*Original Research Article*

BIG DATA & SOCIETY

Big Data & Society
July-December 2017: 1–16

DOI: 10.1177/2053951717716950
journals.sagepub.com/home/bds
SAGE

# Producing "one vast index": Google Book Search as an algorithmic system

**Melissa K Chalmers[1] and Paul N Edwards[2]**

**Abstract**

In 2004, Google embarked on a massive book digitization project. Forty library partners and billions of scanned pages later, Google Book Search has provided searchable text access to millions of books. While many details of Google's conversion processes remain proprietary secret, here we piece together their general outlines by closely examining Google Book Search products, Google patents, and the entanglement of libraries and computer scientists in the longer history of digitization work. We argue that far from simply "scanning" books, Google's efforts may be characterized as algorithmic digitization, strongly shaped by an equation of digital access with full-text searchability. We explore the consequences of Google's algorithmic digitization system for what end users ultimately do and do not see, placing these effects in the context of the multiple technical, material, and legal challenges surrounding Google Book Search. By approaching digitization primarily as a text extraction and indexing challenge—an effort to convert print books into electronically searchable data—GBS enacts one possible future for books, in which they are defined largely by their textual content.

**Keywords**

Algorithmic system, digitization, algorithmic culture, Google, web search, scanning

*Reading a public domain book on the Google Books website is a mundane encounter with text on a screen.[1] In the midst of this experience, the appearance of a hand presents an unsettling disruption (Figure 1). Positioned within the front matter of the Code of Procedure of the State of New York (1862), bright pink rubbers cover three fingers. The hand bears a thick silver ring and matching pink nail polish. The thumb has been partially erased, appearing as a brown, pixelated stripe. The words "Digitized by Google" have been digitally tattooed on the hand's skin.*

Momentarily pulling back the curtain on Google's digitization processes, the hand's presence draws attention both to the book's print origins and to the human and machine labor required to transport (and transform) it from library shelf to laptop screen. This hand belongs to a contract worker hired by Google to turn the pages of more than 20 million books digitally imaged through the Google Book Search Project since 2004. These fingers, skin, nails, and rings appear as visible traces of ongoing processes designed to obviate—and subsequently to erase—human intervention. The dream of automation persists, even as the materials resist.

The hand's ghostly presence also highlights the opacity surrounding Google's undertaking, a disjuncture between the company's techno-utopian public rhetoric and the paucity of public access it provided to the technical specifics of digital conversion. Envisioning a far-reaching public impact, Google CEO Eric Schmidt (2005) described the project's goals:

> Imagine the cultural impact of putting tens of millions of previously inaccessible volumes into one vast index, every word of which is searchable by anyone, rich and

[1]University of Michigan School of Information, USA
[2]Stanford University, USA

**Corresponding author:**
Melissa K Chalmers, University of Michigan School of Information, 105 S. State St., Ann Arbor, MI 48109, USA.
Email: mechalms@umich.edu

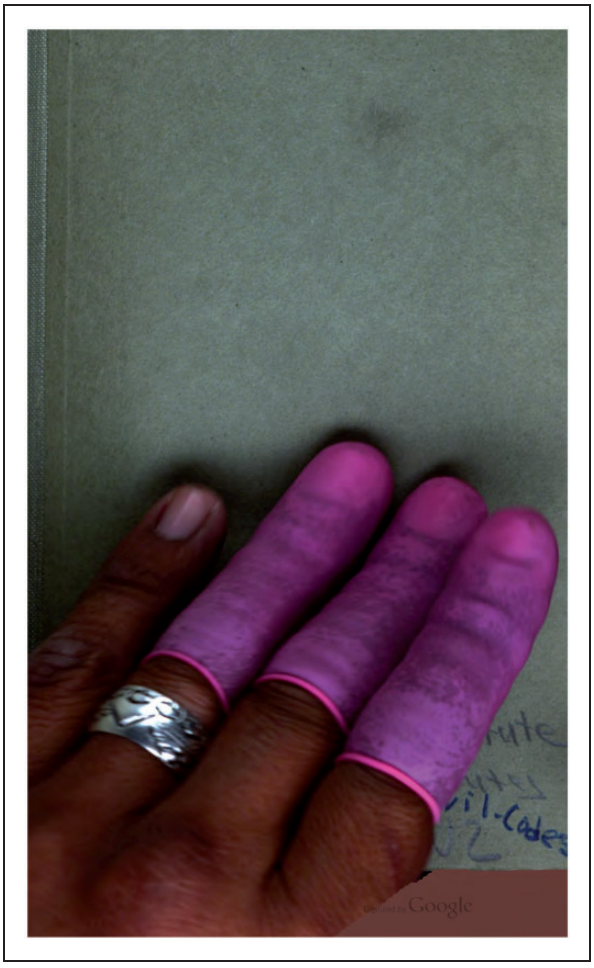


**Figure 1.** Hands scanned by Google (New York, 1862).

> poor, urban and rural, First World and Third, *en toute langue* – and all, of course, entirely for free.

Yet the actual digitization proceeded under a cloud of secrecy, leaving analysts such as ourselves to glean traces of the project's values and processes from public statements, contracts, project webpages, blog posts, presentations, and patent applications—and sometimes from the margins of the page images themselves.

Existing research has investigated many aspects of Google Book Search (hereafter GBS), including its goals, its outputs, and its intellectual property frameworks (Samuelson, 2009). Scholars have considered GBS in the context of the corporate monopolization of cultural heritage (Vaidhyanathan, 2012), the history and future of the book as a physical medium (Darnton, 2009), and the place of digitized books in knowledge infrastructures such as libraries (Jones, 2014; Murrell, 2010). Leetaru (2008) provides a rare analysis of GBS analog–digital conversion processes, while Google employees Langley and Bloomberg (2007) and Vincent (2007) have presented elements of Google's technical workflows to specialized technical research communities.

Here we take a new tack, arguing that Google's approach to digitization was shaped by a confluence of technical and cultural factors that must be understood together. These include Google's corporate commitment to the scalable logic of web search, partner selection parameters, the lingering influence of print intellectual property regimes, and the requirements of Google's highly standardized "mass digitization" processes (Coyle, 2006). This article proposes an alternative descriptor, *algorithmic digitization*, intended to highlight how the algorithms Google uses to scale and automate digitization intertwine with the production logic that governs GBS planning and execution.

Understanding GBS as an algorithmic system foregrounds Google's commitment to scale, standardized processes, automation, and iterative improvement (Gillespie, 2016). These features must also be understood as negotiated translations of varied project, partner, and corporate goals into executable workflows. We first examine how algorithms shape and structure the work of digitization in GBS and consider the effects of algorithmic processing on digitized books accessible to users. We then explore the implications of Google's embrace of an algorithmic solution to the multiple technical, material, and legal challenges posed by GBS. Beyond simply scaling up existing book digitization, Google's algorithmic digitization effort has had the effect of *reimagining* what the intended outcome of such a project should be—with important implications for mediating digital access to print books.

## Books as data: Digital hammer seeks digital nails

Google's corporate mission: "to organize the world's information and make it universally accessible and useful," has remained effectively unchanged since its first appearance on the company's website in late 1999 (Google, Inc., 1999). At the time, it referred chiefly to web search, Google's core business. In December 2004, Google announced an extension to that mission: a massive book digitization project in partnership with five elite research libraries.[2] Since then Google has worked with over 40 library partners to scan over 20 million books, producing billions of pages of searchable text. In 2012, without any formal announcement, Google quietly began to scale back the project, falling short of its aspirations to scan "everything" (Howard, 2012). While it seems unlikely that Google will stop digitizing books completely or jettison its digitized corpus anytime soon, the project's future is currently unknown.

To Google, converting print books into electronically searchable data was GBS's entire raison d'être. Therefore, Google constructed digitization as a step parallel to the web crawling that enabled web search. In contracts with library partners, Google defined digitization as "to convert content from a tangible, analog form into a digital representation of that content" (University of Michigan and Google, Inc., 2005). In practice, this conversion produced a digital surrogate in which multiple representations of a print book exist simultaneously. Each digitized book is comprised of a series of page images, a file containing the book's text, and associated metadata. Layered to produce multiple types of human and machine access—page images, full-text search, and pointers to physical copies held by libraries—each of these elements was produced by separate, yet related, processes.

## *Integrating human values—and labor—into algorithmic systems*

As with many Google endeavors, the company reengineered familiar processes at new levels of technological sophistication. From that perspective, Google's primary innovation on libraries' hand-crafted "boutique" digitization models (which pair careful content selection with preservation-quality scanning) was to approach book digitization as it would any other large-scale data management project: as a challenge of scale, rather than kind. Susan Wojcicki, a product manager for the project, contextualized Google's approach bluntly: "At Google we're good at doing things at scale" (Roush, 2005). In other words, Google turned book digitization into an algorithmic process. Scaled-up scanning required a work process centered in and around algorithms.

Algorithms are complex sequences of instructions expressed in computer code, flowcharts, decision trees, or other structured representations. From Facebook to Google and Amazon, algorithms increasingly shape how we seek information, what information we find, and how we use it. Because algorithms are typically designed to operate with little oversight or intervention, the substantial human labor involved in their creation and deployment remain obscured. Algorithmic invisibility easily slides into a presumed neutrality, and they remain outside users' direct control as they undergo iterative improvement and refinement. Finally, the vast complexity of many algorithms—especially interacting systems of algorithms—can render their behavior impossible for even their designers to predict or understand.

Embedded in systems, algorithms have the power to reconfigure work, life, and even physical spaces (Gillespie, 2016; Golumbia, 2009; Striphas, 2015). Seaver (2013) calls for reframing the questions we ask about algorithmic systems, moving away from conceiving of algorithms as technical objects with cultural consequences and toward the question of "how algorithmic systems define and produce distinctions and relations between technology and culture" in specific settings. Studying algorithmic systems empirically may thus bring together several elements: the technical details of algorithm function; the imbrication of humans (designers, production assistants, users) and human values in algorithmic systems; and the multiple contexts in which algorithms are developed and deployed.

Like many contemporary digital systems, GBS integrated humans as light industrial labor, necessary if inefficient elements of an incompletely automated process. Human labor in GBS was almost entirely physical, heavily routinized, and kept largely out of sight; human expertise resides outside rather than inside Google's system. Partner library employees pulled books from shelves onto carts destined for a Google-managed off-site scanning facility (Palmer, 2005). There, contract workers turned pages positioned under cameras, feeding high-speed image processing workflows around the clock (University of Michigan and Google, Inc., 2005). Directly supervised by the machines they were hired to operate, scanning workers were required to sign nondisclosure agreements but afforded none of the perks of being a Google employee beyond the walls of a private scanning facility (Norman Wilson, 2009). For the time being, at least, human labor in book digitization remains necessary largely because of the material fragility, inconsistency, and variety of print books.

## *Preparing to digitize: Partnerships, goal alignment, selection*

Mass digitization initiatives are often characterized as operating without a selection principle: "everything" must be digitized (Coyle, 2006). In practice, however, partnerships, scaling requirements, intellectual property regimes designed for print, and the particulars of books' material characteristics all challenged Google's universal scanning aspirations.

At the turn of the 21st century, Lynch (2002) observed that cultural heritage institutions mostly understood the *hows* of digitization, even at moderately large scale. The main challenge, he argued, was to optimize processes. Lesk (2003) described the challenges of scale and efficiency more succinctly: "we need the Henry Ford of digitization," i.e. an institution willing to invest vast resources in "digitization on an industrial scale" (Milne, 2008). Google stepped forward to assume this role.[3]

Google courted partners to provide content by incurring nearly all costs of scanning, while carefully avoiding the repository-oriented responsibilities of a library. Each partner library brought its own goals and motivations into the project. The New York Public Library (2004) observed that "without Google's assistance, the cost of digitizing our books — in both time and dollars — would be prohibitive." Other partners spoke of leveraging Google's technical expertise and innovation to inform future institutional digitization efforts (Carr, 2005; Palmer, 2005). Libraries employed different selection criteria, from committing to digitize all holdings (e.g., University of Michigan) to selecting only public domain holdings (e.g., Oxford, NYPL) or special collections (later partners). Most digitization contracts remained private, adding to the secrecy surrounding Google's efforts.

Full-text search quickly emerged as a kind of lowest-common-denominator primary functionality for the project. Using the Internet Archive's Wayback Machine, we can see how Google incrementally modified language relating to the project's goals and mechanisms throughout its first year (Google, Inc., 2004b). The answer to the question "What is the Library Project" evolved from an effort to transport media online (December 2004) to a pledge to make "offline information searchable" (May 2005) to a more ambiguous plan to "include [libraries'] collections... and, *like a card catalog*, show users information about the book plus a few snippets – a few sentences of their search term in context" (November 2005, emphasis added).

The purpose behind these changes became clear in Fall 2005, as the Authors Guild and the Association of American Publishers filed lawsuits alleging copyright infringement (Band, 2009).[4] Google argued that by creating a "comprehensive, searchable, virtual card catalog of all books in all languages," it provided *pointers* to book content rather than *access* to copyright-protected books. The company maintained that scanning-enabled indexing constituted "fair use" under the U.S. Copyright Act (Schmidt, 2005; US Copyright Office, 2016). In November 2005, the project's name changed from Google Print to Google Book Search, reorienting users' frame of reference from the world of paper to the world of the electronic web (Grant, 2005). The change attempted to correct any misperceptions that Google intended to enable access to *user-printed copies* of books and to deemphasize the idea that the project was in the business of copying or of content ownership.

Since December 2004, GBS has provided full access for public domain books. Google consistently downplayed this capability, maintaining that like a bookstore "with a Google twist," readers would use it mainly to *discover* books rather than to actually read them (Google, Inc., 2004a). Yet partners scanning public domain books often referenced online reading as a benefit. This ambiguity perhaps contributed to copyright-related concerns—and misunderstandings—during GBS's early days (Carr, 2005; New York Public Library, 2004).

## A means to an end: Image capture

Once it took custody of partner library books, Google deployed its own selection criteria. In a (rare) concession to the library partners tasked with storing and preserving paper materials, Google used a nondestructive scanning technique. In patents filed in 2003 and 2004, Google provided descriptions of several high-resolution image capture systems designed around the logistical challenges posed by bound documents.[5] The thicker the binding, for example, the less likely a book is to lie flat. In flatbed or overhead scanners, page curvature creates skewed or distorted scanned images. Book cradles or glass platens can flatten page surfaces, but these labor-intensive tools slow down scanning and can damage book spines. Google addressed this page curvature problem computationally, through a combination of 3D imaging and downstream image processing algorithms. That decision shaped and complicated Google's workflow.

In the patent schematic shown in Figure 2, two cameras (305, 310) are positioned to capture two-dimensional images of opposing pages of a bound book (301). Simultaneously, an infrared (IR) projector (325) superimposes a pattern on the book's surface, enabling an IR stereoscopic camera (315) to generate a three-dimensional map of each page (Lefevere and Saric, 2009). Using a dewarping algorithm, Google can subsequently detect page curvature in these 3D page maps and correct by straightening and stretching text (Lefevere and Saric, 2008).

Scanning produces bitmapped images that represent the pages of a print book as a grid of pixels for online viewing. Unlike text, this imaged content cannot be searched and remains "opaque to the algorithmic eyes of the machine" (Kirschenbaum, 2003). As a next step after scanning, Google might have adopted existing library-based preservation best practices for imaged content. Or it could have created new standards around 3D book imaging (Langley and Bloomberg, 2007; Leetaru, 2008). Instead, Google chose to transform the raw 3D page maps described above—rich in information, but unwieldy for end users due to file size and format—into "clean and small images for efficient web serving" (Vincent, 2007).

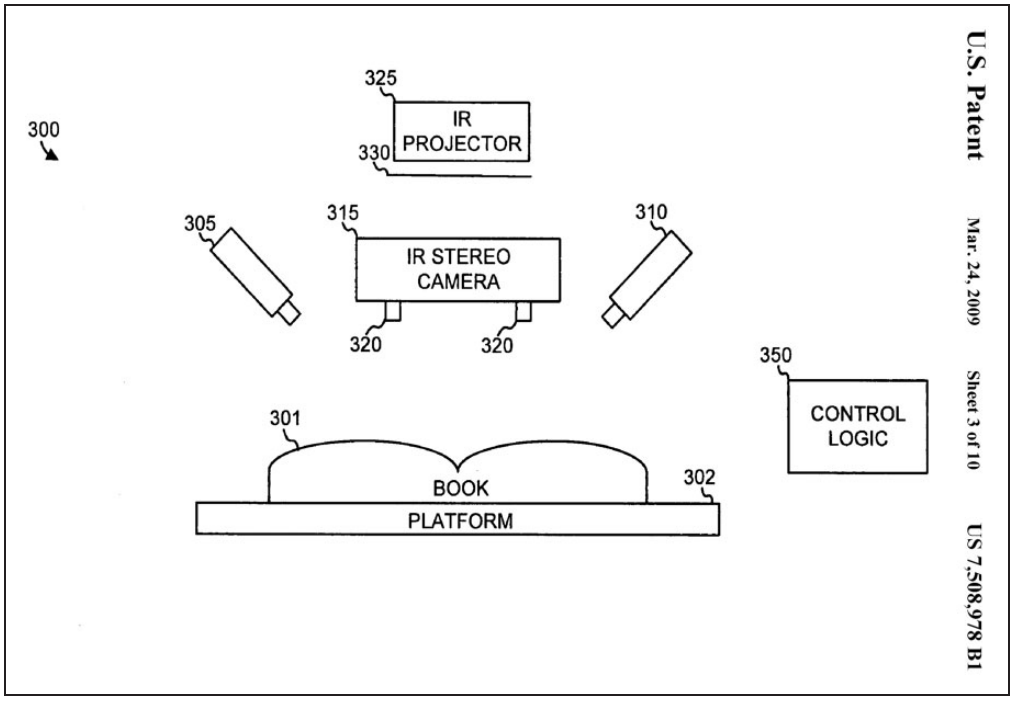


**Figure 2.** System for optically scanning documents (Lefevere and Saric, 2009).

## *Producing a machine-readable index: Image processing*

For GBS, then, imaging ultimately represented a key yet preliminary step toward text-searchable books on the web. The project's image processing workflows thus acquired a dual imperative. It had to produce both (a) two-dimensional page images for web delivery, and (b) machine-readable—and therefore searchable—text. "[O]ur general approach here has been to just get the books scanned, because until they are digitized and OCR is done, you aren't even in the game," Google Books engineering director James Crawford observed in 2010 (Madrigal, 2010). The "game" here, of course, is search. In a web search engine, crawled page content and metadata are parsed and stored in an index, a list of words accompanied by their locations. Indexing quickly became the key mechanism (and metaphor) through which Google sought to unlock the content of books for web search.

To produce its full-text index, Google converted page images to text using optical character recognition (OCR). OCR software uses pattern recognition to identify alphanumeric characters on scanned page images and encode them as machine-readable characters. Originally used to automate processing of highly standardized business documents such as bank checks, over the past 60 years OCR has become integral to organizing and accessing digital information previously stored in analog form (Holihan, 2006; Schantz, 1982). Through OCR, imaged documents gain new functionality, as text may be searched, aggregated, mined for patterns, or converted to audio formats for visually impaired users.

Tanner et al. (2009) argue that by providing search functionality for large digitized corpora at low cost, automated OCR systems have been a key driver of large-scale text digitization. GBS leveraged decades of computing research related to OCR. Through the 1990s, boutique library digitization efforts had addressed the question of quality mainly by establishing *image-centric* digitization standards (e.g., scanner specifications and calibration, test targets, resolution) (Baird, 2003). Rooted in libraries' traditions of ensuring long-term *visual* access to materials through reformatting (e.g., copying, microfilming), these practices relied on labor-intensive visual inspection for quality control. By contrast, pattern recognition research developed systems for algorithmically assessing quality, measured by accurate recognition of printed characters and document structure (Le Bourgeois et al., 2004; Lin, 2006).

Google adopted this framing of digitization as a text extraction challenge, optimizing its processes to produce the clean, high-contrast page images necessary for accurate OCR. The GBS processing pipeline relied heavily on OCR to automate not only image processing and quality control but also volume-level metadata extraction. Google's Vincent (2007) described the

digitized corpus as algorithmic "document understanding and analysis on a massive scale."

## Books bite back: Bookness as bug, not feature

In their commitment to scale and standardized procedure, algorithmic systems often prioritize system requirements over the needs of individual inputs (e.g., books) or users. Google's search engine, for example, has come under criticism for failing to prioritize authoritative or accurate search results. In December 2016, the *Guardian* reported that a Google query on "Did the Holocaust happen?" returned a Holocaust denial website as the first result. A Google spokesperson maintained that

> [w]hile it might seem tempting to fix the results of an individual query by hand, that approach does not scale to the many different variants of that query and the queries that we have not yet seen. So we prefer to take a *scalable algorithmic approach* to fix problems, rather than removing these one by one. (Cadwalladr, 2016, emphasis added)

Google's acknowledgment here of the trade-offs it faces between scale and granularity highlights questions of algorithmic accountability (Pasquale, 2015).

Google's system also exposes tensions between the standardization required to scale digitization processes and the flexibility needed to accommodate the diverse output of print publication history. It is perhaps no surprise that books, unlike business documents created to meet OCR requirements, persistently resisted the structure imposed on them by Google's homogenizing processes.

Bound books evolved over centuries from earlier writing formats such as scrolls and codices. But in Google's conversion system, the hard-won features of bound books—the very things that made them convenient, efficient, and durable media for so long—were treated as bugs rather than features. Google routinely excluded materials from scanning due to size or condition. These included very large or small books as well as books with tight bindings, tipped-in photographs and illustrations, foldout maps, or uncataloged material (Coyle, 2006). Very old, brittle, or otherwise fragile books were also excluded (Ceynowa, 2009; Milne, 2008). Many of the rejected books remain undigitized, while others have joined lengthy queues within libraries' ongoing internal digitization programs.

As a sampling process in which some, but not all, features of an analog signal are chosen for digital capture and representation, digitization is always accompanied by both information loss and information gain (Terras, 2008). In GBS, lost information includes the physical size, weight, or structure of a volume; the texture and color of its pages; and the sensory experience of navigating its contents. Nontextual book features such as illustrations, as well as marginalia and other evidence of print books' physical histories of use, are often distorted or auto-cropped out of Google's screen-based representations. As for information gain, image capture, and processing embed traces of the digitization process into digitized objects.

The quality of Google's digitization output has been systematically evaluated through empirical research and widely critiqued in informal venues such as blogs. While useful in characterizing quality concerns in the digitized corpus, this work generally does not consider how and why digitization processes shape outputs. The following examples illustrate commonly identified problems, but they also extend existing analyses by emphasizing the role of algorithms in concretizing relationships among system inputs, conversion processes, and outputs. These types of problems remain endemic in the GBS corpus not because they are unsolvable, but rather because they have been accepted as trade-offs. Their solutions do not fit easily into Google's priorities and workflows, even as their persistence challenges efforts to automate quality assurance processes.

### *Visual content*

Output-based evaluations of large-scale book digitization have found that except when catastrophic (rare), most text-oriented page scanning or image processing errors result in thin, thick, blurry, or skewed text that may frustrate or annoy readers but does not render them entirely unreadable (Conway, 2013; James, 2010). Objects such as fingers or clamps also appear commonly in scans but often do not obstruct text significantly.

In Figure 3, the very tiny book *Mother Goose's Melody* has been housed in a binder to prevent it from being lost on a library shelf. While the library-created *cover* fits Google's selection criteria and has provided a frame size for image capture, several material elements usually cropped out of Google-digitized page images have crept into the frame due to the size mismatch between the cover and the actual book. These include a call slip and university label, metal book-securing clamps, and the page-turner's hands. When extra-textual features are detected and removed algorithmically—without the help of a human eye—they often leave new artifacts behind. We see some of these less familiar traces here: the stretched appearance of book pages caused by the dewarping algorithm, and the finger incompletely removed by another algorithm. Further, the system has evidently misrecognized some

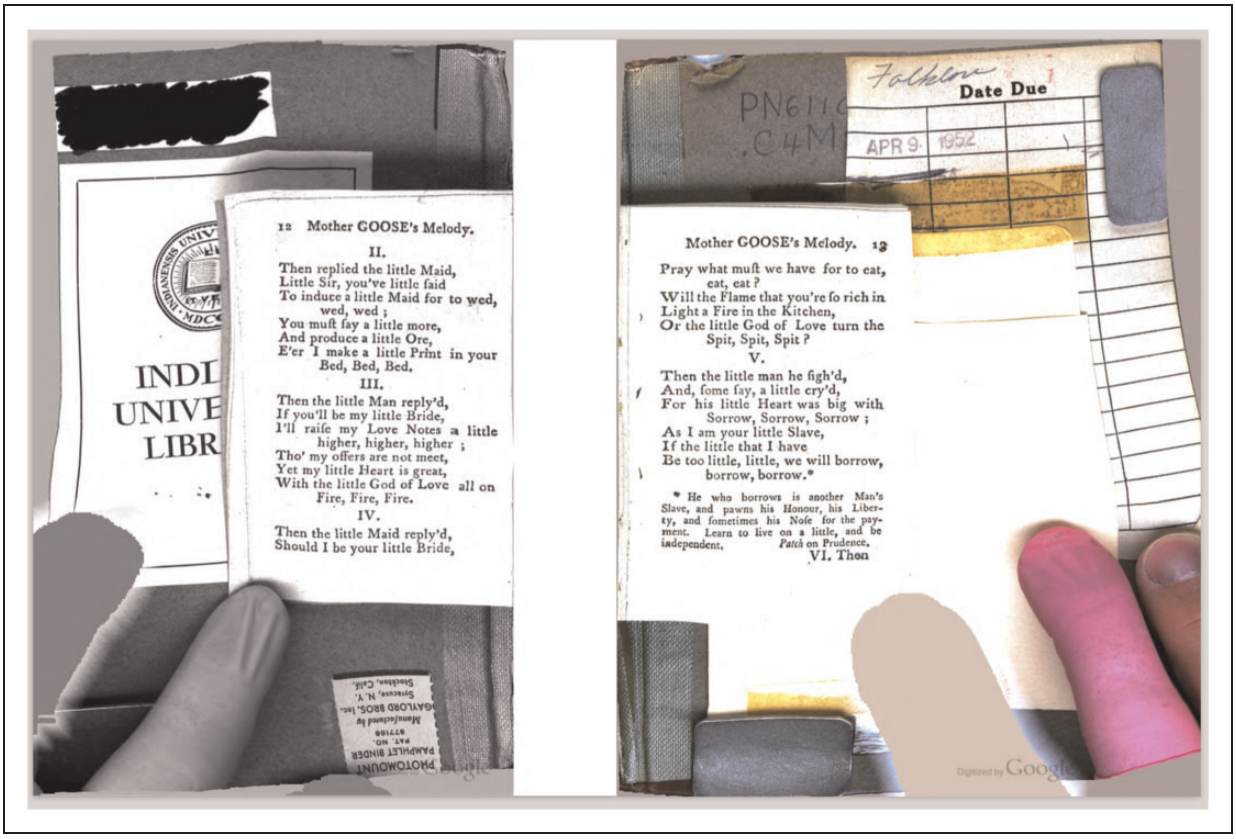


**Figure 3.** Imaging a tiny book (Thomas and Shakespeare, 1945).

aging yellow tape as a color illustration, causing most of the page images throughout the right side of the book to be rendered in color. While this book is an example of a relatively rare "bad book" (Conway, 2013), it aggregates many of the visual quality issues that pervade Google's digitized corpus.

Other material characteristics challenge image processing. These include ornate, unusual, or old fonts; non-Roman characters/scripts; and rice paper, glossy paper, glassine, and tissue paper (Conway, 2013; Weiss and James, 2015). Nontextual content such as illustrations (e.g., woodcuts, engravings, etchings, photographic reproductions, and halftones) also often fare poorly. Halftone reproductions, for example, have been widely used since the 1880s to cheaply reproduce graphic content for print. Placing a screen over an image and dividing it into squares, variably sized and regularly spaced ink dots are used to create the image; the human eye fills in the gaps created by sampling and perceives the image as a continuous tone. Computerized scanning similarly creates a digital image by sampling the dots at regular intervals, but from a different angle; as the two grids meet, this misalignment leaves visual artifacts on the digitized image.

In Figure 4, the grid misalignment has created a psychedelic blue and orange sky, which appears to fascinate the astronomer Hipparchus (Giberne, 1908). These moiré patterns appear throughout the image, along with color aliasing, from wavy striations in the sky and floor to geometric patterns on building columns. Color aliasing occurs when the spatial frequency of the original image is sampled at a rate inadequate to capture all its details. Like moiré, it is a common phenomenon among Google-digitized books that contain engravings or etchings.

While the problem of digitization and moiré has been discussed since at least the 1970s, and corrective measures have been identified (Huang, 1974), no fully automated solution appears to have emerged. In 1996, the Library of Congress acknowledged that moiré mitigation strategies remained unsuitable for production-scale environments (Fleischhauer, 1996). This type of error is predictable, yet intractable, in large-scale book digitization. It is ironic that halftone screening—a technique that facilitated the mass reproduction of photographs for print books and newspapers—became a significant challenge to mass print digitization.


**Figure 4.** Moiré and color aliasing (Giberne, 1908).

Google's automated image processing also often misrecognized features of print books. Initially captured in full color, raw bitmapped images were then processed down to bitonal images for textual content or 8-bit grayscale for illustrated content (University of Michigan Library, 2005). Figure 5 shows a page of text rendered as a grayscale illustration. The thinness of the original rice-paper volume allowed content from adjoining pages to bleed through during scanning. This, combined with the nuanced shading of Chinese characters, caused the system to miscategorize the page (Zhang and Kangxi Emperor of China, 1882).

On the other hand, the same Chinese text often fared poorly when rendered as a bitonal image within the GBS digitization model. Binarization converts a raw color digital image into a bitonal image by using an

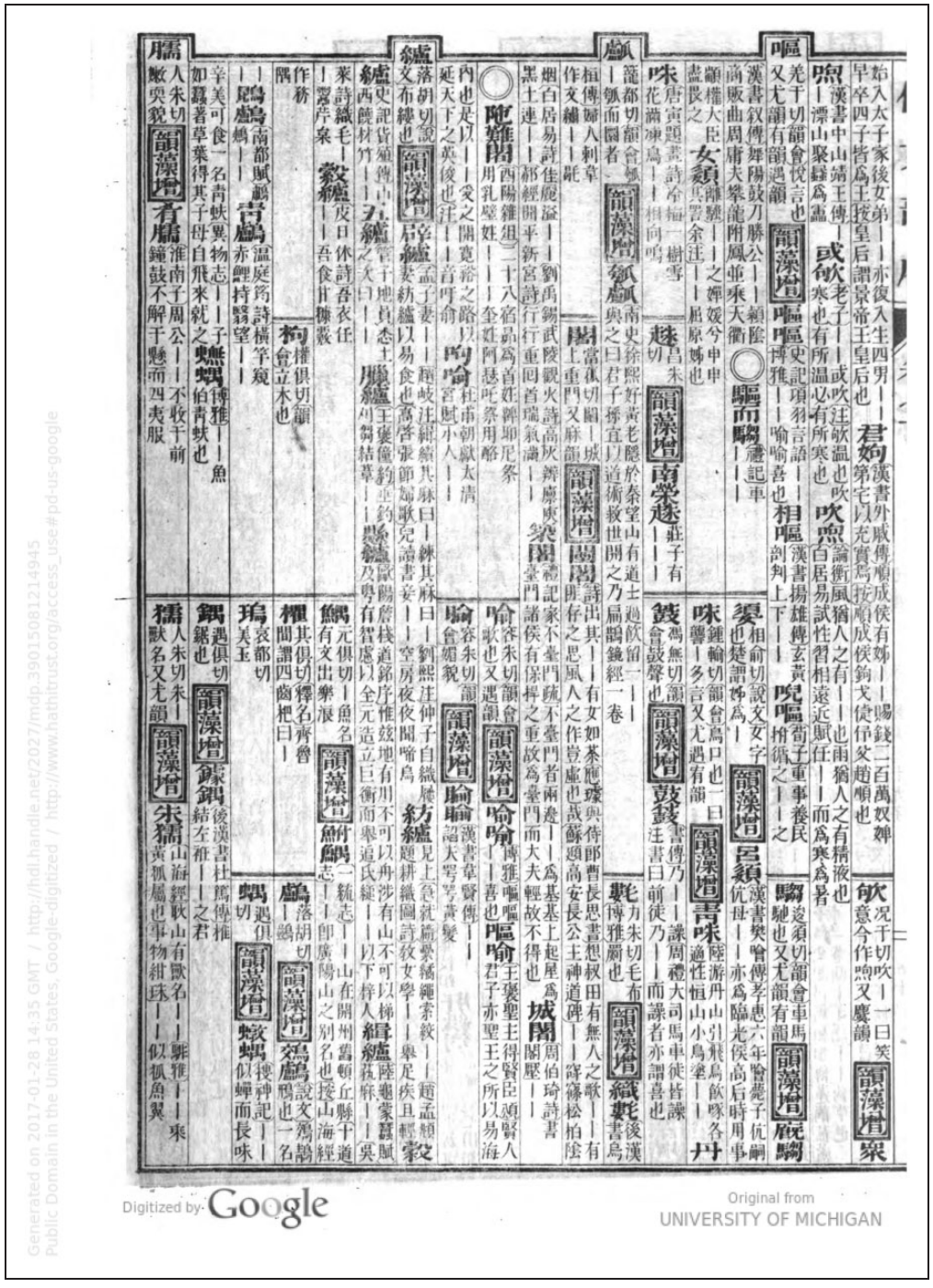


**Figure 5.** Grayscale rendering, Chinese text on rice paper (Zhang and Kangxi Emperor of China, 1882).

automatically determined threshold to differentiate foreground and background. This technique reduces the amount of data contained in full-color scans, thereby speeding up OCR processing and downstream image distribution (Holley, 2009; Vincent, 2007). However, Google's threshold settings often have the effect of darkening, lightening, or erasing nuance from rendered characters. Figure 6, from the same book as the preceding example, illustrates the consequences of automated binarization for calligraphy pen detail.

This problem is avoided by interleaving blank pages to block adjoining page noise, but to do so routinely would slow the scanning process considerably. Further,

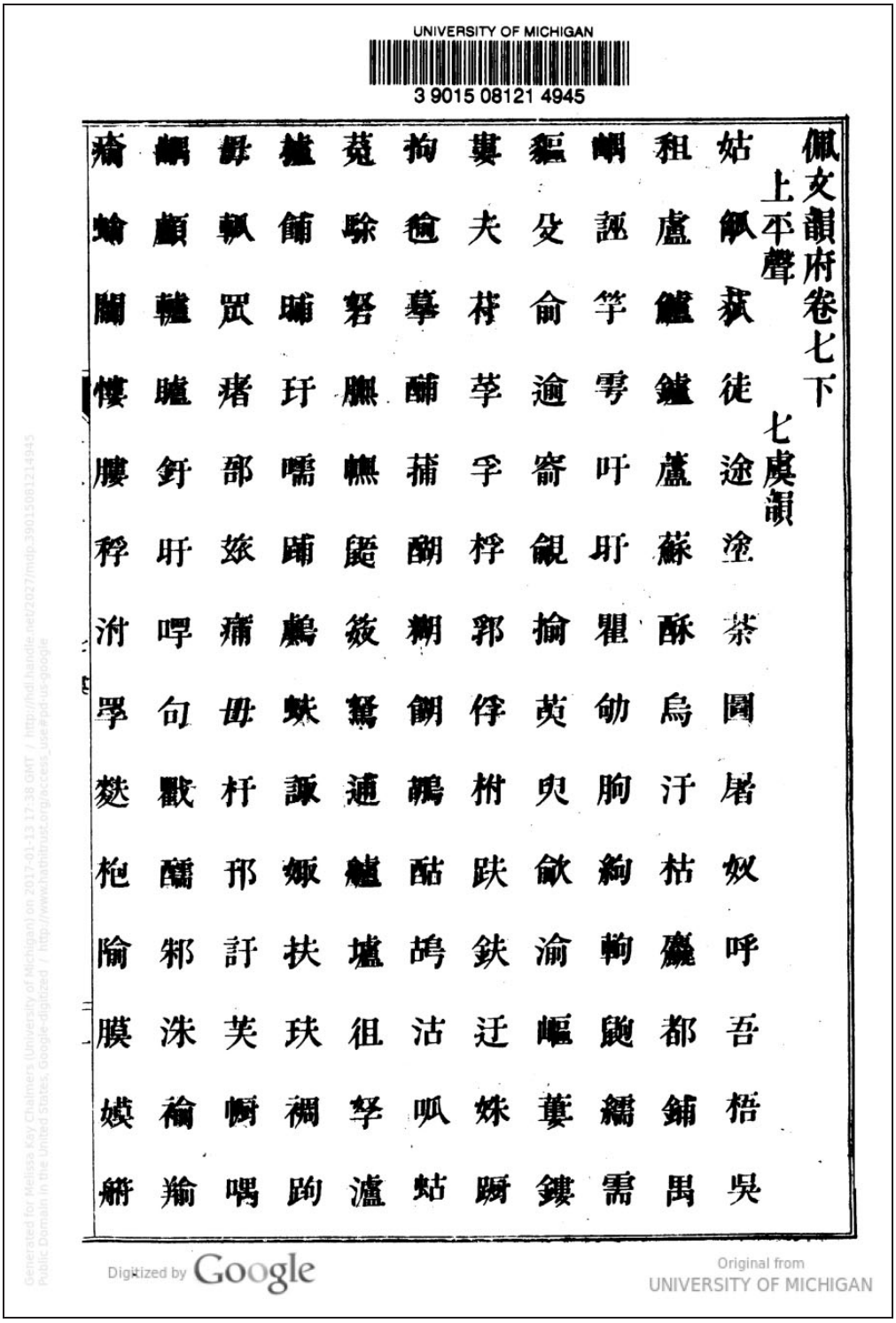


**Figure 6.** Bitonal rendering of Chinese text on rice paper (Zhang and Kangxi Emperor of China, 1882).

without specialized language skills, the original book in hand, or time for careful examination, it can be very difficult to recognize the nature or extent of information loss in a digitized page image. In a related example, Google's standard protocol—scanning books front to back and left to right—often caused books with vertical or right-to-left writing formats to be delivered backward or upside down (Weiss and James, 2015).

### Textual content

Optimizing workflows for OCR does not in itself assure high quality character recognition. Consistent with Google's brute-force approach, corpus indexing (and keyword search) was built upon software-generated uncorrected OCR. Research evaluating OCR in large-scale text digitization reveals widespread accuracy and reliability problems; as with imaging, OCR accuracy is challenged by print material features such as age and condition, printing flaws, rare fonts, textual annotations, and nontext symbols (Holley, 2009; Tanner et al., 2009). OCR also suffers in the presence of imaging quality issues such as page skew, low resolution, bleed through, and insufficient contrast.

Recall the page images of *Mother Goose's Melody* in Figure 3. Surrounded by visual artifacts of the digitization process, the text—a maxim about the value (and challenge) of independence—appears generally readable. However, the OCR provided for the page (Figure 7) reveals numerous problems, from missing words to problems caused by the long s's in the original text.

Human OCR correction, traditionally completed by professionals double-keying texts, is considered the accuracy gold standard but is cost-prohibitive at scale (Tanner et al., 2009). In 2009, Google acquired reCAPTCHA, owner of the web security technology CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) (Von Ahn and Cathcart, 2009). This technology, in widespread use since 2006, asks users to examine digitized images of words OCR cannot interpret. Harnessing the free labor of web users a few seconds at a time, but aggregating to millions of hours, reCAPTCHA has improved the usability of the GBS corpus (for certain languages) while also being fed back into the training sets of machine-learning algorithms. GBS thus fills gaps in its automated quality control system with "human computation," defined by CAPTCHA creator Von Ahn (2005) as treating "human brains as processors in a distributed system" to solve problems that cannot (yet) be undertaken by computers alone.

### Metadata

Scholarly users of Google Books quickly identified problems with its metadata, e.g. item descriptors such as author, publication date, and subject classification contained in traditional library catalogs (Duguid, 2007; Nunberg, 2009; Townsend, 2007). Using his knowledge of canonical texts as a point of departure, Nunberg (2009) conducted searches in the GBS corpus that revealed extensive errors in volume-level metadata. These included a disproportionate number of books listing 1899 as their publication date; anachronistic dates for terms such as "internet"; mixups of author, editor, and/or translator; subject misclassification (e.g., using publishing industry classifications designed to allocate books to shelf space in stores, rather than Library of Congress subject headings); and mis-linking (e.g., mismatch between volume information and page images). James and Weiss's (2012) quantitative assessment supports Nunberg's anecdotal findings. In response, Google acknowledged that it had constructed book metadata records by parsing more than 100 sources of data (Orwant, 2009). These included library catalogs, publishing industry data, third-party metadata providers, and likely data extracted from OCR. If each source contained errors, Google's Jon Orwant acknowledged, the GBS corpus aggregated millions of metadata errors across trillions of individual data fields. (That the most explicit official statement of Google's approach to metadata takes the form of a 3000+ word blog post comment is at once extraordinary and unsurprising.)

Google's metadata mess was quickly—and publicly—cast as a confrontation between old and new information systems for accessing books, evidence of Google's techno-utopian investment in machine intelligence and the power of full-text search to triumph over the centralized library cataloging systems

**Page 13**

Mother GOOSE's Melody. Pray what mull we have for to eat, eat, eat? Will the Flame that you're fo rich in. Light a Fire in the Kitchen, Or the little God of Love turn the Spit, Spit, Spit? V. Then the little man he figh'd, And, fome fay, a little cry'd, For his little Heart was big with Sorrow, Sorrow, Sorrow; As I am your little Slave, If the little that I have Be too little, little, we will borrow, borrow, borrow.* * He who borrows is another Slave, and pawns his Honour, his Man's Liber- ty, and fometimes his Nofe for the pay- ment. Learn to Uve on a little, and be independent. Patch on Prudence. VI. Then

**Figure 7.** OCR produced from page images of *Mother Goose's Melody* p. 13 (Figure 4) (Thomas and Shakespeare, 1945).

constructed painstakingly by librarians (Nunberg, 2009). At a minimum, the pervasiveness of metadata errors drew attention to the irony of Google's public construction of GBS as an "enhanced card catalog." In practice, the need to circumvent license restrictions on bibliographic data significantly shaped Google's approach to metadata. Coyle (2009) and Jones (2014) assert that although Google obtained catalog records from library partners, libraries' contracts with OCLC—a company that produces the union catalog WorldCat—probably prohibited Google from displaying that metadata directly. (For efficiency and consistency, libraries often download catalog records from WorldCat rather than create their own, but OCLC restricts their use.)

Google's metadata problems exposed imperfections in existing book cataloging systems, from the challenges of algorithmically interpreting MARC records to the temporal and geographic limitations of ISBNs to errors in human-catalogued bibliographic data. The incompatibility of legacy catalog systems further challenged Google's attempts to aggregate metadata from multiple sources. Over time, incremental modifications to Google's machine processing substantially improved, identifying and ameliorating systemic metadata problems. Nonetheless, GBS metadata continues to be far from accurate.

### *Integrating books into the web*

Unlike print books, the web is not tied to a single physical device for content delivery. In 2009, Google introduced "mobile editions" of the corpus. The development team explained:

> Imperfect OCR is only the first challenge in the ultimate goal of moving from collections of page images to extracted-text-based books... The technical challenges are daunting, but we'll continue to make enhancements to our OCR and book structure extraction technologies. With this launch, we believe that we've taken an important step toward more universal access to books. (Ratnakar et al., 2009)

By defining books as structured information carriers from which content may be extracted and delivered seamlessly via widely varying devices, Google's focus on mobile technology further distanced digitized books from their print origins.

Approaching books as one among many objects to integrate into web search, Google also projected web-based *expectations of change* onto print books. Search engines crawl the web constantly, capturing changes, additions, and deletions to a massive set of networked pages. A well-justified expectation of constant flux drives this crawling, a scale of change only manageable through constant wholesale capture. By contrast, the pace of change for print media on library shelves is normally much slower. Pages may turn brittle. Users may mark up books, or more rarely, steal them. While Google tried to deploy a "scan once" strategy for initial imaging, when it comes to image *processing* it has treated its book corpus with a disregard for stability borne out of its experience with web pages. Embracing the iterative logic of algorithmic systems, Google routinely updates and replaces scanned content after running it through improved error detection and image quality algorithms (University of Michigan and Google, Inc., 2005). Even if changes to the corpus tend to be small and incremental—algorithms erase a finger in the margins of a scan, restore a missing page, or deliver a once-buried quote in search results—the constant and accumulating changes generate a sense of instability. Google has not consistently provided users with documentation related to this updating (Conway, 2015); the automated work of maintenance and repair remains invisible. It is a tangled, even paradoxical relationship, as the fundamental revisability of algorithms supersedes the print book's material stability and persistence. But while algorithmic logic suggests that the latest version of a page will always be the most accurate, critical traditions rooted in print culture may lead us to ask how GBS defines accuracy and what other characteristics may be altered by real-time updating.

This section has demonstrated that because GBS page images and machine-searchable text are in effect coproduced, an action at one stage of the process can set in motion a cascade of consequences that shape both visual and machine readability in the corpus. At scale, optimizing workflows around textual properties of books ran the risk not only of distorting some books' visual properties but also of defining normative book characteristics. In Google's one-size-fits-most scanning system, decisions about image processing may have a disproportionate effect on certain aspects of the digitized corpus; the Chinese-language volume described above was one of a set of 50, all digitized by Google at a single location and all subject to the same processing problems.

Objects that are excluded from scanning, or distorted and transformed beyond the point at which they may be used as surrogates for their print originals, become "noncharismatic objects" (Bowker, 2000): by failing to be "collected" through digitization, they are rendered invisible to future digitally based scholarship or use. Further, Google's opportunistic rather than systematic approach to digitization may amplify existing selection biases in physical print collections, overrepresent certain types of publications (Pechenick et al.,

2015), or perpetuate Anglo-American cultural dominance in digital cultural heritage (Jeanneney, 2008).

## Mediating access: Indexing the world, one piece of text at a time

By constructing books as data, GBS inserts them into a networked world where algorithms increasingly mediate human access to information. In the project's wake, the dream of digitizing "everything" has taken hold, recalibrating the sense of what is possible and what is expected for both individual web users and cultural heritage institutions.

This article is the first piece of a larger, ongoing study of several large-scale cultural heritage digitization projects, including the Internet Archive and genealogy organization FamilySearch. This project seeks to join an existing critique oriented toward material culture and labor process with an emerging critique of algorithmic culture. "Algorithmic digitization" thus serves us as a sensitizing concept emphasizing relationships between inputs, materials, labor, processes, outputs, use, and users. We use it here to consider opportunities and limitations in Google's approach to providing universal access to information.

Understanding GBS as an algorithmic system renders visible multiple tensions in the project: between Google's universalizing public rhetoric about the project and the technical processes that must translate these ambiguous visions into workflows; between the competing goals of stakeholders such as Google, publishers, authors, and libraries; between aspirations of scale and the specialized needs of individual end users or books; between the materiality of the print book and that of the computer; and between the invisible, iterative authority of algorithms and that of human visual experience or expertise.

As we have seen, notable limitations stem from Google's choices in resolving these tensions. Imperfection is unavoidable in large-scale book digitization. Yet the vocabulary of error is often too static to be useful, since error is always relative to a particular user and/or purpose. Gooding (2013) argues that large-scale cultural heritage digitization sacrifices quality to serve scale. We have shown that while intuitively appealing, this argument is too simplistic. It tends to align "quality" with the needs and values of traditional readers, thus privileging visual access. In doing so it ignores the extent to which quantity and quality are mutually constitutive in building a digitization economy of scale and misses the careful calibration of trade-offs between multiple forms of access to books afforded by digitization. It misunderstands the measures by which the project itself has defined and evaluated quality. Finally, it overstates Google's concern with end users more generally.

We must, then, attend carefully to how Google's algorithmic system supports some users' requirements while simultaneously rendering others difficult or impossible to meet. For example, "visible page texture"—from marginalia to other signs of aging or use inscribed on the printed page—may be useful information or a mark of authenticity for some users, yet it is defined as noise for automated image processing. A situated understanding of these details exposes limitations to GBS's suitability as a flexible, general-use collection that can meet the needs of a range of stakeholders, such as readers (Duguid, 2007; Nunberg, 2009), researchers conducting quantitative analyses of cultural trends (Michel et al., 2011), or cultural heritage institutions.

Further, the opacity of Google's processes has contributed to widespread critique of libraries and other memory institutions "outsourcing the risk and responsibility" for digitization to a private company (Vaidhyanathan, 2012). Google's "black box outsourcing model" (Leetaru, 2008) frames agreements with content providers as partnerships rather than customer–client relationships. These partners give up some control over project parameters, tacitly agree to participate in the digitizer's larger projects or agendas, and remain dependent on the digitizer's continued interest and investment in digitization. As smaller institutions and collections gain access to digitization through this privatized model, the risks grow. Google's digitization model conceals the resource-intensive nature of digitization, from the invisible labor of professional librarians, contract workers, and end users filling in the gaps created by incomplete automation to unanswered questions of long-term maintenance or preservation of digital assets. It may thus discourage cultural heritage institutions from budgeting sufficiently for their own digitization infrastructures. This will doubtless leave some institutions unprepared to maintain their traditional stewardship roles with respect to digital content.

Just as users (individuals or institutions) benefit or suffer from Google's reliance on algorithmic processing differently, so too are print books unevenly affected. Google's highly proceduralized scanning workflows (perhaps inadvertently) imposed a normative idea of the form and content of the English language book on the digitization process. With its construction of digitization as a text extraction and indexing challenge, Google further distanced itself from library-based understanding of the value of scanned page images as surrogates for print originals. Instead, the above analysis has revealed several ways in which Google aligned GBS with other iterative, algorithmic systems—from Google Streetview to 23 & Me—created to bring physical objects, information systems, and even human

bodies within the visual and computational logics of the web.

Today, books maintain an uneasy parallel existence, caught between the world of the web and the world of Gutenberg. GBS highlights the uneven rates of change and competing logics of these two worlds, the technological and legal frameworks that may produce, organize, and mediate access to print and digital information differently but that digitization forces together. Google shaped the processes and outputs of GBS to respect the constraints of copyright law, for example. Yet it simultaneously sought to circumvent print-based permissions management by emphasizing functionality that resonated with its web- and scale-centric mission but had no direct parallel with print.

GBS has provided searchable text access to millions of books. The weight of this remarkable achievement must not be denied or underestimated. Yet by equating digital access with full-text search, the GBS corpus has created a future for books in which they are defined principally by their textual content. Google's workflows have elided other (historical, artifactual, material) properties of books that, when absent, threaten to disrupt or reframe the relationship between a digitized surrogate and its print original. As print libraries fade into the deep background of our brave new digital world, much has been lost that cannot be regained.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. Google Book Search: http://books.google.com.
2. The original five libraries were Harvard, Stanford, the University of Michigan, New York Public Library, and the Bodleian Library at Oxford University.
3. While not the first, GBS was the biggest and most controversial of several large cultural heritage digitization projects undertaken by entities such as Yahoo, Microsoft, Google, and the Internet Archive in the early 2000s (St. Clair, 2008).
4. The Association of American Publishers lawsuit was settled privately in 2011, while in 2015 the Second Circuit Court of Appeals upheld a 2013 lower court judgment rejecting the Authors Guild's copyright infringement claims and affirming Google's scanning as transformative and therefore "fair use."
5. While patents provide only generic system descriptions, they provide sufficient detail for high-level reverse engineering of Google's processes. Journalists' accounts and output-oriented research provide anecdotal verification (Clements, 2009; Shankland, 2009).

## References

Baird HS (2003) Digital libraries and document image analysis. In: *Seventh international conference on document analysis and recognition*, Los Alamitos, CA, 4–6 August 2013, pp.2–14. IEEE.

Band J (2009) The long and winding road to the Google Books settlement. *The John Marshall Review of Intellectual Property Law* 9(2): 227–329.

Bowker GC (2000) Biodiversity datadiversity. *Social Studies of Science* 30(5): 643–683.

Cadwalladr C (2016) How to bump Holocaust deniers off Google's top spot? Pay Google. *The Guardian*, 17 December. Available at: https://www.theguardian.com/technology/2016/dec/17/holocaust-deniers-google-search-top-spot (accessed 1 February 2017).

Carr R (2005) *Oxford-Google Mass-Digitisation Programme*. Washington, DC. Available at: http://www.bodley.ox.ac.uk/librarian/rpc/CNIGoogle/CNIGoogle.htm (accessed 1 February 2017).

Ceynowa K (2009) Mass digitization for research and study. *IFLA Journal* 35(1): 17–24.

Clements M (2009) The secret of Google's book scanning machine revealed. *National Public Radio website*. Available at: http://www.npr.org/sections/library/2009/04/the_granting_of_patent_7508978.html (accessed 7 February 2017).

Conway P (2013) Preserving imperfection: Assessing the incidence of digital imaging error in HathiTrust. *Preservation, Digital Technology and Culture* 42(1): 17–30.

Conway P (2015) Digital transformations and the archival nature of surrogates. *Archival Science* 15(1): 51–69.

Coyle K (2006) Mass digitization of books. *The Journal of Academic Librarianship* 32(6): 641–645.

Coyle K (2009) Google Books metadata and library functions. *Coyle's InFormation*. Available at: http://kcoyle.blogspot.com/2009/09/google-books-metadata-and-library.html (accessed 19 April 2017).

Darnton R (2009) *The Case for Books: Past, Present, and Future*. New York: Public Affairs.

Duguid P (2007) Inheritance and loss? A brief survey of Google Books. *First Monday* 12(8).

Fleischhauer C (1996) Digital Formats for Content Reproductions. Library of Congress. Available at: http://memory.loc.gov/ammem/formatold.html (accessed 16 June 2017).

Giberne A (1908) *The Story of the Sun, Moon, and Stars*. Chicago, IL: Thompson & Thomas. Available at: https://books.google.com/books?id=KY8AAAAAMAAJ (accessed 1 February 2017).

Gillespie T (2016) Algorithms. In: Peters B (ed.) *Digital Keywords*. Princeton, NJ: Princeton University Press, pp. 18–30.

Golumbia D (2009) *The Cultural Logic of Computation*. Cambridge, MA: Harvard University Press.

Gooding P (2013) Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing* 28(3): 425–431.

Google, Inc. (1999) Company info. Available at: https://web.archive.org/web/19991105194818/http://www.google.com/company.html (accessed 1 February 2017).

Google, Inc. (2004a) What is Google Print? *About Google Print (Beta)*. Available at: https://web.archive.org/web/20041214092414/http://print.google.com/ (accessed 10 February 2017).

Google, Inc. (2004b) What is the library project? *Google Print Library Project*. Available at: https://web.archive.org/web/*/http://print.google.com/googleprint/library.html (accessed 1 December 2017).

Grant J (2005) Judging book search by its cover. *Official Google Blog*. Available at: https://googleblog.blogspot.com/2005/11/judging-book-search-by-its-cover.html (accessed 1 February 2017).

Holihan C (2006) Google seeks help with recognition. *Business Week Online*. Available at: http://www.bloomberg.com/bw/stories/2006-09-06/google-seeks-help-with-recognition (accessed 1 February 2017).

Holley R (2009) How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine* 15(3/4).

Howard J (2012) Google begins to scale back its scanning of books from university libraries. *The Chronicle of Higher Education*.

Huang TS (1974) Digital transmission of halftone pictures. *Computer Graphics and Image Processing* 3(3): 195–202.

James R (2010) An assessment of the legibility of Google Books. *Journal of Access Services* 7(4): 223–228.

James R and Weiss A (2012) An assessment of Google Books' metadata. *Journal of Library Metadata* 12(1): 15–22.

Jeanneney J-N (2008) *Google and the Myth of Universal Knowledge*. Chicago: University of Chicago Press.

Jones EA (2014) *Constructing the universal library*. PhD Thesis, University of Washington, USA.

Kirschenbaum MG (2003) The word as image in an age of digital reproduction. In: Hocks ME and Kendrick M (eds) *Eloquent Images*. Cambridge: MIT Press, pp. 137–156.

Langley A and Bloomberg DS (2007) Google Books: Making the public domain universally accessible. In: *Proceedings of SPIE-IS&T Electronic Imaging*, 26–29 January 2007, San Jose, CA: International Society for Optics and Photonics.

Le Bourgeois F, Trinh E, Allier B, et al. (2004) Document image analysis solutions for digital libraries. In: *First international workshop on document image analysis for libraries*, Palo Alto, CA, 23–24 January 2004, pp.2–24. IEEE.

Leetaru K (2008) Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday* 13(10).

Lefevere F-M and Saric M (2008) De-warping of scanned images. Patent 7463772, USA.

Lefevere F-M and Saric M (2009) Detection of grooves in scanned images. Patent 7508978, USA.

Lesk M (2003) The price of digitization: New cost models for cultural and educational institutions. Available at: http://www.ninch.org/forum/price.lesk.report.html (accessed 1 February 2017).

Lin XF (2006) Quality assurance in high volume document digitization: a survey. In: *Second international conference on document image analysis for libraries*, Lyon, France, 27–28 April 2006, pp.311–319. IEEE.

Lynch C (2002) Digital collections, digital libraries & the digitization of cultural heritage information. *Microform and Imaging Review* 31(4): 131–145.

Madrigal AC (2010) Inside the Google Books Algorithm. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2010/11/inside-the-google-books-algorithm/65422/ (accessed 3 February 2017).

Michel J-B, Shen YK, Aiden AP, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182.

Milne R (2008) From "boutique" to mass digitization: The Google Library Project at Oxford. In: Earnshaw R and Vince J (eds) *Digital Convergence – Libraries of the Future*. London: Springer, pp. 3–9.

Murrell M (2010) Digital+library: Mass book digitization as collective inquiry. *New York Law School Law Review* 55: 221–249.

New York Public Library (2004) NYPL partners with Google to make books available online. Available at: https://web.archive.org/web/20050923130755/http://nypl.org/press/google.cfm (accessed 3 December 2016).

New York S of (1862) *Code of Procedure of the State of New York*. New York: George S. Diossy. Available at: https://books.google.com/books?printsec=frontcover&id=aD0KAAAAIAAJ (accessed 1 February 2017).

Norman Wilson A (2009) Workers leaving the Googleplex. Available at: http://www.andrewnormanwilson.com/WorkersGoogleplex.html (accessed 7 August 2016).

Nunberg G (2009) Google Books: A metadata train wreck. *Language Log*. Available at: http://languagelog.ldc.upenn.edu/nll/?p=1701 (accessed 4 February 2017).

Orwant J (2009) Re: Google Books: A metadata train wreck. *Language Log*. Available at: http://languagelog.ldc.upenn.edu/nll/?p=1701#comment-41758 (accessed 24 April 2017).

Palmer B (2005) Deals with Google to accelerate library digitization projects for Stanford, others. *Stanford Report*, 12 January. Available at: http://news.stanford.edu/news/2005/january12/google-0112.html (accessed 3 February 2017).

Pasquale F (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Pechenick EA, Danforth CM and Dodds PS (2015) Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE* 10(10): e0137041.

Ratnakar V, Poncin G, Bedger B, et al. (2009) 1.5 million books in your pocket. *Google Book Search blog*. Available at: http://booksearch.blogspot.com/2009/02/15-million-books-in-your-pocket.html (accessed 29 January 2017).

Roush W (2005) The infinite library. *MIT Technology Review* 108(5): 54–59.

Samuelson P (2009) Google Book Search and the future of books in cyberspace. *Minnesota Law Review* 94: 1308–1374.

Schantz HF (1982) *The History of OCR*. Manchester Center, VT: Recognition Technologies Users Association.

Schmidt E (2005) Books of revelation. *Wall Street Journal* 18 October.

Seaver N (2013) Knowing algorithms. Cambridge, MA. Available at: http://nickseaver.net/papers/seaverMiT8.pdf (accessed 1 February 2017).

Shankland S (2009) Patent reveals Google's book-scanning advantage. *CNET*. Available at: https://www.cnet.com/news/patent-reveals-googles-book-scanning-advantage/ (accessed 30 January 2017).

St. Clair G (2008) The Million Book project in relation to Google. *Journal of Library Administration* 47(1–2): 151–163.

Striphas T (2015) Algorithmic culture. *European Journal of Cultural Studies* 18(4–5): 395–412.

Tanner S, Muñoz T and Ros PH (2009) Measuring mass text digitization quality and usefulness. *D-Lib Magazine* 15(7/8): 1082–9873.

Terras MM (2008) *Digital Images for the Information Professional*. Burlington, VT: Ashgate Publishing.

Thomas I and Shakespeare W (1945) Mother Goose's melody. New York: G. Melcher. Available at: https://books.google.com/books?id=OG7YAAAAMAAJ (accessed 7 February 2017).

Townsend RB (2007) Google Books: What's not to like? *American Historical Association blog*. Available at: http://blog.historians.org/2007/04/google-books-whats-not-to-like/ (accessed 7 February 2017).

University of Michigan and Google, Inc. (2005) UM-Google Cooperative Agreement. Available at: www.lib.umich.edu/mdp/um-google-cooperative-agreement.pdf (accessed 7 February 2017).

University of Michigan Library (2005) UM Library/Google digitization partnership FAQ. Available at: http://www.lib.umich.edu/files/services/mdp/faq.pdf (accessed 7 February 2017).

US Copyright Office (2016) Fair use. Available at: http://copyright.gov/fair-use/more-info.html (accessed 7 February 2017).

Vaidhyanathan S (2012) *The Googlization of Everything*. Berkeley, CA: University of California Press.

Vincent L (2007) Google Book Search: Document understanding on a massive scale. In: *Ninth international conference on document analysis and recognition*, Parena, Brazil, 23–26 September 2007, pp.819–823. IEEE.

Von Ahn L (2005) *Human computation*. PhD Thesis, Carnegie Mellon University, Pittsburgh, PA.

von Ahn L and Cathcart W (2009) Teaching computers to read: Google acquires reCAPTCHA. *Official Google Blog*. Available at: https://googleblog.blogspot.com/2009/09/teaching-computers-to-read-google.html (accessed 30 January 2017).

Weiss A and James R (2015) Comparing the access to and legibility of Japanese language texts in massive digital libraries. In: *International conference on culture and computing*, Kyoto, Japan, 17–19 October 2015, pp.57–63. IEEE.

Zhang Y and Kangxi Emperor of China (1882) Pei wen yun fu. Available at: http://hdl.handle.net/2027/mdp.39015081214945 (accessed 7 February 2017).