature of how content is created, managed, and distributed in a digital-first age. From the outset *Books.Files* has also added a component we have not considered adequately: **preservation**, particularly over time periods well beyond what technologies have had to support to date. We were happy to help bring some of our members to discussions that broadened our understanding of the challenges we'll face trying to understand today's history 25, 50, or 100 years from now.

As you'll see, more work needs to be done to address those challenges. Some of what we must do parallels what BISG is already working on, particularly in areas like workflow. More work is needed to bring the different perspectives of industry and academia together on a more regular and productive basis. In partnership with Professor Kirschenbaum and with the support of The Mellon Foundation, we are happy to have contributed to *Books.Files*, and we look forward to building on its recommendations, so that future generations will know the book business as we do today.

Brian O'Leary
Executive Director
Book Industry Study Group

I.

Background: Why Books.Files?

DAWN OF THE DAM

Twenty years ago, word that one of the (then) Big Six was storing digital copies of texts, artwork, and marketing materials for all of its newly published books in a database—with plans to do the same for significant swaths of its backlist—was notable enough to garner multiple columns of coverage in the *Wall Street Journal* (Rose). While newspapers and magazines were embracing digitization, book publishing had lagged behind: as the article noted, in a typical publisher's office hard copy manuscripts piled up everywhere, editors still marked up drafts by hand, and even work submitted electronically by an author was still printed out and retyped by in-house word processing specialists. Books themselves were printed from films and plates which served as the intermediary between the camera-ready copy created by typesetters and the final product. Meanwhile, unsold stock piled up in warehouses until eventually the book was remaindered at which point most books went out of print for good.

Nonetheless, by the year 2000, it was clear that the publishing landscape was changing. The rise of Amazon was certainly one major catalyst: not only from the standpoint of sales, but also marketing. Publishers found themselves tasked with needing to keep digital catalog information, artwork, and chapter samples at the ready for use by online retailers. Print on demand was also on the horizon, promising a deep backlist of titles that would never go out of print and—equally importantly—never be overstocked. And, of course, ebooks were suddenly in vogue, with now-forgotten devices like the Rocket, Cybook, and SoftBook preceding the Kindle among consumers. Indeed, Simon and Schuster had just launched Stephen King's ebook experiment *Riding the Bullet*: a half-million people paid $2.50 apiece and downloaded it within a day. All of this mitigated in favor of Random House, then the largest trade publisher in the world, exploring solutions for maintaining the content of its books in digital form in perpetuity.

The Random House venture described in such novel terms by the *Wall Street Journal* article was an early exemplar of what is familiarly known as a Digital Asset Management system. Nowadays every publisher has a DAM of one sort or another, ranging from commercial en-

terprise systems complete with trade shows to (at the other end of the spectrum) a GitHub account and a Slack channel. This present-day reality is an outgrowth of larger shifts in the industry and the media landscape in which it is embedded. As early as 1999, an article in *Publishing Research Quarterly* observed that publishing "is coming to mean producing digital content which can subsequently be delivered in different media, rather than producing books or journals" (Wright 87). Repurposing content for different media or even for different formats within the same medium (an online excerpt vs. an ebook, for example) was invariably time-consuming and expensive. Much better to adopt a digital-first approach where the digital representation of the content was regarded as primary and available for adaption to any number of potential containers. In the process, workflows could be rationalized, supply chains consolidated, and publishers could exert greater control over their assets. A concomitant (and underappreciated) development was the evolution of database technology from consisting solely of alphanumeric fields to storing (or indexing) actual digital objects in the form of documents, stylesheets, fonts, images, and even multimedia like audio and video. Thereafter, there would be no element of a book that could not be warehoused and retrieved from a centralized digital store.

Servers in a data center. The "cloud" is always just somebody else's hard drive
Photograph by Baltic Servers

That was the promise, anyway. The reality, of course, was somewhat different. While DAMs are now universal there is little in the way of standardization or interoperability across the industry. Newer platforms and systems are inevitably incompatible with older ones, and migration is rarely seamless or easy. Nonetheless, it is not hyperbole to say that the impact of Digital Asset Management software has been among the most profound in publishing and printing history. Historians of print and publishing understand that there have been many disruptive technologies over the years: the printing press and moveable type to be sure, but also the advent of machine-made paper, the Linotype, and the laser printer, to name just a few. To this long list must now be added the general conversion of publishing to a 21st century media industry whose day-to-day work is that of digital content management.

Regardless of one's role in the industry—acquisitions editor, copyeditor, designer, typesetter, production manager, distributor, sales representative, and of course author—a large measure of the experience of making books now consists in sitting in front of a computer and working with digital files. This point has been documented by John B. Thompson in his interviews with industry figures: what do people in the book business do all day, Thompson asks? They create, edit, and manage digital files, his respondents tell him (2012; 352-8). Only at one particular point in the supply chain, the printing plant, does a phase shift occur, and digital files are converted into tangible objects possessed of actual weight and volume (see sidebar, "Paper, Ink, Water").

With the industry's imperative to maintain its capital assets in digital form, a change has also occurred in the nature of publishers' archives. As Laura Millar has documented, publishers' archives (along with an author's personal papers) have traditionally been an invaluable resource for literary and historical scholarship as well as journalism, biography, and other fields. Archival sources illuminate the "story behind the book," as Millar puts it, allowing insight into the creative processes of authors, artists, editors, agents, and others. The book industry itself is an important social, cultural, and economic institution whose records deserve to be preserved for the public good. For all of those reasons, research libraries have long had precedent for acquiring publishers' archives, cataloging and preserving their contents, and making these collections available to their patrons. By the time they arrived in such a setting, however, a publisher's archive would have had little bearing on its current business. The archives would be of a resolutely historical nature, the ossified paper bones of persons and projects long gone.

The book industry is an important social, cultural, and economic institution whose records deserve to be preserved for the public good

After around the year 2000, however, with the rise of digital content management, a publisher's archives became increasingly coterminous with its digital assets. Indeed, many DAMs overtly market themselves as "archival" solutions, *archive* here being a term that has entered the computer industry to denote systems aimed at long-term storage—but very different from an archive in the institutional sense of a library or repository. Moreover, the distinction between current and legacy projects was quick to collapse as books remained in print indefinitely. The digital components of the book *were* the book, with printed copies produced as needed from their digital surrogates. Printing itself was transformed by processes such as computer-to-plate (CTP) technology and high-quality digital inkjet printing. Access to the "archives" was therefore suddenly akin to asking for permission to comb through a publisher's hard drives and servers—not likely to be granted! The rise of "archives" in the form of DAMs and other content management systems muddied the question of what a relationship with a traditional archive was still good for.

Of course, publishers still generated correspondence and records and all of the traditional stuff of business archives—what Millar terms *administrative* records as well as the *operational* records tied to the actual assets of a firm. But administrative records are also now digital in nature, generally part and parcel of the same internal data systems and platforms, and (crucially) subject to the same corporate policies governing information handling. An editor might have an email from a literary luminary sitting in their inbox right next to a memo about the company picnic. The email system doesn't discriminate, and neither most likely do corporate policies governing retention of the email. Just as the digital world flattens all content to a universal stream of ones and zeroes, so too does the increasingly corporatized conglomerate landscape of publishing flatten the concept of archival value, so that all documents and records are now controlled (and often embargoed or purged) under equally restrictive protocols.

All of this is also taking place amid a period of rapid globalization for the book industry, with the geographical diffusion of its supply chain enabled—in part—by the ease with which digital files can be transmitted across continents and oceans. This too has implications for

archives and preservation. As one authority on supply chain logistics writes, "It's not like there's a control tower overseeing supply networks. Instead, each node has to talk only to its neighboring node, passing goods through a system that, considered in its entirety, is staggeringly complex. Supply chains are robust precisely because they're decentralized and self-healing" (Posner). But decentralization and local contingency do not lend themselves to collective memory.

The confluence of a massive industry shift in publishing strategies and technologies around the year 2000, together with the concomitant implications for what now constitutes a publisher's archives; the challenges thus posed to cultural heritage institutions and to scholarship; and—behind it all—the backdrop of ever-increasing corporate consolidation and casualized globalization; is thus the problem space of the *Books.Files* project and this report.

ABOUT THE PROJECT

Supported for one year by a grant from The Andrew W. Mellon Foundation (but produced over the course of two, 2017-2019), *Books.Files* was an initial exploratory project aimed at assessing the archival value of digital assets in the contemporary publishing industry for stakeholders in the cultural heritage sector (libraries, archives, and academia) as well as in the industry itself. It is not a project aimed at developing technical solutions or even recommending best practices. The scope and ambition is instead more modest but also more sweeping. More sweeping in that it seeks to assess what is and isn't unique about the present moment in the context of centuries of history in publishing, printing, and bookmaking; but more modest in that it looks to initiate conversations and uncover issues, challenges and opportunities rather than concluding or resolving them. We hope the project, and in particular this report—the project's primary public deliverable—becomes the basis of further inquiry and conversation, and further work.

Activities and Scope The cornerstone of the project was an invitational meeting (convening) held March 31-April 1 2018 at the Pierpont Morgan Library in Manhattan. Some two dozen attendees were divided roughly equally between academics (and a curator) working in fields such as publishing studies, the history of the book, and media studies, and representatives from commercial publishing, including several key influencers and thought leaders (see Participants for a complete list of attendees). The objective was to bring members of the scholarly communities that have traditionally been invested in safeguarding and studying the material traces and remains of bookmaking into contact with today's professionals in the industry. This kind of contact between academia and publishing is rare, despite the firsthand experience of many academics in their dealings with publishers as authors; most scholars have only the vaguest sense of what contemporary bookmaking actually entails. Publishers, for their part, had no idea academics interested in such things even existed, let alone had stakes in questions like archival preservation. The convening featured a series of case studies drawn from scholars' research, describing needs and concerns; other topics ranged from workflow and technology to copyright and legal considerations, as well as publishing's convergence with other media and the future of the industry. While the convening itself was an "off-the-record" occasion, the two days of conversation informs this report throughout.

Following the convening, additional research and project work consisted in several different site visits as well as expert interviews undertaken by the principal investigator. Owing to confidentiality, individual interview subjects are not identified here, but they included acqui-

sitions editors, production editors, book designers, book packagers, and printers—a dozen individuals in all. Their insights and expertise likewise inform this report throughout. The site visits, meanwhile, included the Manhattan offices of one of the Big Five, a Manhattan-based book design and book packaging firm, a print production firm located in Washington DC, and a printing plant in the Midwest. Visits ranged from an hour or two to a half-day. In each instance, conversations were conducted and the physical environment was observed.

In all of this, choices as regards participants, interview subjects, and site visits were often highly contingent and inevitably constrained. Who could make an introduction to whom, who returned an email or a voicemail message, and so on, determined participation and representation to no small degree. From the outset, we decided to look to books (specifically) as opposed to newspapers, magazines, calendars, and innumerable other publishing formats. Within the domain of books, the focus has been on trade fiction and non-fiction. We did seek some balance between smaller and independent presses and the Big Five, but with a limited budget and limited time the range of participants cannot claim to represent all of publishing or all of book publishing, or even hardly all of trade publishing. It lacks representation from university presses, textbook publishing, and religious publishing. Moreover, *Books.Files* can make no claim to represent the state of industry practices worldwide, and certainly not in the Global South or developing nations. *Books.Files* must therefore be regarded as a very partial snapshot, albeit one that hopefully manages to surface some useful—possibly even generalizable—observations despite its obvious and severe limitations in scope and representation.

Participants at the 2018 convening at the Pierpont Morgan Library, New York
Photograph by Stephanie Sapienza

Audience The primary audience for this report is two-fold. First, professionals in the publishing industry. Our sense is that questions of archival preservation—and certainly questions of posterity and the cultural record—are beyond the purview of most involved in the day-to-day business of making books. Therefore, we hope that *Books.Files* reaches an industry audience for whom it raises awareness and provokes some useful discussion. It will be of interest to a publisher's production managers, chief technology officers, editors, designers, and legal counsel, as well as in-house record keepers.

The report simultaneously aspires to be useful to archivists and collection development personnel from the range of academic and research library institutions who have historically had a stake in collecting publishers' archives. While the digital preservation and archival community has had successes in working with born-digital content in an individual's papers—for example, an author's manuscripts in the form of documents saved to hard drive—the digital workflow that begins once a manuscript leaves the author's hands is neither well understood

nor addressed by existing standards and practices. It also seeks to speak directly to scholars working in the field known as the history of the book, where interest in contemporary publishing has tended to lack behind research into earlier periods—due in no small part to the opacity of the industry and general lack of accessibility to a contemporary archive.

Before questions of preservation and access can be effectively addressed there is the need for further basic contact and communication between publishers, academics, and archivists. The most important initial step therefore consists not in tool development or prescriptive recommendations, but rather facilitating ongoing encounters between these two constituencies. Doing so will allow publishers to hear from the scholars about what kinds of content scholars might want to see preserved and have access to, and allow the scholars to understand what is and is not within the realm of the possible given legal matters, workflow efficiencies, and other considerations. With this report, *Books.Files* seeks to inaugurate that conversation.

Before preservation can be addressed there is a need for further communication between publishers, academics, and archivists

Personnel and Support The principal investigator of the project and primary author of this report is Matthew Kirschenbaum, Professor of English and Digital Studies at the University of Maryland. Errors, mistakes, or misunderstandings are his alone. Kirschenbaum was further assisted by Brian O'Leary, Executive Director of the Book Industry Study Group. (See About the Investigators for more information.) Additionally, some project participants contributed individually credited sidebars to this report.

Supported by a grant from The Andrew W. Mellon Foundation, *Books.Files* is a collaboration between the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland (in College Park), and the Book Industry Study Group (BISG), based in Manhattan. Founded at the University of Maryland in 1999, MITH is a leading digital humanities center that pursues disciplinary innovation and institutional transformation through applied research, public programming, and educational opportunities. The Book Industry Study Group, Inc. is the industry's leading trade association for information, standards and research.

[SIDEBAR]

Collecting Born-Digital Material Within Publishers' Archives: A Curatorial Perspective

Dr. Karla Nielsen, Ph.D., MSLIS

Although considered "collateral evidence" to more conservative descriptive bibliographers, publishers' archives contain some of the most exciting and compelling research material for book and literary historians and for a new generation of critical bibliographers.

Collecting institutions find collecting publishers' archives inherently challenging just with respect to the analog materials. Many of the factors that impact how repositories can collect publishers' records apply to both analog and digital materials. The collections tend to be large—up to thousands of linear feet in physical form—and require an equally large investment of storage space and archival processing time. For the most part the records of larger firms have been collected by well-resourced special collections libraries at prestigious R1 universities: with the papers of Richard L. Simon and M. Lincoln Schuster, Harper & Brother/Harper & Row, and Random House at Columbia University, Henry Holt and Charles Scribner's Sons at Princeton and Alfred A. Knopf at the Harry Ransom Center at the University of Texas at Austin.

The archives sometime come to the library in several accessions over many decades, which means that archival processing norms differ from the oldest to the most recently processed materials, resulting in uneven description across the collection, and metadata migration challenges. For example, it is difficult to use optical character recognition (OCR) on a typewritten finding aid for conversation to encoded archival description (EAD) when the document has been heavily annotated for decades. Research interests also change over time, and the materials considered interesting to scholars today, such as ledgers documenting printing costs or sizes of print run, could have been excluded either by the creators, curators, or archivists at earlier moments.

Publishers' archives contain the work of many different creators: some in-house, some clients, some contract workers

Moreover, the archives represent the work of the business only as evenly as a company's record management policies and the individual employees' filing habits permit. Publishers' archives contain the work of many different creators: some in-house, some clients, some contract workers. Sometimes the work of entire departments, for example Production, or of key individuals, are not represented in an archive. Some editors take their correspondence with them when they leave. Sometimes the papers of the company's founders or prominent editors or designers are collected separately, even by different repositories.

There are many format types in these collections even on the analog side: galleys, art, marketing brochures, correspondence, internal reports and memos, print orders, sales reports. On the born-digital side, a variety of file formats are represented as well, many of them proprietary: Quark, InDesign, WordPerfect, MSWord, Excel, Filemaker Pro. Because of the emphasis on design work, many people in publishing use Macintosh computers rather than PCs. Design files are often large and exist in many variants, and the scholarly significance of the variants can be difficult to determine without an informed person opening each file, especially if the files names are not descriptive.

Future scholars will surely be interested in the proliferation of published outputs at the turn of the 20th century, such as ebooks and audio books, but because these were emerging media the preservation standards trailed by decades. It took the library profession some time to understand that it was more important to preserve degrading magnetic tape than the "brittle" acidic nineteenth-century paper that was a cause celèbre during the 1980s. Many of these materials arrive at repositories needing immediate remediation for which there may not be immediate funding.

If hard drives are brought in, the files must be weeded before commitments are made for digital preservation. Libraries also take in floppy disks, CDs—the entire gamut of outmoded digital storage formats. It can be difficult to

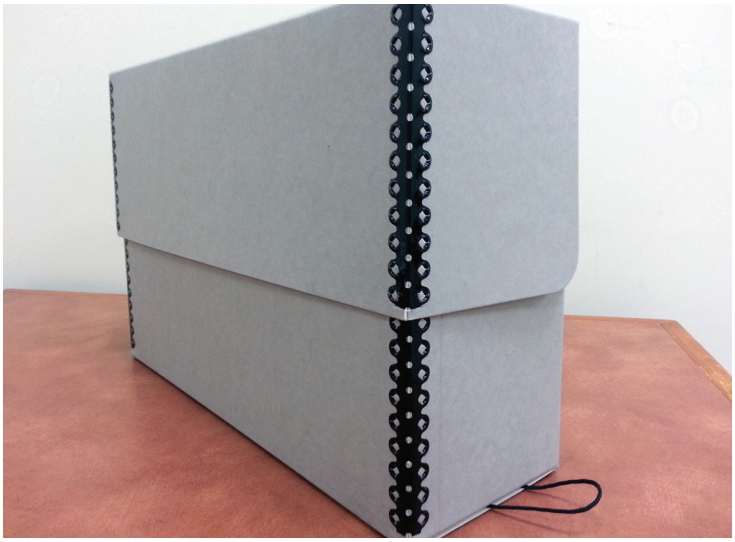
retrieve relevant files from the laptops or desktops of individuals who did not save them centrally. Many repositories are just creating the workflows for appraising digital files but selective deselection must be done lest digital storage costs become prohibitive. Over time digital storage will become less expensive, but as with analog materials we anticipate that storage space will remain an issue, especially because digital preservation best practices require retaining multiple copies of each file selected for retention.

Email provides an opportunity but also a challenge. Arguably, the discursive space that email occupies between often more formal snail mail communication and conversation is a boon to publishing historians of the last twenty years. Publishing is a business that relies on building trust and relationships, and much of the evidence of those ties is lost because there is no record of in-person meetings or phone conversations. Some of the evidence would be recorded over email.

Many creators will not turn over email without assurance that they will be able to review what is made accessible. Many people use email in a way that blurs personal and professional roles and have not retained a memory of what their email client has archived. When turning over archival material, the prospect of unexpected exposure or even embarrassment is always there and email pronounces those concerns for many creators. The collecting institution may have to commit to working iteratively with the company to decide what to make public.

Research libraries are collecting email, and can export files from a variety of web-based email platforms, but not all collecting repositories have the capacity to acquire email, and even fewer have created scalable methods for making email accessible to researchers. A notable exemplar is the email archive of the Carcanet Press at the University of Manchester (see sidebar "User Experience"). This is a boutique project documenting a small literary press and it remains to be seen which of its work is scalable for larger presses and larger corpuses of material, but it's an exciting glimpse.

There is also the matter of what materials the company is willing to include in an archive open to researchers. Because of the relationship building on which publishing depends, and the long-term nature of many of the relationships, publishing houses may be uncomfortable sharing information about authors or outside vendors who did not know that they would be archived. And while authors may expect to become public figures and objects of scholarly scrutiny, many who work in publishing did not.



A Hollinger box for storing archival records
Photograph by William Denton

As the larger American publishing houses were bought by international media conglomerates in the 1970s and 80s, their perspective on the archive often changed, with companies preferring to pay to store their archives as potentially monetizable assets rather than donate them to research libraries and open them to researchers. Some of the publishers who were donating or selling their materials to collecting repositories, for example HarperCollins and Random House, ceased after these mergers. Some of the publishers have in-house archivists, but the people in those posts must prioritize internal needs to those of outside researchers, and may not have the time or authority to make materials available to scholars.

At this point, almost all publishing, even of print books, is digitally mediated. The factors that make archiving publishers' records challenging are often also those that make it worth the effort. They provide important insight into the work proceeding and informing the final product brought to the public and document the work of the array of agents who work to connect authors and readers:

editors, designers, publicists, printers, sales directors, marketers, booksellers. Special collections libraries tend to work with (and sometimes as) book or printing historians and aware of the factors that have informed the gaps in the historical record of the last 500 or arguably 800 years of publishing. Concerted efforts are being made to document the shifts to the publishing industry over the last thirty years and this brief sidebar lays out some of the factors that are shaping what will be represented.

Karla Nielsen *is the Curator of Literary Collections at the Huntington Library in San Marino, CA, which holds one particularly rich publisher's archive, the Merrymount Press. She was responsible for acquiring and stewarding a wide range of publishers' archives (Random House, Harper & Bros., Kulchur, Dalkey Archive Press, Granary Books) in her previous post as the Curator of Literature in the Rare Book & Manuscript Library at Columbia University in the City of New York.*

II.
Understanding Workflows: How a Book Is Made

THE BIRDS AND THE BEES: WHERE DO BOOKS COME FROM?

You are an author who is publishing a non-fiction trade book. You vividly remember the moment—somehow exhilarating and anticlimactic at the same time—when, about ten months ago, you sent your completed manuscript to your editor as a Word file attached to an email. For some weeks after there was an intensive back and forth, exchanging copies of that file—each of which grew longer and more convoluted file names—with your editor's edits and comments in Track Changes, your responses to the comments, your editor's responses to your responses, and so on. Finally, everything was resolved to your mutual satisfaction and the manuscript was truly done. It's not like nothing at all then happened in the interim. You received further notes and queries from the freelance copyeditor your publisher hired, queries about the index, requests to approve artwork and a jacket design, and, most thrillingly, a PDF that was a virtual simulacrum of your book, typeset and laid out just as it would appear in print. Suddenly, your own prose seemed estranged from you. The book, you realized, was no longer just "yours." Meanwhile, the marketing department at the publisher was spinning into motion, with inquiries about advance copies, your social media platform, your availability for interviews, and more. You actually found yourself looking forward to your email every day.

Still, it comes as something of a surprise when one afternoon you hear a heavy grinding of gears and a delivery truck pulls to a stop in front of your house. There's the solid thump of a good-sized package landing on your stoop. You're not expecting anything today, you think, as you make your way to the front door. Still, a tingle of excitement—*could this be it?*

The box is not from your publisher—it's from a distribution company whose name you don't recognize. You open the package and sure enough, there's your book, two dozen author's copies, just as your contract promised. The sight of them is startling. That sense of estrangement you felt when you first saw the PDF is now magnified a thousandfold. You once knew this work intimately, you think, but now seemingly not at all. The glossy sheen of the cover, the texture of the pages, the book's size, the heft of it in your hand, the way your prose

manifests in a stately typeface so different from the one on screen where you spent so many hours lingering over it; the copyright declaration in small print, the publisher's imprint, the ISBN and bar codes, other inscrutable marks and glyphs—all of these are foreign to you. In time, your prior sense of the book and this strange new artifact will merge, and you will once again think of it as your own. But how did your book, which is to say your manuscript—that gnarly Word file, filled with ugly green and red squiggles, notes to yourself and to your editor, ragged lines and orphaned pages—how did it get to be this, this wonderful but alien thing in your hands? What actually happened in that ten-month interim, that purgatory at the publisher?

Author Shannon Pufahl with her novel in the publisher's warehouse
Photograph from @RiverheadBooks (Twitter Oct. 10, 2019)

In a sentence, the answer is this: from the time the Word document landed in your editor's inbox to the time when it was delivered as a "preflighted" print-ready PDF to a contracted commercial printer your book underwent a series of operations and transformations as a collection of digital files. In the industry, these operations and transformations are known as "touches." Every time someone opens, edits, and saves a file, it is a touch. "Your" book accrued hundreds, perhaps thousands of touches—from dozens of strangers' hands—as what began as a Word file became a series of other documents and formats, such as InDesign (INDD), XML, EPUB, and PDF. Last but not least, before it became an object in your hands, your book was a set of "plates"—thin metal sheets with photographic renderings of the individual pages from which actual pages are printed on larger-sized paper sheets on a massive web-fed offset litho press. Your book also grew its own library of ancillary documentation: first, of course, your contract and other preliminaries, then, all of that email correspondence with your editor; next, more arcane forms of documentation that only publishers know or care about: transmittal forms and art logs and spec sheets, for example; marketing plans, tip sheets for sellers, and lists of names and addresses for review copies. Copies of absolutely *everything* proliferated, in email, on local hard drives, on local area networks and in-house servers, on tape backups, and in the so-called "cloud." Your book travelled the world in its digital incarnations as files: to a copy editor, to an indexer, to a jacket designer, to a typesetter or packager, to a color separation firm for its artwork, and then to a printer. There the book finally assumed its physical form (see sidebar, "Paper, Water, Ink"), leaving the printer on pallets destined for a truck or shipping container, and ultimately to a distributor (which is where the cardboard box now on your doorstep came from). All of these entities can be located more or less anywhere in terms of physical geography.

You, as an author, don't know any of this, nor do you necessarily care. Your job is just to write the books, not to make them. As people in the business sometimes say, authors don't

write books, they write manuscripts. Books come into being as the work of many hands, not just one. But whose hands exactly, and what are these hands (for the most part on keyboards, right up until the moment the book is printed) actually doing?

UNDERSTANDING WORKFLOWS

Publishers refer to their procedures for making books as their *workflow*. Workflows are what give us the artifacts and objects—the assets, which is to say the digital files—that become candidates for archival preservation. As John B. Thompson summarizes:

> The entire publishing process, from the point at which an author turns words and sentences into keystrokes which are captured in a digital file to the point at which the final book is printed, has been turned, step by step, into what we could call a "digital workflow." Behind the scenes, the book has become a digital file, a database, that is worked on and manipulated in various ways by various people until eventually it is ready to be printed, which today can also be done digitally. (2019; 256)

Thompson refers to this as a "hidden revolution" in publishing because it is all but invisible to the average reader (customer). Ebooks notwithstanding, once they are in a reader's hands most books are still just books, looking and behaving much as books have for decades if not centuries. But for an archivist assessing a publisher's operational records from the standpoint of preservation or for the future scholar or historian who wishes to take advantage of those records, the digital workflows of this hidden revolution have profound implications. To understand the prospects for preserving digital assets it is necessary to understand something about the workflows that produce them.

A recent white paper from the Book Industry Study Group entitled *Fixing the Flux: Challenges and Opportunities in Publishing Workflows* defines workflow as "the combined impact of decisions made about process, tools, and organizational structure" (3). Here "process" is glossed as the discrete steps required to accomplish a task, "tools" are generally (but not exclusively) understood as technologies, and "organizational structure" refers to people and their roles and relationships. These somewhat abstract terms are visualized as three circles, each with a portion of their internal area overlapping with the other two. The report further emphasizes the interdependent or interlocking nature of these areas: "Even a small change in one area can require big shifts elsewhere" (5). This has the salutary effect of reminding us that tools and technologies are never independent of people, and that both technology and people work in the context of agreed upon goals and outcomes.

To understand preserving digital assets it is necessary to understand something about the workflows that produce them

Looked at in just a slightly different way, we can say here that workflows are externalizations of the different roles and responsibilities in the production of a book. They are also manifestations of material constraints: technical, logistical, legalistic, and of course fiscal. Workflows are perhaps best thought of as negotiations: solutions and practices arrived at in order to maximize efficiency (and, one hopes, quality) while minimizing costs and resource

consumption. Workflows are not static; they evolve as technical, logistical, legalistic, or fiscal considerations change. For example, there was once a time when it was not unusual for an author's entire manuscript to be rekeyed by typists employed by the publisher, even if it was delivered electronically by modem or disk. This was a time-consuming and expensive part of the workflow. As Microsoft Word became a de facto standard within the industry, publishers began mandating that manuscripts be submitted in a DOC format that could go straight to production after editorial was done. This expedited one segment of the workflow, but also created new frictions and fail points: extricating the text from Word introduces its own considerations and complications.

Workflows can be delineated in different levels of detail. Most reductively, a typical publishing workflow first involves editorial interventions in the author's manuscript as submitted, followed by additions and interventions from the production and marketing departments. The production department renders the book in its authoritative version as one or more digital files which will become the basis for manufacturing it as a physical codex, as well as for the book's distribution in ebook, audio, and other formats. The marketing department creates ancillary documentation that will become part of the larger product package that is also the book, including press kits, tip sheets, excerpts, advertisements, and the like. These too subsist as digital files which must be managed and maintained.

Much of a publisher's expertise now consists in its ability to execute elaborate and precise transformations on digital files

But that is only a very high-level view. More concretely, it is individual digital files (and their formats) that make up the day-to-day reality of a workflow. Workflows revolve around files, and more specifically file types. Different file types (formats) serve different functions. Indeed, much of a publisher's expertise now consists in its ability to execute elaborate and precise transformations on digital files. Each file type supports a function that is necessary at some point in the workflow, but that necessity has a limit—other necessities then take its place. For example, during the editorial phase, content—the actual text of the book—is the focus. It is therefore important that the manuscript be in a format that lends itself to easy editing and tracking revisions, as well as sharing and other forms of collaboration. For better or for worse, Microsoft Word generally performs that role. But you can't just press "P" and print a Word file to make it into a book; in order to shape and present the text in the manner we are accustomed to in printed books it must next be transformed into another file type, one that is suitable for applications which offer robust tools for formatting and layout. For a long time the industry standard was QuarkXPress; more recently, it is Adobe InDesign.

In a typical workflow, the InDesign file is where the book takes shape as an actual book: layout, typeset text, and artwork are all brought together here, with the final product being a print-ready PDF which will be sent to the printer, wherever they may be in the world. Alternately, there is the "XML First" workflow, so called because the Word document is converted to an XML file with tagging to create an explicit representation of the book's structure and content. This XML document becomes the basis for any number of future iterations of that content, whether an eBook, a webpage, or other. The XML can itself be imported into InDesign for layout and art. Between these two basic workflows there are many variations and permutations.

But even the preceding is still a great simplification, omitting numerous individual steps and artifacts. For example, the process of handing off a manuscript from editorial to production involves what is typically known as a transmittal form, which is a structured document intended to capture the editor's vision of the project in such a way as it can be realized by formatting, art, packaging, and design. Generated by the workflow, the transmittal form thus becomes just one of many new documents (which is to say files) now closely associated with the book. Likewise, the technical process of flowing a document from Word into Adobe InDesign, or generating XML from the InDesign file (or the Word document) will have its own attendant processes and sub-processes, all of which are part of the overall workflow. Creating an ebook, likewise, has a workflow; so too does creating an accessible ebook for the visually-impaired. And so on. Some of these processes will be documented and standardized throughout the company, but others may take the form of tacit knowledge, bespoke practices evolved through individual experiences—tricks, hacks, shortcuts, and workarounds. These may never be explicitly documented but they form an invaluable part of maintaining the workflow. Ultimately workflows are only as good as the people implementing them.

An example to prove the point: in early 2020 copies of Desmond Cole's *The Skin We're In: A Year of Black Resistance and Power* were shipped to bookstores with jackets that omitted the word "Black" from the subtitle. The book, published by Doubleday Canada, is an account of racial injustices suffered by Black Canadians; "Black" is thus not only a key word in the subtitle, its absence recapitulates the very racial injustice the book seeks to document. On social media, some speculated that the omission must have been deliberate, either because the publisher wished to avoid controversy or because someone was perpetrating a deliberate act of vandalism. As reported in *Quill and Quire*, a statement from Doubleday Canada affirms that the error occurred in the process of preparing the jackets for the printer, that new jackets were printed and distributed as soon as the error was discovered, and that "at no point was there an alternative title being considered." The statement goes on to note: "We also made immediate changes to our internal processes to ensure an error like this doesn't happen again."

We may never know the precise explanation—whose hands were on the file at what point in the workflow acting out of what motivation. But whether innocuous or otherwise, the mistake resulted in harm. Doubleday's commitment to "immediate changes" to its "internal processes" is a commitment to revising workflow as we have been describing it here, ensuring (presumably) greater security over access to files and further measures for quality control.

But no publisher has as much oversight of their workflow as they might like. Much happens through contractors and third parties. The actual printing of books always happens externally, sometimes on the other side of the world (see sidebar, "Paper, Ink, Water"). Copyediting, jacket and interior design, and indexing are all typically outsourced. Indeed, sometimes the entire production process is outsourced to a contractor. Entities known as book packagers specialize in putting together the complex elements (artwork, permissions, expert writing) that certain kinds of projects—a coffee table book about battleships, for instance—require, delivering a print-ready product to the publisher on spec. All of these actors and entities are typically invisible to the reader; they are often not even credited in the published book. Indeed, often their identities will be protected by corporate policy. Each will have their own idiosyncratic file retention system. (Put more

prosaically: how often does a freelance book designer clean out her Dropbox account?) And yet, each of them creates records (in the form of digital assets) which are part of the history of that individual book, and each of is potentially a source for archival records of that book's history.

It should be obvious by this point that measures in the interest of collecting digital records for posterity do not have a set place in typical publishing workflows. For example, could the aforementioned transmittal form for some books conceivably be of interest to future researchers? Absolutely it could. But from a busy editor's standpoint, the transmittal form might be regarded as disposable once the book is launched; absent any explicit process for safeguarding it, the odds of its being locatable in five years (let alone fifty) are slim. The same is true for every individual document and artifact that the workflow produces: If the workflow itself does not include an explicit provision for archival preservation, most individual files will be treated as so much ephemera, subject to the whims and vicissitudes of individual habit. Perhaps the best (or only) prospect for survival in such instances is if it gets printed out, destined one day for off-site storage if the publisher has such a facility.

How often does a freelance designer clean out her Dropbox account?

Archivists must understand the ins and outs of workflows if they are to advise publishers as to strategies for collecting and safeguarding digital assets. Yet workflows themselves are constantly changing as new technologies come in to play, and their particulars are generally treated as guild knowledge—if not exactly top secret, then accessible only to those professionally involved in the industry. The workflow is the window onto the specific kinds of digital assets—the individual file types and different document genres—that may become candidates for archival preservation. Documenting workflows and making them available to cultural heritage professionals is thus a low-stakes way for publishers to open a conversation about their own archival records and what might be collected for posterity.

This is further in keeping with the first two recommendations of the aforementioned BISG *Fixing the Flux* report on workflows: to make them "visible" using maps, pictures, and flowcharts, "including defined start/end states and deliverables," and—crucially—to then *share* those maps and documentation (15). More than implicit is also a third of that report's recommendations, which is simply on the importance of conversation about workflow to improve mutual communication and understanding. Our emphasis here is on expanding the audience for those conversations to also include archivists, cultural heritage professionals, and scholars. If so then all of us might understand a little bit more about what a book actually is nowadays, what it takes to make one, and how to remember its history.

[SIDEBAR]

Lifecycles of Digital Files and Staff Labor at Ooligan Press in Portland, Oregon

Dr. Kathi Inman Berens, Ph.D.

Featured people:
Marina Garcia, 2nd year Project Manager
Ivy Knight, 1st year Project Manager
Jenny Kimura, Design Manager
Hanna Ziegler, Cover Co-designer
Madison Schultz, Managing Editor
Abbey Gaterud, Publisher, Senior Instructor at Portland State University English Department

The Master's in Book Publishing program at Portland State University is the only one in North America where students own and operate a full-scale trade press that publishes four books annually. These are distributed by Ingram Publishing Services, one of the largest book distributors in the world; an Ooligan 2nd-year project manager leads the sales call to pitch her book to Ingram. Ooligan books, available anywhere books are sold, win regional awards (such as the Oregon Book Award, twice), and starred reviews in *Publisher's Weekly* and *Kirkus*.

Ooligan Press staff turnover is 50% annually because of graduation. In the spring, first-year students apply for management roles that start over the summer. Management positions radically expand a student's responsibility for making, moving, storing, bundling, and archiving digital files. At a weekly press-wide meeting in the spring, second-year managers answer timorous questions from the first-years applying to replace them. "What do I need to know to be a manager?" One asked. A manager replied: "Every book is on fire."

In an environment where digital files are urgently needed until launch and rarely needed after it, file "housekeeping" isn't a top priority. The team that launches a book becomes immediately responsible for the next acquisition. In general, publishers spend a lot of energy on the next new thing and don't always make time to tidy up digital files into archival folders. At Ooligan, populating the book's archival folder follows a precise checklist and is the project manager's final responsibility at the end of the book's production cycle. Publisher's assistants facilitate the archiving process by setting up the folder and, at the end, confirming that the folder is complete.

Book publishers work extensively with digital files, and they benefit from both project management software and in-house protocols (such as naming conventions) that facilitate ease of finding things amid the hundreds of files associated with any given book. At Ooligan Press, Google Drive stores most of the files and Trello organizes them into a visual display. Trello acts more as a directory pointing to Google Drive than file storage itself. Book files are made using Word, InDesign, Photoshop, Illustrator, XML code and MOBI (for ebooks). These files are "final" products each of which represent the composite of often dozens of drafts, whether it's four final iterations of a book cover (front cover hi-res tif, CMYK; front cover hi-res jpg, CMYK; front cover web-optimized jpg, RGB; full cover hi-res tif, CMYK), or the compiled manuscript versions in Word, or the various interior design files, or the approximately 2500 emails exchanged about a book.

Ooligan Press has a *gratis* entry-level enterprise account with Trello, which Publisher Abbey Gaterud likens to a "bulletin board" where instead of sticky notes, Trello "cards" are attached with digital files. Trello is a vertically scrolling framework where each column represent a different part of a book's production timeline. At Ooligan, a book's Trello board templatizes workflow: when a new book is acquired, its Trello board is made from a template originally built by Gaterud. But how that template gets filled and customized depends on the practices of project team members and particularly the leadership of the 2nd-year project manager, who is ultimately responsible for the book's delivery to the printer. For *Odsburg*, a literary fiction about an anthropologist who happens upon

The Ooligan Press imprint
Image courtesy of Ooligan Press

the surreal town of Odsburg, Washington, Marina Garcia was project manager for most of the book's production cycle which culminated when *Odsburg* published on 29 October 2019. Garcia, who worked on *Odsburg* for 4.5 of her total seven school terms, designed file management with an eye toward softening the impact of staff turnover. "I stored all artifacts that I felt the team would need on Trello because it was important that the current and future Odsburg team have access to the different components for publicity and production." Garcia worked between 8-17 hours weekly for 45 weeks, a mean of 562 hours in addition to coursework.

For *Odsburg*, the current Trello board one month before launch organizes materials into the following columns, left to right: "Project Materials," "Collateral," "Weekly Assignments," "Summer 2019," "Fall 2019," and "Done." Each column has many cards, each usually specifying one task, sometimes with multiple steps and files attached. As tasks are completed, team members move the card to the appropriate new location, as when a finalized book cover is moved from the design department's board to the "Project Materials" column on Odsburg project board;

Book files are made using Word, InDesign, Photoshop, Illustrator, XML code and MOBI (for ebooks)

or when a task is moved within Odsburg from, say, "Summer 2019" to "Done."

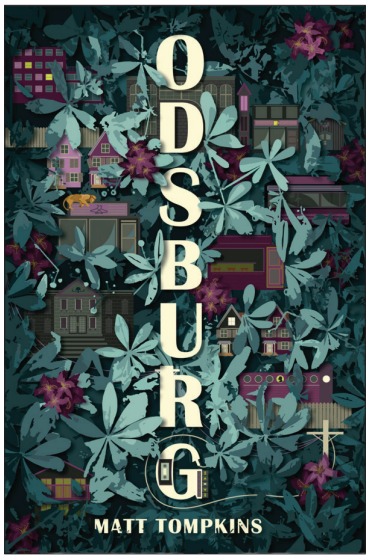
Odsburg project manager Ivy Knight, now a second-year who collaborated with Garcia, notes that how people sort objects in Trello can be idiosyncratic. "When I need a specific file that I didn't create, I don't go looking for it on a card any more. Guessing the file name and searching for it is faster and more reliable than trying to figure out what board and card it might be on." Files might start as attachments to department card (marketing, design, digital) but then need to be transferred to the appropriate card on the Odsburg board when the department has produced a "finished" file. Couple this sometimes-confusing migration with the fact that 50% of the staff is new each fall while also adjusting to a full load of master's-level academic coursework, and the Trello learning curve can be daunting. "I have so many questions about what all these different elements are that I just don't know what's all here," said one just-promoted manager. Inefficiencies are not uncommon while team members and managers are onboarding—duplicating work, for example, because unaware a resource already exists such as templates for blurb requests and review letters. From the student perspective during their first term, it's heady to go suddenly from loving books to running a full-scale publishing house. "You have the keys to the press," Gaterud announces at orientation. "Don't burn it down." (She pauses for effect and smiles: "I wouldn't let you burn it down.")

The upshot is that, at Ooligan and perhaps for many presses, working with digital files in project management software is less structured and automated than it would seem. People rely on email as a backup storage system associated with particular people based on their jobs at the press. Remembering "who touched the file" and scanning email for communications from that person can be a way to find missing stuff. "I made 4 files for my first *Odsburg* cover concept," says Jenny Kimura, Design Department manager, "and maybe 2-3 with Hanna when we combined our files. When we finalized the cover, however, we were passing files back and forth, making new files updating old ones, and I lost track of how many we sent. My email says at least 8 different files—Illustrator files, zip files, etc." Jenny spent "upwards of 75 hours on *Odsburg's* cover if you count each round, sending feedback on others' designs, and when Hanna and I finalized the design over the month of December. You'd have to ask Des how many hours they spent on shaping up the interior, but probably about 50-75 on my side of things, which included creating the special galley interior, testing out some of the found docs, giving feedback on others' found docs, and coordinating with [managing editor] Maddie and [project manager] Marina about the artistic direction."

"Design work is only ever done when you send it to the printer," observes Hanna Ziegler, who co-designed the cover with Kimura. "Until then everyone will have to touch it and give their opinion, maybe argue about something, but once the printer has it, it's officially done." Ziegler also designed several "found" documents (the novel features a number of visual elements), plus marketing collateral,

chapbook cover, and galley cover, in addition to assisting other designers on their projects.

"Timing is everything when working on a large, shifting team," notes Ivy Knight, the first-year project manager who will see *Odsburg* through to launch and archive the files. "Taking time to organize files and plan timelines is never time wasted. But researching contacts for reviews, venues, specialty markets and the like way ahead of time is a bad idea because that info can be out of date. Assets created far ahead of when they're needed have a tendency to be hard to find, especially when the person who made them just graduated."



The final jacket design for the Ooligan Press's Odsburg
Image courtesy of Ooligan Press

The greatest challenge to *Odsburg*? "We couldn't market the book as short stories because it isn't a short story collection, but we couldn't market it as a sole narrator either," says Garcia. "Needless to say, the project team became particularly adept at finding the surreal marketing spot between character-driven and town-driven marketing copy." *Odsburg's* lack of generic markers made it hard to decide upon cover specifications. At the weekly "exec" meeting, where the entire press gathers on Mondays during lunch, the press deliberated about how well the three finalists for the cover decision interpreted the design brief made by Kimura. Kimura shrunk the finalist covers down to see how they'd look as thumbnails; the press debated how emotionally light or dark the cover should be so as not to signal a kid's book. "Despite the initial struggles that surrounded this," Garcia says, "Ooligan's collaborative nature truly won out in Hanna and Jenny's work. It was amazing--their work not only represented the town of Odsburg, but also laid a foundation on which to build future visual artifacts for marketing. Sometimes to find what works, you need to talk and talk and dream, and then get a little weird. Strategic absurdity was key in producing this book."

Odsburg was particularly challenging because the "found" objects that the narrator discovers each required many drafts. Most were built in InDesign. (All master's candidates are required to learn Adobe design software, and many Ooligan Press lab students contributed to this design effort.) Some "found objects" were hand drawn and scanned. Managing editor Madison Schultz wrote on physical napkins for the segment "A Woman Walks Into a Bar." "From an editorial standpoint, my primary concern was making sure the artifacts were error-free prior to scanning them and adding them to the InDesign file, because if something handwritten had an error, we'd have to have the creator remake it, then rescan it and replace it in the file... Luckily we only had one document that needed to be modified, and it was relatively painless to do." Schultz's main jobs are guiding the book through developmental edit after acquisition, communicating those to author Matt Tompkins, overseeing a team of copyeditors, and compiling copyedits into one document returned to Tompkins for review. (A copy chief oversees the manuscript's conformity with the *Chicago Manual of Style*.) Schultz and her team use Word because its editorial and review features are more robust than those in Google Docs.

The *Odsburg* team varied in their habits of overwriting files or saving each draft. Knight saves everything; she adds a date to the end of the file name to distinguish drafts. Kimura saves items on her computer and loads only the final documents to Google Drive. Ziegler stores files everywhere: "personal computer, appropriate Trello card, Ooligan Google Drive, email."

Distributed storage between Trello, Google Drive, email, and team members' personal computers can complicate matters if down the line the publisher decides to reissue a book. This was the case when Gaterud decided to make available as print-on-demand the 2010 title *Brew to Bikes*. It had one long print run and then was made into an ebook. In that process, "we made a lot of corrections to the text, so the ebook was the most current version," Gaterud notes. "We took the ebook file and exported XML to make a new print version" which digital editors manually cleaned up for POD. While it would seem that software to automate such clean up would be

desirable, to Gaterud's knowledge it doesn't exist. "Book publishing is not the place that's going to care about [automated versioning] because books tend to not be in print that long. Archivists or librarians would care about collecting all that information along the way. But book publishers—I mean the book has one shot and it's probably not gonna make it. The vast majority of books don't go through the [versioning] process again." And if they do, it's because the book sold exceptionally well, so the cost of manually updating the files would be recouped in the second round of sales. If files go missing over the years, an email search may or may not yield the answer. "Email is a kind of backup," says Gaterud. "But we probably have 700 accounts in our Google domain, and who knows who touched something in the last few years? Time makes it much harder."

Texting via phone is a frequent, vital, and unarchived component of Ooligan's production process

Unlike many regional small presses, which run on lean staff labor, Ooligan has large labor force. PSU's approximately 60 book publishing master's candidates are required to take either Ooligan studio (four credits) or lab (1 credit—can be taken multiple times) toward their degree. Most students opt to invest much more time in Ooligan than that, since nothing sends the academic lessons home like making a book start-to-finish. Such hands-on experience also accounts for why PSU book publishing grads are in high demand: 90% of them are working in their field of choice 6 months after graduation; 30% land publishing jobs within one month.

A lot of press work gets done in informal settings while students gather at desks before class starts, or on lunch breaks between classes, or after classes before (many) students commute to paying jobs. Texting via phone is a frequent, vital, and unarchived component of Ooligan's production process. Live synchronicity allows for real-time collaboration and deliberation that cuts down on the number of emails staff would otherwise be obliged to send to each other. Oolies gather in two weekly meetings, one press-wide, one project- or department-based. Oolies work in either a department (acquisitions, design, editorial, marketing, digital, outreach) or a book project like *Odsburg*. Departments collaborate with book project teams to make the digital artifacts.

Ooligan staffers aren't just classmates; they're also colleagues running and managing a business in an industry that runs on relationships and sells objects that are very expensive to fix if something goes wrong. The love of books first brings students together in the graduate program; loyalty, trust, understanding, and mutual dependence prompts the students to tattoo themselves with the Ooligan Press logo, start new presses and freelance agencies together, and engage in a myriad of personal and extracurricular activities together. The files they make, actualized in the book itself and individually credited on a colophon, become a tangible part of who they are—and what they became—working with each other on digital files.

Kathi Inman Berens *is Assistant Professor of Book Publishing and Digital Humanities at Portland State University, where she works on digital-born bestsellers. A longtime scholar and artist in the electronic literature community, Berens joined PSU's book publishing faculty in 2015 after completing a Fulbright in Norway. Ooligan Press, a commercial trade press that publishes three books annually and distributes through Ingram, is staffed entirely by PSU book publishing Master's candidates. Ooligan books have won many awards and earn starred reviews in* Publisher's Weekly *and* Kirkus. *Berens's role is to teach students digital contexts and emergent practices in book publishing and book culture. Her essays have been published by Oxford University Press, Bloomsbury,* Hyperrhiz: the Journal of New Media, *and* the Los Angeles Review of Books, *among others.*

III.

Understanding Assets: File,Version, and Format

FILES, FLONGS, AND FILMS

Books.Files proceeds from the assumption that there is at least one simple, uncontestable fact that obtains for any book produced with commercial press processes in the last twenty years, and which will continue to obtain for the foreseeable future. That fact is this: a book is a file, which is to say it is a persistent digital asset stored in a digital repository somewhere.

But to call a book a "file" risks too tidy an impression. A book is an assemblage of digital assets, consisting, in practice, of multiple files and formats collected in a digital asset management system, or DAM. These files contain artwork, fonts, stylesheets, metadata, and of course the text. A book is thus also a network, since these digital assets must be orchestrated to interact with one another in structured and predictable ways in order to generate desired outputs, such as an EPUB or an XHTML file. It follows from this that a book in the tangible sense—the thing that we actually hold—is in fact one medium-specific output derived from an interrelated network of digital assets. A "book" is thus the born-digital potential for a file to become a book first, and a physical, tangible object in our hands only secondarily. Every new book on our shelves has its shadow in a digital file, or more precisely a set of digital files consisting of the various assets needed to bring the book into being. A physical book nowadays is a surrogate for a digital master. Put still another way, the bookmaking process today bears more resemblance to 3D printing—which uses hot liquid filament ejected from a computer-controlled nozzle to fabricate physical objects based on specifications in a digital file—than it does to books printed using hot metal, when slugs would be melted back down at the end of a job, or (for letterpress) the forme broken down and the type distributed back into its case.

Because type was expensive and almost always in short supply, no printer could afford to leave it "standing" (that is, locked in a chase and ready to print) once an edition was done. To reprint a title months or years later therefore meant starting the typesetting process all over again from scratch, one letter at a time. By the early nineteenth century, however, techniques existed for creating a papier-mâché mould—called a flong—of a forme of type in order to produce a metal plate—a stereotype—with the text of the entire page preserved

in relief. The stereotype plate would then serve as the basis for subsequent printings. (Our modern use of the word "stereotype" derives from this notion of an object making copies of itself.) The point is that printers have long understood the advantage of keeping print-ready settings of type in formats that could be stored indefinitely.

With the rise of photo-offset printing in the twentieth century, film replaced stereotype plates. But the principle remained the same: a printed book is a derivative object created by a set of mechanical processes referencing a representation of the book in its idealized form.

A flong is made from a forme of type; the stereotype printing plate will then be cast from the flong
Photograph courtesy of Deutsche Fotothek

Digital files are the films and flongs of today. The collection of digital assets stored is in the DAM is the master from which the book is made (even if it in turn yields a plate for the press). But digital files differ from flongs and films in the most fundamental ways. They are not physical objects but symbolically coded assemblages. Any given sequence of such symbols has no inherent meaning. An 8-bit string of ones and zeros could represent an ASCII character or the color value of a pixel or an arithmetical operator—indeed, it could represent all of these and more in independent contexts. Abby Smith puts it this way:

> When all data are recorded as 0's and 1's, there is, essentially, no object that exists outside of the act of retrieval. The demand for access creates the "object," that is, the act of retrieval precipitates the temporary reassembling of 0's and 1's into a meaningful sequence that can be decoded by software and hardware. A digital art-exhibition catalog, digital comic books, or digital pornography all present themselves as the same, all are literally indistinguishable one from another during the storage, unlike, say, a book on a shelf. (1, 6)

Once software like QuarkXPress in the late 1980s and PDF in the early 1990s was embedded throughout the industry, publishers began transmitting digital files to printers. But the files were then regarded as something for the printer to manage—a publisher would have had no more use for print-ready files than they had for actual films. As a result, when print-on-demand became a reality, publishers suddenly found themselves bereft of their own capital assets. Printers, for their part, were not in the habit of archiving and inventorying digital files in a systematic way. Files for titles that were frequently reprinted would be kept ready to hand, but most were relegated to tape backup systems, idiosyncratic and unwieldly and vulnerable to decay and obsolescence. In short, there was no easy way for a commercial printer to suddenly locate and return the tens of thousands of files that might comprise a given publisher's backlist. Publishers were suddenly faced with the necessity of controlling their own content, and so Digital Asset Management systems entered into the industry.

VERSIONS AND FORMATS

Long before there were files (or at least digital files), publishers retained designated "file copies" of every book in their list. A file copy was simply a representative copy of the book as it was published, collected and shelved as an exemplar of itself. The practice persists to this day, often furnishing the most distinctive feature of a publisher's office environment. Even amid the high-rise floorplans of one of the Big Five—cubicles and computer screens under fluorescents as far as the eye can see—one can find pleasant library-like nooks fitted out with books in cases and comfortable reading chairs. This in itself speaks to a certain kind of archival impulse, the knowledge that the publisher is, incrementally, adding to the world's collective store of words and wisdom in a way that will outlive any individual product cycle.

The file copy also once served as the publisher's most available source of a book's actual text. In an era when new printings and new editions demanded new settings of type and new films it was typically the version of record for the book in question; indeed, collectors sometimes come across file copies that have been annotated with running records of reprints and new editions, turning that individual instance of the book into an invaluable archive of the book's own history.

But while the tradition of the ceremonial file copy on a well-appointed shelf endures, the actual version of record for any book nowadays is, of course, its digital instantiation, which practically speaking is a set of files in the publisher's DAM (or sometimes stored with the printer). So, if a book enjoys a second printing and there are errors—typos—to be corrected or new information to be added, where does the definitive version of the text of come from? Whose hands, exactly, on what file (exactly) stored where (exactly) are doing the actual editing? What if there are multiple authorized versions of the book, for example British vs. American English? (To say nothing of translations.) What happens if a change has to be made at the printer at the very last minute? (It happens—maybe a typo or correction, but maybe also a color adjustment to an image.) Are there procedures and protocols in place for ensuring that the most up to date version of the file then makes its way back to the publisher? And that the publisher then duly ingests it into their own systems so that the next time the book is printed it is printed from the correct version?

File copies of books on display in the publisher's Sixth Avenue office
Photograph by Matthew Kirschenbaum

All of these questions are rooted in the extreme malleability of digital files—their susceptibility to "touches" at different times and places—and likewise, their capacity to proliferate as copies and copies of copies that may or may not be truly identical to one another. As Brian O'Leary notes (see sidebar, "Going Global"), files can generally be opened and edited at numerous different points in the supply chain. In the case of a file that undergoes an eleventh hour correction

at the printer (John B. Thompson offers such an example [2012; 356-7]) you then have different versions of the book and its text in circulation, with no guarantee that any one individual is fully in command of their relationship to one another. There are ways to mitigate this of course but at the end of the day systems are only as good as the people who oversee them. Just as bibliographers will pour over the evidence of different printings and editions to reconstruct how a critical variant was introduced into an early modern text, so one can also imagine future generations of scholars attempting to piece a book's supply chain back together so as to determine whose fingers were on the keyboard at the moment a given edit was made.

Claudia Rankine's *Citizen* offers a compelling case study. Since its publication in 2014 it has been through eighteen printings (and counting) and sold some 200,000 copies, placing it at the top of the list for its publisher (Graywolf) and making it nothing short of a blockbuster in the contemporary poetry world where sales numbers sometimes fail to even clear double digits. Subtitled "An American Lyric," *Citizen* deals head-on with questions of race and racism in American history and society. It is also a carefully designed multimodal text, incorporating color photographs and drawings, URLs to online content, and unconventional page layouts. Even the heavy weight and stock of the book's 80# matte-coated paper is significant in the context of its themes, as the reader is repeatedly confronted with the stark contrast of black type "against a sharp white background," in Zora Neale Hurston's phrase.

Page 134 of *Citizen* displays of a list of Black Americans dead as the result of police violence, the print gradually fading to white by the very bottom. On at least five of the occasions when *Citizen* has been reprinted, Rankine has intervened to add the names of additional victims. A tweet from one bookseller remarks: "I am crushed by the changes made in every new printing of Claudia Rankine's *Citizen*" (@KennyCoble). Publishers, of course, typically try to avoid such changes, but the circumstances here are clearly exceptional.

But how, exactly, does page 134 get revised from one printing to the next? Someone edits a file somewhere, but who? And where? And what file, where? Stored on what hard drive or server? Is it Rankine herself who executes the keystrokes? Someone at Graywolf? The book's printer, Versa Press? The book packaging firm that Graywolf contracts to manage its production work? How are versions tracked and archived? Whose responsibility is it to maintain all those files, anyway? These may seem like forced or esoteric questions, but given the sensitivities of the content the idea of a "touch" takes on new meaning and significance. Here one could argue that to "touch" the book's file in this way is a profound and intimate act.

From this we can see that *Citizen*'s continued revisions, so meaningful to so many readers, are directly enabled by the publishing workflow and supply chain. The book's strong sales, coupled with the capacity of commercial printers to do smaller and smaller runs within acceptable margins (so that a new printing is undertaken each time the book sells out), creates the opening—the touch point—for Rankine to revise and update the page.[1] But the example also serves to dramatize how essential versioning and version control are to publishing as an industry now predicated upon effective digital content management. In a DAM, version control is implemented through a combination of permissions (files can be locked to different constituencies at different points in the workflow—for example, the acquisitions editor might

[1] Communication with Graywolf confirmed to me that any changes to the page are made (of course) only at Rankine's explicit behest. The changes are keyed into an InDesign file by personnel at BookMobile (the production contractor), from whence the files are transferred to Versa Press, all of this under the supervision of Graywolf. Copies of the PDFs are retained at all three locales, with the InDesign file at BookMobile.

be locked out once production takes over). Most DAMs also maintain every file's version history, meaning that users can roll back its current representation to access it in a prior state. DAMs are thus invaluable in regulating access while the file is in-house at the publisher, but every time it is outsourced to a contractor—for copyedits, for artwork, layout, printing—there is the potential for loss of control and mistakes.

Moreover, digital files are not interchangeable with one another simply because they are digital files. A file takes on properties and behaviors relationally and contextually; that is, as a function of its *format*, which is typically manifest to the end user by its suffix. Ebooks alone have a variety of different formats: EPUB, MOBI, IBA, and more. Any given "book" in a DAM may therefore exist in numerous different formats, as a constellation or network of files; consequently, the DAM must manage not only the history of a file as it evolves over time, but also the relationship of different files in different formats to one another. The book's final jacket art, for instance, may be stored as a PSD (Photoshop) file with layers intact but also as a TIFF which is a non-proprietary standard; the TIFF, in turn, may be used to derive JPEGs or PNGs that are passed to vendors for when the book is sold in online storefronts. All of these may look the same to the human eye, but they manifest very different computational affordances. Digital does not always mean interoperable, and standards and common interchange formats emerged only gradually.

There is no digital equivalent to a shoebox full of photos (or flongs) found under the eaves in the attic

Or to take another example: EPUB is currently the standard for delivery of content to ebook readers. It is used by all major platforms and vendors aside from Amazon's Kindle. EPUB is also an XML based format, meaning that it is machine-readable and also generally human readable. Because the ebook edition of a book can be continuously pushed out to consumers as a bitstream, it is easy to issue corrections and updates. (Usually, however, such changes are invisible to the consumer, who has no way of knowing which version of the text is manifesting on the screen in front of them.) Increasingly, this means that the EPUB file becomes the version of record for the book. If the publisher wishes to retain a separate format-independent rendition of the book, any changes or updates in the EPUB must then be back-propagated to the original XML in order to keep versions consistent. How exactly this happens is likely to consist in a set of bespoke practices which may or may not be fully documented even for internal use.

Given the pressures of production schedules and the day-to-day exigencies of the business, the focus of a DAM is of necessity going to be on optimizing workflows and relations with supply chain vendors, not on provisions for independent access by future archivists and scholars. Nor is it likely that such access would be welcome from the standpoint of corporate culture, where it would be regarded as intrusive or invasive. Flongs and old-fashioned stereotypes were hardly cherished objects, and those that survived did so through happenstance and the physical durability of the objects themselves. But there is no digital equivalent to a shoebox full of photos (or flongs) found under the eaves in the attic. Digital objects require constant forward migration to newer formats if they are to remain functional. (Open formats such as XML are hardly a panacea in this regard and may only prolong the inevitable.) It is unlikely that files whose formats have passed beyond utility for production will remain acces-

sible and available over the long-term time horizons archival thinking demands.

For the archivist, the challenge is to use knowledge of the publisher's workflow to understand the full range of files that are created, then make decisions not only about which files to collect, but which versions of those files and in which formats. All of this will be complicated by questions about intellectual property and corporate disclosure, as well as questions about storage, indexing, and end user access. These are daunting challenges, not only technological but also cultural and legalistic.

Here are some specific types of digital assets archivists might wish to develop strategies for collecting. Again, following Millar's distinction, these are "operational" records (directly tied to the products of the organization, in this case books) as opposed to administrative records, which would be materials related to the internal history of the organization itself. They are presented here "blue-sky" fashion, for sake of consideration and discussion, without regard for technical feasibility, organizational sensitivity, or prospects for actual access.

- Contracts (with authors, vendors, or others)
- Editorial correspondence with authors (email or other formats, including text messages)
- Manuscript drafts with editorial comments and emendations (Track Changes or otherwise)
- Production-related correspondence (email and otherwise)
- Transmittal sheets and other internal workflow documentation, including project management instruments like Gantt charts
- Jacket art, including rejected or alternative designs
- Production materials in various stages, from development to proof (InDesign or PDF files, for example)
- Spec sheets (to designers, printers, book packagers or production firms, and other contractors)
- Tip sheets, press kits, and other marketing materials
- Correspondence and contracts related to international rights
- Reviews, awards, and other materials related to the post-publication life of the book

Few of the items on this list are categorically different from the kinds of materials libraries have collected from publishers in the past (see sidebar, "A Curatorial Perspective"). Emails map to correspondence, PDFs and InDesign files map to galleys and proofs, a contract is a contract and an art sample is an art sample. What is categorically different is that the archival object is also now—and will indefinitely remain—a functional capital asset, a dynamic entity whose ownership (like a print-ready PDF) confers an ability to reproduce the work in ways a mere hard copy never did. This, coupled with the extreme secrecy and securitization that attends much of the industry now, mitigates against a publisher simply dumping such materials on an eagerly awaiting archival institution.

From the collecting institution's standpoint, it is not at all obvious what one would do with gigabytes or terabytes of digital data, possibility unindexed, some of it containing sensitive personal or legal information, much of it in proprietary and (eventually) obsolescent file formats. How would such material be cataloged and accessioned, stored and preserved? What would patron access look like? These are real questions, but questions beyond the purview of this report—the foundational premise of which is that if the conversation isn't at least begun there will be no chance of any of this history surviving at all.

[SIDEBAR]

Going Global: The Supply Chain in Book Publishing

Dr. Brian O'Leary, Ph.D.

In publishing, the supply chain includes authors, publishers, manufacturers, wholesalers and distributors, retailers, libraries, and industry partners who provide goods and services to one or more parts of the supply chain. All of these segments work together to deliver content to readers in physical and digital formats, through a variety of bricks-and-mortar, institutional, and online channels.

The notion of a supply chain is not new to publishing. What was once considered to be simply "production and distribution" has grown to include sourcing (particularly paper and printing), supplier management, procurement, and both outbound and inbound shipping. As publishers have moved to outsource elements of their operations, understanding how distribution, retailing, and returns work has grown more important.

Supply-chain management also includes demand planning, something that requires coordination across multiple parts of the industry. In the fourth quarter of 2018, a number of "big" books consumed more paper and printing capacity than expected, forcing publishers, printers, and retailers to change release dates, limit order quantities, and in some cases miss delivering on retail orders. The situation is a good, if painful, example of how the supply chain affects business success.

Over time, the supply chain has become increasingly global. Multi-national companies have acquired many national publishers. All large publishers sell rights and publish outside of their home territories. Manufacturing, notably printing, has several global centers, depending on the type of printing, degree of quality, and the extent to which human labor is needed.

The move to digital workflows has both facilitated and in some cases pushed the globalization of book publishing. Activities that were once the province of local specialists, notably composition and page layout, have either been moved in-house or migrated offshore, where costs are lower.

Color reproduction for higher-quality books has been done internationally for decades, but the advent of high-speed transmission options for digital files has accelerated the offshore production of color books for children and other audiences. Although publishers make efforts to send only final files, book production workflows are often iterative, and the truly final version of a book is often made and maintained by the printer, not the publisher.

Demand for digital formats has also led publishers to offshore the creation of EPUB files. After the Kindle was introduced in 2007, ebook purchases grew from a rounding error to 20% or more of the market. The growth caught many publishers by surprise, and they turned to offshore vendors to create ebook versions of both back-list and front-list titles. Quality control on these versions was not consistent, and the versions of record were often stored at multiple digital distributors.

In the fourth quarter of 2018, a number of "big" books consumed more paper and printing capacity than expected

Maintaining multiple versions of a single title requires coordination when updates are made, a process that is complicated by the use of offshore vendors. Over the past decade, some publishers have taken the creation of front-list digital books in house, typically using a sequential process than creates the print version first.

In the United States and some international markets, legal and market demands for accessible content have led publishers to use offshore partners to create the digital files that deliver accessible features. Unlike ebooks, which are typically faithful reproductions of a print product, accessible ebooks include content that describes navigation, the intent of images and illustration, and other assistive features. The workflows employed to create these enhanced versions vary widely.

The lack of consistent, widely employed content workflows, coupled with the globalization of both production and distribution of digital files (whether they end up as printed or digital products), challenge the industry's ability

to preserve the history of a published work. With digital-first workflows, editorial changes can and often do occur after files leave the direct control of authors and publishers.

Communications about the changes take place through email, collaboration workgroups, voice conversations, and other means, none of which are preserved or archived with the manuscript or production file. Many of the changes are made by vendors that have been contracted to do one thing—typically, print—and have added file maintenance and archiving as service offerings.

When paper-based workflows were the norm, things like editorial versions, transmittal notes, and requests for changes were documented and filed in folders, boxes, and archives. The vendor community was more limited and local, making it easier to standardize how work was done. Agreements about the storage of reference materials might have been as simple as a handshake or a phone call.

In a global supply chain, all vendors can open and edit digital versions of the books they handle

In a global supply chain, the number of vendors has grown substantially, and all can open and edit digital versions of the books they handle. To some extent, these problems carry over from poorly understood workflows that governed the creation of paper-first manuscripts. The extent to which versions of a manuscript and its related materials were preserved depended on the mindset of a particular house and the commitment of its staff to keep materials during and after the editorial process.

The move to a global supply chain has weakened the control that an editor or production manager can exert over a specific title. Relationships that had been local and negotiated have changed, largely in pursuit of lower costs and specific skill sets. The emphasis has shifted to price and conformance to schedule, without clear agreements on how work gets done. As a result, even a long-standing commitment to preservation is challenged.

Storage of interim and final files has also moved to digital warehouses maintained by printers, content aggregators, and distributors. Many printers have eliminated the use of film, giving them the ability to make copies of a book only as demand warrants it—even one at a time. The increasingly global market for books also shifts the model from "print, then distribute" (incurring costs to ship books from a printer to a retail store or consumer) to "distribute, then print". In the latter scheme, files move across borders, typically as print-ready PDFs, waiting in multiple digital warehouses. This makes every book potentially unique, as there is no master copy, just a file that can be edited whenever a publisher chooses.

Keeping up with these developments demands a sophisticated approach to supply-chain management. Paper and printing are influenced by global availability, and distribution and online sales are critical considerations when working to reach both traditional and emerging markets. With these factors in play, the value of looking at supply-chain issues as a core part of publishing is likely to grow in the next decade.

Brian O'Leary *is Executive Director of the Book Industry Study Group.*

IV.

"They Do Not See the Point of Us": Academic Interests

THE *CLOUD ATLAS* CONUNDRUM

Martin Paul Eve, Professor of Literature at Birkbeck College, University of London, was puzzled by something strange he'd noticed about the book he was reading. The book was David Mitchell's novel *Cloud Atlas* (2004), a bestseller that has won literary awards (including being shortlisted for the Booker Prize) and was subsequently made into a film starring Tom Hanks and Halle Berry. The rare cross-over title with both genuine popular appeal and undeniable literary merit—it is structured like a Russian doll built of seven nested stories each written in a different narrative voice with different genre conventions and even different dialects of English—*Cloud Atlas* has been the subject of neighborhood book clubs and doctoral dissertations alike. It is as good a candidate as any for a recent work of trade fiction whose readership is likely to continue to endure.

What Eve noticed was that not everyone who was reading *Cloud Atlas* was reading the same book. Specifically, his Sceptre UK edition of the novel differed in a number of respects from the American edition, published by Random House and also the source of the Kindle ebook he was reading in tandem. The differences were not the stuff of standard localization: as Eve demonstrates, there are variants in wording on nearly every page, some trivial but others less so; the storyline "An Orison of Sonmi~451" has the most extensive changes, and may be said to have been rewritten entirely. But no one, it seemed, had ever remarked on these differences in the dozen or so years the novel had been in print. Moreover, the text of the French translation is derived from the Random House (and Kindle) edition; but the German, Italian, and Japanese editions are translations of the UK text. The film script, meanwhile, hews to the American version. Eve argues persuasively that the differences have substantial implications for how one reads and interprets the novel as a whole. In a very real sense then, *Cloud Atlas* is not one novel but two. (As its readers will know, this is very much in keeping with some of the book's underlying themes about storytelling.)

Some might be led to suspect that Mitchell created the two different versions deliberately, as a kind of creative flourish for some clever reader like Eve to uncover. But this is not so. The

answer turns out to be more mundane. What happened, Eve tells us, was this: David Mitchell, not yet famous, was working with two separate editors, one at Random House and one at Sceptre, which is an imprint of Hodder and Stoughton, a division of Hachette UK. During the editorial process Mitchell's editor at Random House changed jobs and the manuscript was consequently sidelined there for many months. In the meantime, production of the Sceptre edition proceeded apace, with content editing and copyediting of the manuscript. Eventually Mitchell was assigned a new editor at Random House and the manuscript went through its own content and copyedits there. Neither Mitchell nor anyone else ever thought to reconcile the two (Eve 45-8). (Mitchell characterizes himself as inexperienced with publishing at the time and unconcerned with what he terms "a lot of faff" [Eve 46].) In other words, there were two different and entirely self-contained workflows.

It is not therefore as if one edition of the novel were an outright mistake or a bowdlerized or corrupted version; both have fair claim as to representing Mitchell's intentions as an author. What is interesting, however, is not just what Eve found out about *Cloud Atlas*, but how he found out about it: he sent an email to David Mitchell and Mitchell was kind enough to answer, furnishing the explanation which Eve recounts and which we have just summarized above.

But what if David Mitchell wasn't so conscientious about answering his email? What if a question arises for an author who is no longer living? (As Mitchell himself has said, he never dreamed that the book would still be in print all these years later, or that anyone would care enough about it to study it so closely [Eve 46].) That is the point at which a researcher normally turns to the archives. For a 21st century book like *Cloud Atlas* the "archive" would consist, in practice, of the digital files contained in the various publishers' respective Digital Asset Management systems.

Most likely, Eve would have attempted to make contact with the publishers; perhaps he might have eventually reached one of the several editors who had worked on the project. But it is doubtful any one person other than Mitchell has the whole story at his or her disposal, and Eve's prospects for hands-on access to the actual emails and digital files that would allow the episode to be reconstructed in detail would be close to nil—after all, publishers are not in the habit of opening up their internal systems to a random English professor who comes calling.

ACADEMIC INTERESTS

Cloud Atlas offers an unusually clear-cut instance of what is at stake in the question of the preservation of digital assets in the publishing industry. But surely it is also an extreme anomaly? In most cases, why would an academic or anyone else possibly want access to a publisher's files? Isn't the book itself, the actual artifact in the hands of readers, its own best witness? Isn't everything else just the means to that end?

The reality, of course, is that we're often as fascinated by the process of making or creating things as much or even more so than the end-product. This is a phenomenon we can observe across many forms of media and entertainment. Deleted scenes and the director's cut in film for example, not to mention bloopers reels and behind the scenes interviews; demos and alternate takes in music, the stuff of so many boxed sets; gamers, meanwhile, love so-called Easter eggs (hidden messages and puzzles left behind by the game's developers), and some may even try to peek inside a game's source code to see what there is to see. Witness also the ending of Greta Gerwig's 2019 film adaptation of Louisa May Alcott's *Little Women*, where

Jo peeks in on the pressmen working to typeset and bind her words. The viewer is treated to beautifully composed close-ups of movable type and tools for folding and gluing and sewing. The message is clear: Jo's manuscript has become a book, and witnessing the actual process of its making is the emotional climax of the film.

With books and publishing, though, we tend to see this only infrequently or in special circumstances—for example, Michael Pietsch's commentary about editing David Foster Wallace's posthumously published *The Pale King*. Some general readers may also be familiar with the academic controversies over what constitutes the authoritative text of James Joyce's *Ulysses*, or even Shakespeare's *Hamlet* or *Lear*. Scholars have long prepared what are called "critical editions" of important literary texts, where variants from different manuscripts and sources are available for comparison.

Books are as messy to make as anything else in this world, and knowing how they are made can illuminate questions beyond their covers

As such editions remind us, books are as messy to make as anything else in this world, and knowing something about *how* they are made can illuminate questions that extend far beyond their covers. Consider that the great bibliographer Donald F. McKenzie used a box of neglected accounting records from the Cambridge University Press to reconstruct certain very particular details of how books were made and printed in the first decade of the eighteenth century: chief amongst McKenzie's insights from these most mundane of documents was a practice he called concurrent printing, meaning that rather than any one given book being printed from start to finish before work on the next began, individual portions of different books were printed concurrently with one another, so that a shop was usually printing many different books at once. Bibliographical evidence having bearing on specific aspects of the printing of any one book is therefore as often to be found in *other* books printed in the same shop at the same time as it is between the boards of any one single volume.

McKenzie's own research was published in two magisterial folios from Cambridge, replete with charts, tables, illustrations, and fold-out inserts, all of which demanded the full measure of commitment and skill from the press's copyeditors, book designers, compositors, press workers, and binders. Fittingly, he dedicated his work to the typically anonymous "makers of this book."

But anyone wishing to undertake the sort of research McKenzie has performed for an earlier historical era would now face formidable technical, practical, and legalistic barriers. To offer another example alongside of Eve, Alan Galey, a scholar at the University of Toronto, has painstakingly reconstructed the complex circumstances surrounding the publication of Johanna Skibsrud's prize-winning novel *The Sentimentalists* (2009). Originally published in a limited edition by Gaspereau Press (a highly regarded small press), following the novel's receipt of the prestigious Scotiabank Giller Prize it became instantly in wide demand—but all but unavailable to the public, except in a hastily produced ebook edition. (It was eventually reprinted by Douglas and MacIntyre.)

Galey details a number of inconsistencies between the original small press edition (which included a letterpress-printed jacket) and the subsequent digital rendition of the book. These range from differences in typography to a coding error which resulted in the

epigraphs preceding each chapter being presented out of sequence. All of these factors arguably (and sometimes demonstrably) affected the reading experience and the essence of the text, resulting in an important novel reaching the majority of its readers in corrupt form. Galey details the various methods by which he arrived at these findings, including—crucially—the need to circumvent the DRM on the EPUB copy of the book in order to examine the underlying HTML code in which the text is presented (see Galey's sidebar for more on the ethical and legal considerations involved). But crucially for our purposes, he also arrives at a moment in which he relies on very specific affordances of the digital files in question in order to understand how the errors afflicting the ebook edition of the novel occurred. In particular, he notes that the cover image for the book, as displayed in the now exposed code of the EPUB file, is Sentimentalistsfinalforfilm_0003_001.jpg. This, as Galey notes, "suggests that Gaspereau was involved in the production of the original EPUB file, at least to the extent of making its typesetting files available" (233).

That there are two different versions of *Cloud Atlas* in circulation make it a more interesting book, not less

Galey's research is a meticulous, technically virtuoso exploration of the complexities surrounding contemporary bookmaking across varying platforms, formats, and publishers. Quite possibly he has arrived at a deeper understanding of the considerations involved in the publication of the novel than any individual involved in its production. Yet, when we discussed examples like *The Sentimentalists* or Mitchell's *Cloud Atlas* at the 2018 convening in New York City, some publishing representatives initially bristled: their take was that the academics were looking to sensationalize or skewer them for their errors—calling them out or issuing a scolding from the sanctity of the ivory tower. This is a sensitivity of which good researchers are well aware: the notion that, as one put it, they might be "dining out" on the professional missteps of others. But of course, this is not the case: anyone who has studied the long history of books and bookmaking appreciates and understands that mistakes and mishaps are part of the warp and woof not just of the publishing industry but of textuality itself—the fundamental condition of all written material. That there are two different versions of *Cloud Atlas* in circulation make it a *more* interesting book, not less.

"Both the practice and study of human culture comprise a network of symbolic exchanges," writes Jerome McGann. "Because human beings are not angels," he continues, "these exchanges always involve material negotiations" (3). Which is to say: we live in a fallen world. Scholars don't want to lay blame or cast aspersions. They do want to illuminate the ways in which books (no different from any other creative production) embody the ever-changing contradictions and complex motivations that govern all human enterprise. Doing so allows us to better understand not only books but the cultures and economies in which they participate. Scholars also have an interest in ensuring that the historical record is able to furnish an accurate account of why and how certain books came to be the way they are. Consider again the dust jacket for Deborah Cole's *The Skin We're In*, described above. While many or most of the jackets were replaced in a timely manner, it is possible that some few defective copies entered into circulation. A future researcher might be interested in this episode—not to dig up dirt on Doubleday Canada, but because, accidentally or otherwise, the work now belongs to

a longer history and wider context of books whose titles, jacket designs, and contents have been altered in ways that impact readership and reception.

The history of the book (also known as "book history" and "book studies," as well as *l'histoire du livre* to the French) is the name of the field where such topics are pursued. It arose as an academic discipline during the Cold War. Its practitioners study books as physical objects and cultural artifacts, as opposed to simply "texts" (the book's "content"). Practically speaking, this means an interest in everything from the technologies and conditions of bookmaking as a trade and industry from its origins to present day, the circulation and dissemination of books both geographically and temporally, and the habits of authors, readers, printers, editors, agents, and all of the other individuals involved in the lifecycle of a book.[2]

Ironically (or perhaps predictably) interest in the history of the book has grown, rather than diminished, in the era of ebooks and digital publishing. "Far from displacing sewn or glued blocks of paper, the digital era seems to have invested these objects with new glamour," notes leading book historian Leah Price (19-20). To date, however, there has been relatively little organized contact between scholars working in the history of the book and the contemporary publishing industry itself. One academic issues this lament: "They [commercial publishers] do not see the point of us" (Eaglestone 1096). A major aim of the *Books.Files* project is to begin offering a corrective.

But the "history of the book" is not just of historical—or even academic—interest. The lifecycle of a book is not always predictable, and unforeseen circumstances have been known to thrust neglected titles back into the spotlight. Originally published in 1935, Sinclair Lewis's dusty political fable *It Can't Happen Here* made Amazon's bestseller list in the weeks following Donald Trump's electoral victory; on the day of Trump's inauguration Penguin Modern Classics reissued it in a new edition. Perhaps an even starker example is Ahmed Rashid's non-fiction study entitled *Taliban: Militant Islam, Oil and Fundamentalism in Central Asia*, published with no particular fanfare by Yale University Press in 2000; but a year later, in the aftermath of 9/11, it spent five weeks on the New York Times bestseller list and would eventually move 1.5 million copies. If a publisher cannot locate the necessary digital assets when such cases arise then it has to the invest time and resources to recreate them. Moreover, with piracy, counterfeiting, and litigation major concerns it is in publishers' own best interests to ensure that records of their assets and operations are preserved in a sustainable way. Nor can future generations of publishing professionals learn from the industry if there is no knowledge base to pass on—witness the widely publicized shortfalls in paper supply and printing capacity that led to a number of front-list titles being delayed or unavailable during the 2018 holiday season. How can the industry in the future learn from its own past without a sustainable archive of its present?

How can the industry in the future learn from its own past without a sustainable archive of its present?

[2] The history of the book is a field of research defined by two authorities as the "social, cultural, and economic history of authorship, publishing, printing, the book arts, copyright, censorship, bookselling and distribution, libraries, literacy, literary criticism, reading habits, and reader response." See Ezra Greenspan and Jonathan Rose, "An Introduction to Book History," *Book History* 1.1 (1998): ix.

[SIDEBAR]

User Experience and Access to Born-digital Data Produced by Publishers: The Case of Carcanet Press

Dr. Lise Jaillant, Ph.D.

Publishers often treat their archives as rubbish. Fortunately, we still have a large amount of records from firms such as Random House or Chatto & Windus due to two main reasons: the determination of enterprising archivists to preserve materials they thought valuable; and the incentivization of publishers who were promised prestige and financial rewards if they transferred their collections to university libraries. The pressure to preserve publishers' records rarely came from scholars and other users. As a publishing historian, I stand on the shoulders of giants in my field: but these great scholars did little to gather the collections that made possible their own scholarship. With archivists in the driving seat, the question of access was often relegated to the "desirable" rather than "essential" criteria. At the University of Reading, for example, users need to ask Random House UK for permissions to consult archival documents, which severely restricts access.

This short report presents the work we have been doing to facilitate access to the archive of Carcanet, a leading poetry publisher in the UK. (Founded in 1969 by Michael Schmidt and Peter Jones, Carcanet moved from Oxford to Manchester in 1972. The press went on to build a diverse list, including poetry in translation and by neglected women poets. Among the distinguished writers associated with Carcanet are Elizabeth Jennings, Ted Hughes and many others.) It addresses two issues at the core of the "dark" archive situation: first, technical issues to make born-digital records available; and second, issues relating to the confidentiality or sensitivity of these documents. This work, funded my AHRC Leadership Fellowship (2018-2020), builds on an earlier project that aimed to bring together archivists and scholars, for which I received a British Academy Rising Star Engagement Award (2017-2018).[1] After two years of discussions and collaborative work with archivists, I am convinced that we need to move fast (and avoid breaking things). Open data respectful of privacy is possible, and the first step is to quickly build prototypes to give access to archival records.

Publishers often treat their archives as rubbish

Since the late 1970s, the John Rylands Library in Manchester has acquired the Carcanet Press archive on a yearly basis. In the past three decades, this collection has become hybrid: it is now composed of paper records but also emails and other born-digital documents. The vast majority of the paper archive is uncatalogued and closed to researchers; and the digital part of the collection is a "dark" archive, open only to a handful of staff. My AHRC-funded Project Archivist, who is based at the Rylands Library, has access to the entire collection. In Summer 2019, she prepared a selection of 200 emails that she thought would be interesting for me to see. She then submitted the selection to Michael Schmidt, the founder of Carcanet Press, for approval. Schmidt requested that some materials be closed or redacted for confidentiality reasons. The redacted selection of emails was then sent to me as a PDF, with email attachments in a separate ZIP folder.

For archivists, only basic technical skills are necessary to provide access to emails and other born-digital archives. There is no need to build a complicated system, or to buy expensive tools. Creating a PDF is enough to allow users to see content that will be useful for their research. I am not saying that this is perfect: as a researcher, I wish I could download thousands of emails and do some data analysis. But even a small selection of data is better than no data at all. Technical issues at the core of the "dark" archive problem can be easily resolved if we change our mindset and embrace imperfection. A prototype can be

[1] For more information on these projects, see: Jaillant, Lise. 'After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives'. *Archives and Manuscripts*, vol. 47, no. 3, Sept. 2019, pp. 285–304, doi:10.1080/01576895.2019.1640555.

improved over time, whereas a closed archive remains static and inaccessible.

The second issue (the confidentiality and sensitivity of some born-digital documents) can also be addressed with a change of mindset. Archivists at the John Rylands Library were understandably nervous when they gave me access to the selected Carcanet emails. Even after redaction, emails often contain information that the sender had not intended for public release. This is particularly problematic in light of the GDPR (General Data Protection Regulation) that applies to UK and European collections. But it is essential to embrace risk and trust that researchers will make good use of the data they access. And for users, it is important to respect privacy. Each time I saw CLOSED or [.....REDACTED.....] on the PDF, I wished I could see the entire message. It reminded me of the asterisks used for censored passages in early-twentieth-century books. Yet, I also realize that I would not want people to access all of my emails. Closure and redaction are reasonable measures, as long as the user is informed of the withdrawal of information.

Many libraries and archival collections are now experimenting with new systems to make their digital collections more accessible. For example, ePADD (an open-source software developed at Stanford University) is a valuable tool to discover born-digital materials, but researchers still need to travel to Special Collections to consult relevant records.[2] For archival repositories with limited staff time and funding, one solution is to create PDFs based on certain themes and to make them available to users after obtaining permissions. This is a low-tech solution that nearly all institutions could implement rapidly to respond to user needs. "Our users are crying out for faster access," argued Mark Greene and Dennis Meissner in their influential article "More Product, Less Process."[3] To resolve the "dark" archive problem, archivists need to start with the users and quickly work backwards. But unlocking born-digital data is not a one-way process. We need more collaboration between archivists and users. We also need more empathy: the ability to understand the concerns of archivists, and the needs of users[4] of born-digital collections.

Even a small selection of data is better than no data at all

Dr Lise Jaillant *is Professor at Loughborough University. She has a background in publishing history and digital humanities. She was the first researcher to access the emails of the writer Ian McEwan and her work has been recognized by a British Academy Rising Star award. She is currently Principal Investigator for a major Arts and Humanities Research Council Leadership Fellowship. This two-year project focuses on the poetry publisher Carcanet and its digital archive, which is currently closed to other researchers. For news and updates on Lise Jaillant's AHRC project, see www.poetrysurvival.com and follow her on Twitter: @lisejaillant*

[2] For more information on ePADD, see Schneider, J., et al. 'Appraising, Processing, and Providing Access to Email in Contemporary Literary Archives'. Archives and Manuscripts, vol. 47, no. 3, Sept. 2019, pp. 305–26, doi:10.1080/01576895.2019.1622138.

[3] Greene, Mark, and Dennis Meissner. 'More Product, Less Process: Revamping Traditional Archival Processing'. *The American Archivist*, vol. 68, no. 2, Sept. 2005, p. 235, doi:10.17723/aarc.68.2.c741823776k65863.

[4]For examples of collaboration, see Kirschenbaum, Matthew, et al. *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*. National Endowment for the Humanities Office of Digital Humanities, May 2009, http://drum.lib.umd.edu/handle/1903/9787; and *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. CLIR, 2010, https://www.clir.org/pubs/reports/pub149/

[SIDEBAR]

Analyzing E-books in the Age of Digital Locks: Challenges and Strategies

Dr. Alan Galey, Ph.D.

What can an e-book reveal about the history and social contexts of its making, or the collaborative nature of its construction? This is the kind of question that bibliographers have been asking—and answering—with regard to printed books for many years, and it is a viable question for ebooks as well. However, ebooks are made of code organized into files, and it is nearly impossible to answer a question like this if those files are not accessible, along with digital publishers' records generally. Scholars in fields ranging from analytical bibliography to book history to video game studies have emphasized the importance of first-hand analysis of digital objects at the level of code, and not just what we see on the screen. If we wish to understand the relationships between an ebook's form and functionality, or if we need to account for an apparent error in an ebook's construction, or if we are curious about plans for an ebook's design that may have been abandoned but left vestigial traces in the code, we will need to look for evidence that can only be found within files that are increasingly walled off behind digital locks. In this sidebar, I'll consider the challenges facing the code-level study of ebooks in a world of Digital Rights Management (DRM) systems, in which they are increasingly published with digital locks (known formally as Technical Protection Measures, or TPM) that impede direct access to them as primary evidence.

For the most part, digital locks are deployed under a narrow paradigm that assumes all possible uses of a text are knowable in advance

Ironically, digital locks themselves may be easily broken with tools that are not difficult to find on the web; the greater challenge, which is my focus here, is that *the act* of breaking TPM—or sharing the tools to do so—may fall within a grey area of copyright policy and law. Dan Burk, who works at the intersection of copyright law and digital materiality, articulates the crux of the problem: "Lacking the deliberative nuance of human agency, DRM lacks the flexibility to accommodate access or usage that is unforeseen, unexpected, or unanticipated" (2010, 231). For the most part, DRM and digital locks are deployed under a narrowly positivist paradigm that assumes all possible uses of a text are knowable and codifiable in advance. Those who study books and reading, in all their forms, know that's not true and never has been.

In the United States, the Digital Millennium Copyright Act (DMCA) broadly prohibits the circumvention of digital locks on copyrighted materials, regardless of intention, and prohibits trafficking in technologies that facilitate circumvention. Section 1201 of the DMCA provides for exceptions to the DMCA's anti-circumvention prohibitions, and those rules are revised every three years in conjunction with the Librarian of Congress. The European Union's Copyright Directive contains similar prohibitions, and it also has a mechanism for member states to establish valid exceptions (e.g. breaking digital locks on an ebook to enable screen-reading software to work, often necessary for readers with visual disabilities). However, in the EU and the United States, there has been widespread concern that even with these mechanisms, DRM nonetheless inhibits uses of digital objects that should be—and in many case, *are*—legal and protected by the doctrines of fair use and fair dealing.

As a digital bibliographer based in Canada, I do my work in a country where these issues are far from settled. From 2017 to 2019, the Canadian Copyright Act underwent a statutory review, and a Parliamentary committee travelled the country to receive feedback from stakeholders. My own 2012 study of an ebook's source code was one of many examples presented to the committee to support the idea that there are valid reasons for TPM circumvention.[1] Remarkably, the committee's final report (released

[1] Alan Galey, "The Enkindling Reciter: E-Books in the Bibliographical Imagination," *Book History* 15 (2012): 210-47.

in June 2019) recommends a balanced approached to TPM circumvention, including a non-exhaustive ("such as...") approach to enumerating reasonable exceptions to copyright. Even fierce defenders of the public domain such as Michael Geist received the report optimistically, but whether its recommendations will become Canadian law is another question—and, even so, that may be cold comfort outside the borders of my home country.[2]

So where does this leave someone who wants to sit down and dig into the code of an ebook right now, to see what they can learn? A university-based digital bibliographer wishing to examine the source code of an ebook may well know precisely *how* to access its code, but may be more uncertain as to *whether* she can do so without violating copyright law, policies of universities or funding agencies, or the terms of End User License Agreements. The stakes are even higher for those in positions of precarity, and the chilling effects of uncertainty about DRM circumvention for research purposes are very real. I'll conclude with an outline of five possible responses to this scenario that I've identified (and named in the spirit of tvtropes.org), though none of them may be adequate on their own. I also hasten to add that these are descriptions of practices, not recommendations. Scholars contemplating these kinds of strategies should always seek advice from someone qualified and authorized to provide it, such as a university copyright librarian.

1. *The "what happens in Vegas..." approach: breaking digital locks in the course of one's research, but omitting discussion of how one broke them, or acknowledging that one broke them at all.* This has the advantage of being a genuine path to knowledge about the artifact under study, and it provides the researcher with evidence that can answer many bibliographical questions. The downside is that one can't be fully transparent about one's methods, one can't do this with students (or with peer workshops like those at SHARP or Rare Book School), and one might still be breaking the law.

2. *The "Thor Heyerdahl" approach: instead of breaking digital locks on the object under study, building a replica as a kind of manipulable model, and hoping it behaves analogously to one's real object of study*. It is much easier to create an ebook using an open standard like EPUB than it was for Thor Heyerdahl to build and sail his experimental ship, the *Kon-Tiki*, and one can test hypotheses about ebooks in safer environments than the waters of the South Pacific. This approach can work quite well in an educational context, but only with relatively simple digital objects using open standards like EPUB, and conclusions based upon it must rely on probability and conjecture rather than empirical evidence.

3. *The "Spotify teardown" approach: modelling the algorithms that govern a digital system by manipulating its inputs and examining the results*. This strategy takes its name from the recent book *Spotify Teardown: Inside the Black Box of Streaming Music*, written by a group of researchers who wanted to understand the algorithms that govern Spotify's behavior as a music distribution platform. Their multi-pronged set of methods included creating their own music label for research purposes, and using it to upload files that tested Spotify's behaviors in various ways.[3] This approach can work for those interested not just in a single digital thing, like an ebook, but in systems that circulate many digital things. Disadvantages include those for the "Thor Heyerdahl" approach mentioned above, and the possibility of legal pressure from the company under study (which the *Spotify Teardown* authors—and their funding agency in Sweden—defied successfully).

4. *The "grateful lurker" approach: documenting how online communities who care about certain kinds of digital artifacts are discussing them, curating them, and sometimes breaking them open to understand how they work—and how they share their evidence online*. This strategy works especially well for video games, many of which have thriving online communities of modders: people whose work to repurpose video game engines to create new games often leads to discoveries about the original game's development process, usually in the form of abandoned design features or digital assets that

[2] Michel Geist, "The Authoritative Canadian Copyright Review: Industry Committee Issues Balanced, Forward-Looking Report on the Future of Canadian Copyright Law," June 3, 2019, http://www.michaelgeist.ca/2019/06/the-authoritative-canadian-copyright-review-report-industry-committee/

[3] Maria Eriksson, Rasmus Fleischer, Anna Johannson, Pelle Snickars, Patrick Vonderau, *Spotify Teardown: Inside the Black Box of Streaming Music* (Cambridge, MA: MIT Press, 2019), 69–78.

the developers neglected to remove from the source code. I have also adapted this approach to the study of digitally curated musical recordings, though ebooks may not benefit from the same levels of dedicated online communities.[4] The main advantage, of course, is that someone else is doing the digital lock-breaking—and they may document their methods and analysis to a reasonably high standard of evidence, sometimes even with informal community peer-review. Another advantage of the "grateful lurker" approach is that it harmonizes nicely with book history's emphasis on reception and what D.F. McKenzie called the "sociology of texts," and can shine a light on the valuable cultural heritage work done by online pro-am communities (i.e. amateurs whose work approaches or reaches a professional standard). A potential disadvantage is that not all online communities want that kind of light shone on them; following research ethics protocols for studying online communities is therefore essential.

Where does this leave someone who wants to sit down and dig into the code of an ebook right now, to see what they can learn?

5. *The "Tom Petty" approach (cf. lyrics to "I Won't Back Down"): eschewing all of the approaches mentioned above, and breaking digital locks openly and unapologetically with the expectation that one is acting reasonably within the limitations to copyright, with no intention of infringement or piracy—and then standing one's ground.* Disadvantages are obvious, as are advantages. Less obvious, but no less real, are the networks of support and advocacy for those whose scholarship sometimes requires the protection of the law.[5] In ideal circumstances, this is not so much a challenge to copyright law as an opportunity to clarify its purpose and limits. Not an approach to try alone, but then again neither is most digital scholarship.

Alan Galey *is Associate Professor in the Faculty of Information at the University of Toronto, and Director of the collaborative program in Book History and Print Culture. His research and teaching are located at the intersection of textual studies, the history of books and reading, and the digital humanities, and his current research focuses on the bibliographical study of born-digital texts and artifacts. Currently he is working on two research projects, a book-length study titled* The Veil of Code: Studies in Born-Digital Bibliography*, and a set of open-source digital prototypes titled* Visualizing Variation. *For details on his research and teaching, see http://individual.utoronto.ca/alangaley/.*

[4] Alan Galey, "Looking for a Place to Happen: Collective Memory, Digital Music Archiving, and The Tragically Hip," *Archivaria* 86 (2018): 6–43.

[5] A good place to start is Patricia Aufderheide and Peter Jaszi, *Reclaiming Fair Use: How to Put Balance Back in Copyright*, 2nd ed. (University of Chicago Press, 2018), especially their chapter "The Culture of Fear and Doubt, and How to Leave It," 1–16.

V.

An Archive of the Present: Some Recommendations

ARCHIVES OF THE PAST AND THE PRESENT

In *The Nature of the Book*, his acclaimed study of the printing industry in early modern London, the historian Adrian Johns takes us back to the narrow streets and lanes clustering around St. Paul's Cathedral where printers, publishers, and booksellers once congregated. He proceeds to deliver a rich recreation of what it would have been like to live and work in the book trade at the center of the Anglophone publishing world. Every day, tradesmen, city officials, lawyers, clergy, and even authors moved freely back and forth across the thresholds of the pressrooms; paper, ink, candles, rags, bread, and ale all flowed in, and printed sheets issued forth. Printing, Johns's reader soon learns, was about much more than just the books themselves. As he writes:

> Any printed book is . . . both the product of one complex set of social and technological forces and also the starting point for another. In the first place, a large number of people, machines, and materials must converge and act together for it to come into existence at all. . . . But the story of a book evidently does not end with its creation. How it is then put to use, by whom, in what circumstances, and to what effect are all equally complex issues. (3)

Central to Johns's premise is that much of what we now prize about books—their fixity and permanence, the expectation that the text is unchanging and that which was intended by the author—precisely what makes the example of *Cloud Atlas* so unsettling—wasn't always to be taken for granted. An author's text was often regarded as fluid and fungible as a book was printed and reprinted. Type compositors were known to improvise to accommodate their own needs and sentences were sometimes rewritten while type was still in the bed of the press. Piracy, plagiarism, and Bowdlerized editions were rampant. "The very identity of print itself had to be *made*," Johns emphasizes. "It came to be as we now experience it only by virtue of hard work, exercised over generations and across nations" (2). He concludes: "A printed book can be seen as a nexus conjoining a wide range of worlds of work" (3).

Johns was able to conduct much of that research without ever leaving the Fitzwilliam Li-

brary at the University of Cambridge. There he drew on an immense assemblage of public records, registers, ledgers, notes and notices, bills of sale, commissions, inventories, and court filings, as well as drawings, engravings, maps, and more. It is still true that printed books conjoin wide ranges—whole worlds—of work. But the prospects for the kind of archival research Johns, D. F. McKenzie, and many others have undertaken in the service of earlier historical periods is dim.

Indeed, it is possible that future generations will know more about how books were made in the centuries-old hand-press era than how they are being made right now. Any author can tell you who published their book; but how many know who *printed* it, or where it was printed, or using what machines and methods? The real question though is not why today's authors and readers don't know more about how books are made, but why they seemingly don't *need* to know anymore. Johns teaches us that in the hand-press era knowing who printed a book really meant something: it went directly to the question of that particular copy of the book's legitimacy and authority. Today that authority is unquestioned. And yet, as Martin Paul Eve notes, "We are often lulled into a false sense of security in the study of contemporary fiction, believing that the perfection of production techniques would mean that editions must be identical" (45).

A printing shop in Europe in the late 16th century: books were a nexus for "wide worlds of work"
Image courtesy of The Welcome Trust

And no wonder. Most books are now printed behind literally locked doors (see sidebar, "Paper, Ink, Water"). Visitors to production plants are rarely permitted absent specific prior arrangement. This is a function not just of the digitization of content and workflows in publishing, but structural changes in the publishing industry itself. Each of the Big Five—Hachette, HarperCollins, Macmillan, Penguin Random House, and Simon and Schuster, responsible for upwards of 80% of trade publishing in the US—is a subsidiary of a larger global media corporation. As André Schiffren and Mark Crispin Miller have contended, consolidation threatens to have a deleterious effect on the quality and diversity of what gets published. But it also subjects publishing records to the same legal strictures and information handling procedures as obtain elsewhere throughout the conglomerate's corporate hierarchy. Likewise, the globalization of supply chains means that records, to the extent they are maintained at all, are dispersed across a wide array of different actors and entities, each of which may have their own independent policies as regards records retention let alone provisions for archival access. Indeed, intense securitization is the default throughout the supply chain. (Not without reason given the prevalence of piracy for high-profile or lucrative books.)

The future of the history of books lies behind those locked doors and password protected file systems. That is, it is contained within what literary scholar Amy Hungerford has termed

"the archive of the unfolding present" (xi). But it is an archive that is very different from the ones assembled out of records collected from seventeenth- or eighteenth-century London and Cambridge, and it is different again from the way Harper and Row, Doubleday, or John Murray allowed their nineteenth- and twentieth-century archives to be acquired by (respectively) Columbia University, the Library of Congress, and the National Library of Scotland (see sidebar, "A Curatorial Perspective").

Broad swaths of publishing history—and by extension, literary, cultural, and social history—are currently at risk. Without provisions for safeguarding the digital assets that form the basis of that archival record there are innumerable stories and histories that will never be written or told—and certainly not with the richness and rigor that a Johns or McKenzie brought to earlier eras. It matters because the history of books also informs their present—and future. Surely current issues in copyright and DRM can be usefully understood through awareness of the origins of copyright in an earlier era of rampant book piracy; the reading habits of the public are of as much interest to publishers now as they were two hundred years ago (novel reading was once thought to be a particularly pernicious influence on female readers); our courts rely on past precedent (such as Joyce's *Ulysses*) when adjudicating current claims of censorship; and so forth. Without sustainable and accessible archives publishing will lose some of the most important conduits that connect it to other sectors of culture and society, from universities to libraries and museums as well as that portion of the public who cares about cultural heritage and posterity.

RECOMMENDATIONS

We know there are instances where the files behind the book—including editorial correspondence, design concepts, specifications and communications with printers, timelines and tracking sheets, data for sales agents, distributors, and all of the other assets and records that accrue around a project—might be of comparable interest to what we see in film, music, gaming, and other creative cultural industries. But what is worth preserving? Digital galleys and proofs for a novel like *Cloud Atlas* certainly, but what about for a cookbook? Or a travel guide that gets reissued annually? What about the digital equivalents of the bookkeeping records McKenzie relied upon for his forensics of the Cambridge University Press? What about email and editorial correspondence? And how to balance a commitment to posterity with reasonable concerns over privacy, security, and proprietary information? Scholars and archivists will not be able to make such determinations on their own. But we can start by articulating what is of interest and value to us. Assessing the cultural heritage value of digital assets in the publishing industry has been the focus of the *Books.Files* project and report. To that end, to conclude, we present some specific recommendations—to publishers, to archivists and scholars, and to both.

FOR PUBLISHERS:

- *Document and Share Workflows.* In keeping with the recommendations set forth in the BISG's earlier *Fixing the Flux* report, publishers can help by documenting and demystifying workflows whenever opportunities present. Practically speaking this means documenting (mapping, graphing, visualizing) and explaining their workflows, developing a culture in which such workflows are available for peer evaluation and critique, and, whenever possible, sharing such workflows with archivists, scholars, and cultural heritage professionals.
- *Locate Cultural Heritage in the Workflow.* Without some provision in workflows for the eventual relocation of assets to an external institution, it is unlikely those assets will ever

be collected and retained in the first place. Action here begins with the understanding that archiving in the sense it has been discussed in this report—shifting assets and records to external institutions dedicated to the preservation of cultural heritage for posterity and the public good—differs from "archiving" in the now commonplace sense of mere data storage. Many DAMs include provisions for "archiving" content in the latter way. That is different, however, from the expertise and function of an actual archival institution, which revolves around long-term preservation and eventual access from those external to the organization with which the records originated.

- *Involve Archivists.* Archivists have professional training which can benefit publishers in the short-term as well as over the hazy horizon of posterity. A consultation with an archivist might well reveal aspects of a workflow wherein digital assets are placed in unnecessary jeopardy, as well as potentially effective ways in which their longevity may be more reliably guaranteed.

FOR ARCHIVISTS:

- *Understand Bookmaking and Publishing Processes.* Archivists cannot collect (and scholars cannot utilize) what they do not understand. The challenges here include an often insular industry and specialized, always-evolving, sometimes proprietary technologies. Nonetheless, there is a literature documenting the ins and outs of contemporary bookmaking that can function as a starting place for archivists and scholars looking to self-educate (see this report's Works Cited list).
- *Look to Authors. Books.Files* must acknowledge that the challenges to collecting contemporary publishers' archives are daunting. Therefore, it is not unreasonable to lay even more emphasis on the importance of author's papers, where interactions with publishers may be documented to a significant extent. This recommendation extends, of course, to an author's digital "papers," including email and content stored on hard disk and removable media. The digital preservation community has seen some very reasonable success in these areas, and so this is an approach that builds upon strengths and capabilities already present.

FOR BOTH:

- *Start with Email.* Email is a promising place for publishers and archivists cooperatively to begin to address questions of posterity. Editorial correspondence with authors, in particular, is at times intimate, exciting, and important, and has a clear precedent in traditional literary archives. Email has itself been the subject of a recent Mellon Foundation report on *The Future of Email Archives* (2018), and tools exist for harvesting and packaging email for archival processing. If there is a proverbial low-hanging fruit, it is email (see sidebar, "User Experience and Access").
- *Continue the Conversation.* As stated at the outset, *Books.Files* was an exploratory study of an ultimately narrow segment of the contemporary publishing industry that drew upon evidence and observations collected from a limited range of actors and entities. Clearly there is much more work to be done. At the very least, it is our hope that this report has laid the groundwork for future conversations, and that BISG and other industry entities will be motivated to continue them. Likewise, we hope that this report motivates archives and collecting institutions to return to the admittedly daunting challenges posed by contemporary publishers' archives; and that individual scholars are inspired to undertake the kind of work that demonstrates the benefits that access to the unfolding archives of the present can bring.

[SIDEBAR]

Paper, Ink, Water: Visiting a Print Production Facility

Dr. Matthew Kirschenbaum, Ph.D.

At the print production facility, you can smell the ink in the parking lot. A massive water tower emblazoned Kendallville looms overhead. There is an American flag out front. Doors are locked and monitored by visible CCTV cameras. Everyone must sign in and surrender their electronics. The environment inside is noisy; earplugs are mandatory. Visitors must stay within yellow-ruled walking lines.

To get to Kendallville, Indiana, population 9682, you fly into Fort Wayne and then drive about 45 minutes north on State Rt. 3. Like many other places, Kendallville owes its existence to a confluence of natural resources and critical infrastructure. It was founded as a trading and transit center in the mid-19th century along what had been a buffalo migration route on land belonging to the Potawatomi, who were dispossessed by the 1833 Treaty of Chicago. Main Street, Kendallville was laid down as a plank road in 1848. Soon after, rail lines crisscrossed there. Today, an industrial park on the east side of town takes advantage of access to nearby I-69, I-80/90, and the Norfolk Southern. One of the largest commercial printers in the United States maintains two facilities there, one for offset printing and one for digital printing. In fact, there has been printing in Kendallville almost as long as the town has existed: the *Noble County Star* began in 1849 and its shop served surrounding communities with job printing and a weekly.

Inside the facility, the interior floorplan reflects the workflow. Paper and ink are stored at the back of the building, in giant rolls and barrels. "There are three things you need to print," I'm told. "Paper, ink, and water." These ingredients are fed into one of four enormous ManRoland Lithoman IV offset presses, each a $15 million piece of hardware. (The plates are kept wetted down with water to keep ink off of the areas with no image; there are 110 gallons of water in the press at any given time.) Each Lithoman is capable of printing tens of thousands of sheets (and thus hundreds of thousands of pages) per hour. Printed sheets are then dried, folded, cut, and sorted, before being shunted off to separate areas of the floor for binding and casework on equally imposing machines. Software and sensors constantly monitor and correct the print job in progress for color levels and registration. Nonetheless, the people who guide us are deeply knowledgeable themselves about paper stock, coating (sealant), ink, and so forth. They know things the software doesn't, like how to compensate for the ambient environmental effects of a humid Indiana summer and a frigid Midwestern winter. Everything is about throughput: sheets printed per hour, pages cut and bound per hour, boxes and pallets stacked per hour. A motivational sign reads: "SAVING 1 MINUTE AT A TIME."

Outside the print production facility, Kendallville, IN
Photograph by Matthew Kirschenbaum

Although the printer is only one point on the supply chain, there is some reason to think of it as exceptional: the printing plant is the place where the digital entity that is the book in the form of a file is transformed—through processes involving wetting, drying, staining, stretching,

pressing, folding, and cutting—into the physical commodity of the codex. Print-ready PDFs come into the building over fiber optic and saleable books leave it, shrink-wrapped on pallets.

On the day we visit, the plant is printing a title under embargo, meaning physical security has been tightened even more than usual. This is not an uncommon occurrence. Every once in a while, the embargo is for a high-profile title like, say, James Comey's memoir or a new Harry Potter. More often, it's for a textbook, or else a book related to the games industry. This particular book has a release date on Amazon three weeks in the future, which makes looking at it there in the present a little uncanny—were I to buy it in three weeks' time, would I receive one of the same copies I'm gazing at now?

I wanted to go to Kendallville because I wanted to see how books were made. I've spent my life with them, but I realized I didn't really know. But this and places like it around the globe—typically invisible and inaccessible to outsiders—are where the supply chain goes to ground; this is where the work gets done; this is where books are made, where bits become atoms with the help of paper, ink, and water.

Matthew Kirschenbaum *is Professor of English and Digital Studies at the University of Maryland.*

Acknowledgments

Support for this research report was provided in part by a grant from The Andrew W. Mellon Foundation. We thank program officers Patricia Hswe and Donald J. Waters for their interest and assistance. We also thank all of the convening attendees for their time, expertise, and generosity; and likewise extend our sincere appreciation to the hosts of our various site visits and to our interviewees. We thank the Pierpont Morgan Library for use of its facilities. Matthew Kirschenbaum thanks Grace Babukiika and Stephanie Sapienza at the Maryland Institute for Technology in the Humanities for their work in support of this project, and Britt Starr, doctoral candidate in the Department of English, for her research assistance. He also wishes to thank Brian O'Leary for his ongoing interest and commitment, and the sidebar authors for their contributions. Brian O'Leary thanks Maya Fakundiny, formerly operations manager at the Book Industry Study Group. Deborah Rust did the layout and design work for this report, and we are grateful for her skill.

Several paragraphs from section V and Kirschenbaum's sidebar appear in modified form in an essay entitled "Bibliologistics," forthcoming at the time of publication of this report on the Post45 "Contemporaries" (post45.org/contemporaries) website.

About the Investigators

Matthew Kirschenbaum is Professor of English and Digital Studies at the University of Maryland. He has previously collaborated with the Mellon Foundation, the Library of Congress, the National Endowment for the Humanities, and the Institute for Museum and Library Services on reports and white papers assessing born-digital content for archival preservation. He is also the author of two books, most recently *Track Changes: A Literary History of Word Processing* (Harvard University Press, 2016). Kirschenbaum is affiliate faculty with Maryland's College of Information Studies and a member of the teaching faculty at the Rare Book School at the University of Virginia. He has been a Guggenheim and an NEH fellow.

Brian O'Leary is Executive Director of the Book Industry Study Group. Prior, in nearly two decades as a consultant, O'Leary worked with a wide range of book publishers and associations, using his insight and experience to shape their strategy and operations.

Convening Participants

Kathi Inman Berens
Assistant Professor
Portland State University

Martin Paul Eve
Professor
Birkbeck University of London

Alan Galey
Associate Professor
University of Toronto

Amy Hungerford
Professor and Dean
Yale University

Matthew Kirschenbaum
Professor
University of Maryland

Kari Kraus
Associate Professor
University of Maryland

Shannon Mattern
Professor
The New School

John Maxwell
Associate Professor
Simon Frasier University

Karla Nielsen
Curator
Huntington Library

John B. Thompson
Professor
Cambridge University

Angela Bole
Chief Executive Officer
Independent Book Publishers Association

Liza Daly
Strategist

Pablo Defendini
Designer

Phil Madans
Executive Director
Hachette Book Group

Tracey Menzies
Vice President
HarperCollins Publishers

Richard Nash
Strategist

Brian O'Leary
Executive Director
Book Industry Study Group

Robb Pearlman
Associate Publisher
Rizzoli New York

Michael Shea
Senior Vice President
LSC Communications

Robin Sloan
Novelist

Works Cited and Further Reading

Works marked with an asterisk (*) are particularly useful as introductions to current industry practices in book publishing.

* Bullock, Adrian. *Book Production*. London: Routledge, 2012.

Burk, Dan L. "Materiality and Textuality in Digital Rights Management," *Computers and Composition* 27 (2010): 225-234.

Carter, Sue. "Doubleday Canada to Replace Jackets . . ." *Quill and Quire* (Feb. 5, 2020). Online: https://quillandquire.com/omni/doubleday-canada-to-replace-jackets-for-desmond-coles-new-book-after-black-disappears-from-the-subtitle/

* Clark, Giles, and Angus Phillips. *Inside Book Publishing*. 6th Edition. London: Routledge, 2019.

Eaglestone, Roger. "Contemporary Fiction in the Academy: A Manifesto." *Textual Practice 27.7* (2013): 1089-1101.

Eve, Martin Paul. *Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell's Cloud Atlas*. Stanford, CA: Stanford University Press, 2019.

Fixing the Flux: Challenges and Opportunities in Publishing Workflows. White Paper. New York, NY: The Book Industry Study Group, 2019.

The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives. Washington DC: CLIR, 2018. Online: https://www.clir.org/pubs/reports/pub175/

Galey, Alan. "The Enkindling Reciter: E-Books in the Bibliographical Imagination." *Book History 15* (2012): 210-247.

Hungerford, Amy. *Making Literature Now*. Stanford: Stanford University Press, 2016.

Johns, Adrian. *The Nature of the Book: Print and Knowledge in the Making*. Chicago: University of Chicago Press, 1998.

@KennyCoble. "I am crushed by the changes made in every new printing of Claudia Rankine's Citizen. (These pages used to be blank.)" Tweet. Jan. 6, 2015.

McGann, Jerome J. *The Textual Condition*. Princeton: Princeton University Press, 1992.

McKenzie, Donald F. *Cambridge University Press*, 1696-1712. 2 Vols. Cambridge: Cambridge University Press, 1966.

* Millar, Laura. *The Story Behind the Book: Preserving Authors' and Publishers' Archives*. Vancouver, BC: Canadian Centre for Studies in Publishing Press, 2009.

Miller, Mark Crispin. "The Publishing Industry." In Barnouw, Erik, et al. *Conglomerates and the Media*. New York, NY: NYU Press, 1998: 107-133.

* Murray, Simone. *The Digital Literary Sphere: Reading, Writing, and Selling Books in the Internet Era*. Baltimore, MD: Johns Hopkins University Press, 2018.

Phillips, Angus. *Turning the Page: The Evolution of the Book*. London: Routledge, 2014.

* Phillips, Angus and Michael Bhaskar, eds. *The Oxford Handbook of Publishing*, eds. Oxford: Oxford University Press, 2019.

Posner, Miriam. "See No Evil." *Logic* (April 1, 2018). Online: https://logicmag.io/scale/see-no-evil/.

Price, Leah. *What We Talk About When We Talk About Books*. New York: Basic Books, 2019.

Rose, Matthew. "A New Chapter: Book Publishing is Finally Entering the Digital Age," *Wall Street Journal* (May 20, 2000): R15.

Schiffrin, André. *The Business of Books*. London: Verso, 2000.

* Shatzkin, Mike, and Robert Paris Riger. *The Book Business: What Everyone Needs to Know*. Oxford: Oxford University Press, 2019.

Smith, Abby. "Preservation in the Future Tense." *CLIR Issues* 3 (May/June 1998).

* Striphas, Ted. *The Late Age of Print: Everyday Book Culture from Consumerism to Control*. New York, NY: Columbia University Press, 2011.

* Thompson, John B. *Merchants of Culture: The Publishing Business in the 21st Century*. 2nd Edition. New York, NY: Plume, 2012.

Thompson, John B. "Trade Publishing," in *The Oxford Handbook of Publishing*, eds. Angus Phillips and Michael Bhaskar. Oxford: Oxford University Press, 2019: 245-258.

Wright, James. "Exploiting Your Assets." *Publishing Research Quarterly* (Fall 1999): 84-94.