Modeling the Book Trade

3

Within Amazon's warehouses, there is no distinction between a cloth copy of Jami Curl's *Candy Is Magic*, a variety box of Skittles and Starbursts, or a Peppa Pig plush doll. All three items are stored, organized, and shipped in the same fashion.[1] Despite this ambivalence, books remain central to Amazon's retail operations. Books, and specifically ebooks, accumulate and return a greater volume of data than other products. Amazon uses both distribution and consumption to monitor the ebb and flow of the book trade to predict future trends, working with well-established industry standards including ISBN, Online Information Exchange (ONIX), and machine-readable cataloging (MARC) to create a conceptual model of the book and the publishing industry. The Kindle built on the core pillars of Amazon Web Services' virtualization tools detailed in the previous chapter and the company's commitment to model the book based on the wealth of internal and external data it had accumulated.

Warehouses of Metadata

Bezos faced an uphill battle to compete with Barnes & Noble but used the publishing industry's long-standing collaboration to develop ISBNs to Amazon's advantage. F. Gordon Foster, a professor at the London School of Economics, developed the British Standard Book Number (SBN) in 1966 to manage the needs of WH Smith at the request of the British Publishing Association. The standard was quickly adopted by a range of booksellers and publishers for cataloging and warehousing, with the International

Organization for Standardization (ISO) ratifying the standard as International Standard Book Numbers (ISBNs) in 1970.[2] Publishers now use ISBNs as the backbone of spreading metadata to distributors, retailers, and libraries.[3] The metadata standard provides information about a book's publisher and region in a human- and machine-readable format.[4] This data structure also distinguishes between different editions of the same book. For example, the 1991 Picador paperback copy of Alastair Gray's *Lanark* has an ISBN of 0-330-31965-5. The first number indicates the publication region, with 0 representing the Anglophone market. The number 330 refers to Picador as the publisher, with the following five digits noting the specific book in Picador's list. Each ISBN concludes with a check digit to ensure the number was correctly entered through "an elegant and rather ingenious system since it guards not only against inaccurately recorded digits but also against the apparently more common error of transpositions."[5] ISBNs can identify specific editions of books with complex bibliographic histories. In the case of Gray's *Lanark*, my search on Amazon.co.uk in July 2015 returned two Canongate Classics editions published in 2002 and 2007, and the original hardback from 1982 on sale for £87.47. Conversely, Amazon.com records the two Canongate titles, the original, an American edition published by Harvest Books, and the Picador edition. Amazon's search function privileges local and new editions, but as it integrates ISBNs into product page URLs, it is possible to bypass this navigation system to locate specific editions. Book collectors can use the system to ensure they receive the expected version of a book rather than purchasing the latest reprint.

The World Wide Web of 1994 at Amazon's launch was vastly different from the contemporary web, where all information is assumed to be online. Data about the book trade appeared in sources such as Bowker's *Books in Print*, which was available to retailers in print or on CD-ROM and contained a catalog of ISBNs for books still available to purchase from publishers and wholesalers. Retailers used *Books in Print* to supplement their in-store holdings and as a data source for preordering books. Customers could request titles from the database through the retailer, but the information was not publicly available without purchasing access. Information about forthcoming books was only available through trade sources, the popular press, or bookshops. Amazon challenged this informational monopoly by building its online catalog from the *Books in Print* data set to demonstrate the large catalog of titles they *could* order from wholesalers rather than what was currently in stock.

Amazon used *Books in Print* to inflate the number of titles available on its website to make its selection appear larger than that of other retailers.

The boast that Amazon was the "world's largest bookshop" drew a lawsuit from Barnes & Noble in 1997, which was settled later that year when the companies agreed "they would rather compete in the marketplace than in court."[6] Amazon's sleight of hand opened an industry data source for public consumption to allow users with niche interests to discover relevant books. Brick-and-mortar bookshops need to stock books they are confident will sell within a reasonable time frame to justify the shelf space at the expense of discovering more obscure titles. Ordering directly from distributors and warehouses at the point of customer order while displaying a large stock rewrote these rules. Readers who did not feel represented by the limitations of brick-and-mortar retailers saw the extensive database as catering to their tastes.

Amazon reshaped ISBNs as web-based metadata that were discoverable rather than an identifying number for warehousing and distribution, turning an industry standard into a broader public good. Product pages turned ISBNs into searchable entities on the web, acting as unofficial URIs for the book trade. Book metadata were no longer obfuscated through standards such as ONIX or MARC. Others have built on this process, including Daniel Green, the cofounder of CamelCamelCamel, a website that tracks price fluctuations on Amazon, which launched in 2008.[7] This is vital in the world of algorithmically mediated pricing, since automated systems can rapidly increase the cost of a book beyond its actual value, as with Michael Eisen's example of a $23,698,655.93 copy of Peter Lawrence's *The Making of a Fly*.[8] External services including Google Books link to Amazon as a centralized repository of book metadata that would only exist in a more authoritative form on individual publishers' websites. The company's ubiquity across book trade platforms is equivalent to Facebook's integration of "likes" across the web, establishing the company as bookish infrastructure.

In April 1998, David Risher, then lead of product development for Amazon, launched an investigation into selling music, with the new product shop opening later that year.[9] The music industry had no equivalent for ISBNs, so Amazon created a broader cataloging standard. The company decided to issue every item a ten-digit Amazon Standard Identification Number (ASIN), which uses the full range of alphanumeric characters to identify nonbook items.[10] Currently every nonbook ASIN (including ebooks) begins with *B*. This practice can catalog up to three thousand billion unique items. The development of ASIN runs counter to the widespread use of Universal Product Codes (UPC), a barcode standard featuring thirteen or eighteen numbers to identify items from various countries.[11] UPCs integrated a superset of ISBN called ISBN-13, formally introduced

in January 2007, although the format had a longer history starting with the inclusion of barcodes on books in the 1980s.[12] ISBN-13 contains no additional metadata, as the first three digits, usually 978, refer to the country, which in the case of books is represented by the fictitious "Bookland."[13] Preexisting ISBNs could be used with the new prefix and a recalculated check digit. Bookland was also given the prefix 979, doubling the possible number of ISBNs. ISBNs prefixed with 979 are not used in Anglophone markets as of 2018 but have already been adapted by the French ISBN authority, L'AFNIL (Agence Francophone pour la Numérotation Internationale du Livre).[14] French books with ISBN-13s starting with 979 receive an ASIN to identify the product page, and the ISBN is only visible in the "Détails sur le produit."[15]

While ISBNs and ASINs can identify unique editions, Amazon uses a higher-level metadata structure to connect separate editions to reuse valuable data internally. Christopher Weight, a member of Kindle Special Ops, led research on "title sets," which connect similar books based on a probabilistic model of both metadata and textual similarities.[16] Strategic reuse of data also links relevant metadata across regions regardless of the local publisher. For example, reviews are shared across editions to boost the number of voices commenting on a product, although this creates problems when criticism of a poorly edited version is carried over to a more carefully produced edition. The English-language edition of Emily St. John Mandel's *Station Eleven* is available to purchase from nine Amazon regions, with shared data despite the novel's publisher varying between the United States (Vintage / Penguin Random House), Canada (Harper Perennial / HarperCollins), and United Kingdom (Picador/Macmillan). Most often, blurbs and professional reviews will be linked, but if a lower threshold for reviews has not been met, reviews from a more popular edition will be pulled through. It is common practice in book retail to use data sources such as Nielsen BookData or Bowker's *Books in Print* to populate online catalogs. Data will appear consistently across Waterstones, AbeBooks, and Bokus, but Amazon expanded this approach by extracting subsets of its data for different markets and formats.

Although ISBN was an early international identification standard, the book trade struggled to agree on other metadata standards. Bad practice is still common within the trade, encouraging Karina Luke, executive director of Book Industry Communications (the British publishing supply chain organization), to publish a statement condemning superfluous title metadata such as the introduction of "Man Booker Prize winner" or "The explosive next book from . . ." in the subtitle of product pages for Amazon and other online retailers.[17] The disparities between identifying books

and authors have remained an open problem, as author metadata can be inconsistent and frequently requires disambiguation for common names such as John Smith or generates multiple records attached to transliterated names such as Fyodor Dostoyevsky. Projects such as Open Researcher and Contributor ID (ORCID) attempt to create consistent standards for authors in academia but rely on contributors signing up, limiting its appeal as a universal standard.[18] Amazon extended the ASIN standard to include persistent identifiers for both authors and customers, flattening the hierarchical relationship between the objects. For example, George R. R. Martin has an author ASIN of B000APIGH4, whereas the Kindle edition of his novel *A Storm of Swords* is the product number B004PIJEWD. Amazon integrated authors, bibliographic objects, and consumers into a single conceptual model that favors relationships over hierarchies of production. Amazon prioritized connectivity over authority when constructing its computational model of the book trade.[19] When everything is treated as a relational entity, it can be difficult to find meaning. For example, Paul Ford struggled to find patterns in the complexity of Amazon's reviews and product information, since he did not have access to the full relational data underpinning Amazon's data collection and analysis processes.[20] Since Amazon alone has access to the full data structure, it can offer public access to individual elements without jeopardizing its complex data surveillance system.

ASINs are flawed due to the lack of external oversight. The system privileges books as a central part of Amazon's infrastructure, but third parties might not use due diligence to correctly enter metadata. If third parties create an incomplete or faulty listing for a print book, additional records are tied to new ASINs. For example, a search for Ian Bogost, Simon Ferrari, and Bobby Schweizer's *Newsgames* in July 2017 produced nine separate product pages for the title despite the need for only two official product pages relating to the cloth and paperback editions.[21] The remaining seven product pages feature incorrect metadata about authors, titles, and publication date, with some listing the seventeenth century, and the addition of the publisher to the book's title. These product pages are created by automated vendors, which set prices according to other offerings. Most of these prices exceed the cost of purchasing directly as a consumer from the MIT Press, which suggests the merchants do not currently stock the title but will purchase at point of order. The introduction of ASIN as an all-encompassing structure requires tight control on Amazon's part to avoid such duplicitous efforts, but the evidence from faulty *Newsgames* metadata is replicated with other titles. The company does not collaborate with librarians or other information professionals who also

work on similar metadata issues. As a consequence, little consistency exists between metadata fields beyond ASIN. While BIC requires metadata about availability, territorial rights, and a cover image for books published in the United Kingdom, third parties can circumvent this rigor when uploading product data. Retail is less profitable than AWS, so Amazon has little interest in substantially overhauling these systems. These weaknesses are frequently exploited, particularly given the scale of the Amazon marketplace. For example, in 2018 it emerged that CreateSpace was used to launder money by creating fake books for real authors and using their social security number.[22] Just as with roads, electricity, and the internet, the Amazon infrastructure is designed to be used by all with little oversight on *how* it is used until problem cases emerge.

ASINs do not offer human-readable information about products, and no public documentation exists. Nonetheless, we can identify patterns in the standard. For example, Bella Forrest's monthly releases of books in the *A Shade of Vampire* series conform to a sequential pattern of the first four digits of the ASIN. *A Shade of Vampire 10* begins with B00S, and *11* starts with B00T, indicating that numbers are assigned in order but not sequentially. The first ebook editions of J. K. Rowling's *Harry Potter* series begin with B019PIOJ, indicating that they were uploaded nearly simultaneously, although the final two digits are not sequential. There is no check digit, since the ASIN for *The Philosopher's Stone* and *Chamber of Secrets* differ only by the final digit. Unlike ISBNs, publishers cannot register a cluster of ASINs so there is no continuity between publications. For example, one of the later Pottermore publications, the playscript for *Harry Potter and the Cursed Child*, has an ASIN of B073P9348D. There are also clusters around genres: the ASINs of Amazon-published public domain titles begin with B004 regardless of publication date. ASINs reflect Amazon's general commitment to data: volume and obfuscation override fidelity and legibility. Any patterns are circumstances of coincidence rather than the meaningful data points embedded in ISBNs.

Amazon's digitization projects offered further data points in the company's understanding of the book trade. Converting print is complicated by the asymmetric representation of a photograph of the page and the text it contains. As discussed in chapter 1, Amazon prompted publishers to digitize their backlists as part of the "Search Inside the Book" scheme that formed the basis of the Kindle's large launch catalog, and Mechanical Turk was an early attempt to improve the quality of digitized texts. Users will only buy high-quality scanned ebooks, so Amazon invested in accurate modeling of print initially to increase adoption. Texts reported as having bad formatting as a result of poor-quality OCR are removed from the

storefront for review. Teresa Elsey notes that 69 percent of books flagged on the Kindle storefront are the result of users reporting typos.[23]

Gathering data on common mistakes in conversion techniques allowed the company to reuse the textual data of ebooks in other ways. Amazon explored machine learning approaches to digitization through modeling contemporary language use across publishing. Early retail experiments led to the development of Statistically Improbable Phrases, an algorithm that scoured digitized material for phrases that only appeared in a single title. When a user searched for that phrase, it would appear as a result even if the phrase did not appear in the book's title. For example, searches for "nadsat" or "droog" are likely to relate to Anthony Burgess's *A Clockwork Orange*. Amazon engineers in the Digital Products group led by Sonjeev Jahagirdar also focused on using contextual learning information to improve the quality of an initial scan.[24] The patent offers an example of a Mexican restaurant menu, where a probabilistic model can determine that "taco" is more likely than "baco." Such experiments show how data from one part of Amazon's services are often reused in a different context, affirming the company's ambivalence to the book as a unique cultural object.

Amazon: Twenty-First-Century Book Historians?

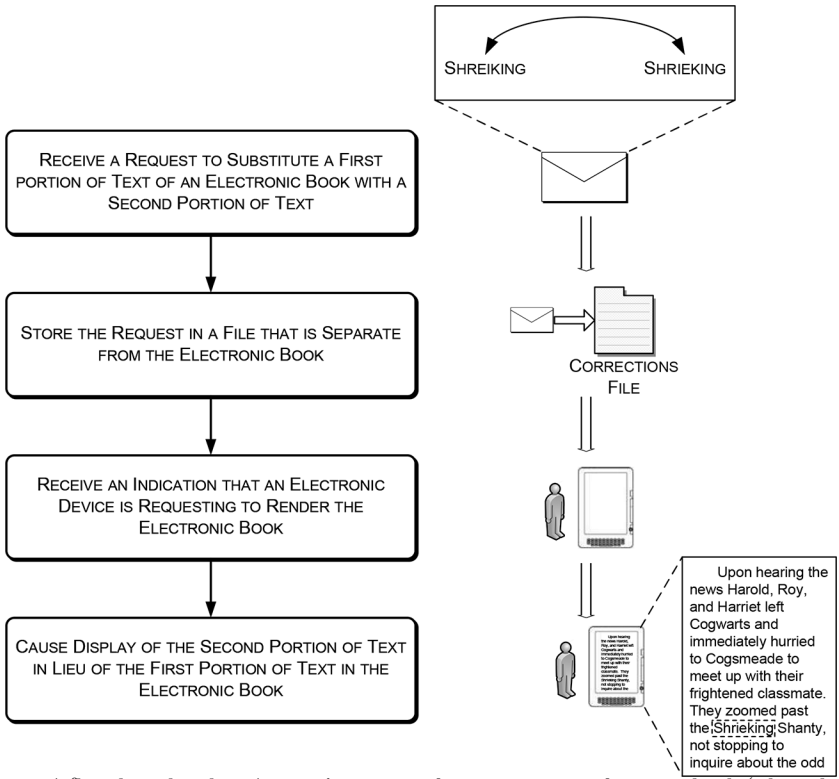
Amazon's infrastructure, which now encompasses many aspects of the production and reception of books, recalls Robert Darnton's model of the book trade as well as library science models such as the Functional Requirements for Bibliographic Records (FRBR).[25] Darnton argues that it is impossible to consider the book outside of the wider context of its production and reception, which takes the form of a network of relationships between objects such as the printing press and human producers in roles such as the author or smuggler.[26] We can trace parallels between Darnton's theoretical model of the book trade through countless years of research in the archives of the Société typographique de Neuchâtel (STN) and how Amazon has accumulated and analyzed a mass of data related to the industry. For example, Darnton's visualization of the relationship between bookshops and literary demand maps distribution in a similar manner to Amazon's profiling of warehouse locations and next-day delivery priorities. Both models analyze the needs of local markets and sociopolitical constraints such as taxation and policy.[27]

FRBR complements models of the book trade by outlining the relationship between the idea embedded within text and its containers. The model emerged from the challenges of cataloging a "book" that might exist

in various forms (audio, electronic, print) and editions (hardback, EPUB, PDF). For example, Margaret Atwood's *The Handmaid's Tale* is available as a hardcover, paperback, ebook, and audiobook, as well adaptations as an opera, television show, and film. These different formats can take a different approach to Atwood's work but are unified by the central idea of her original novel. The underlying logic of ASINs and title sets runs parallel to FRBR, as Amazon links various editions across regions and formats. The purpose of each project is different: Amazon uses "title authority" to maximize the marketing potential for a single text, while FRBR attempts to tackle ontological issues around preservation.

The Kindle pushed Amazon toward addressing traditional bibliographic problems such as linking different editions of the same book internally, called "title authority," and delivering errata to copies of a book already downloaded on readers' devices.[28] Both problems are central to the bibliographic impulse to map the development of a book and discrepancies in editions that can lead to different interpretations of the same text. Amazon's interests were more pragmatic, as the ability to map text across editions and corrections would enable consistent popular highlights to appear across editions and other Whispersync functions to act seamlessly regardless of updates or variation in regions. Janna Hamaker and colleagues on the Kindle's Reading Experience team conceptualized title authority in terms of "similarity" and "alignment."[29] An algorithm would first consider whether the books are similar, and then check the "alignment" of the text word by word on a global and local level. The algorithm produces a granular list of changes within individual editions that shows the development of a book over time that can otherwise be lost. These changes can be significant yet go unnoticed, such as Martin Eve's discovery of the discrepancies between the British and American editions of David Mitchell's *Cloud Atlas*, where Mitchell made different corrections between the two editions owing to changes in the publishers' staff.[30]

Documenting changes in content is simple compared to the potentially more intrusive effort of issuing errata. In *Merchants of Culture*, the sociologist of publishing John Thompson lauded the ability to update ebooks in digital publishing models, which could immediately correct any mistakes by altering titles on users' devices.[31] While texts can be updated automatically, users need to opt in to the scheme; otherwise Amazon deprioritizes updatability and allows readers to elect if and when to update their content. This produces ancillary files specifically containing errata, the equivalent to the print practice of errata slips if the errors were identified before the end of the production process. Amazon's default approach to errata and title authority demonstrates a commitment toward the authenticity



3.1 A flowchart detailing Amazon's process of removing errata from an ebook (adapted from patent filing). Source: Ward et al., Selecting content-enhanced applications, 9.

of documents. The book is a discrete object that can be supplemented by other files but largely exists as an individual entity. Although Amazon's data collection practices flatten relationships between objects, producers, and consumers, the actual object is still treated with reverence.

User-Generated Data

Amazon benefits from connecting bibliographic data with empirical evidence of reception. The company's methods unintentionally tested the most contentious point of Darnton's model of the book trade: the feedback loop between readers and authors. Since the company used Bowker's *Books in Print* to extend its catalog in 1994, others have followed the same steps of curation. User-generated data, including diachronic patterns in purchases, reviews, and search inputs, are more valuable commodities than bibliographic records, as they are generated by the interaction between

user and platform. Information about the book trade such as the ISBN standard has a finite boundary and can be relatively easy to capture, share, and analyze. The implicit data of consumption generate a larger data set: each book format is allocated a single ISBN, but thousands of people can purchase, read, and review that edition via Amazon, creating a substantially larger volume of data. Shoshana Zuboff has described this form of data as "behavioral surplus," which is collected and mined by large technology companies to extract value from otherwise free or cheap services.[32] Bezos understood the data's value and considered selling premium memberships to third parties in exchange for access to Amazon's cache of data. Nonsubscribers would be at a disadvantage, as Bezos stated they "wouldn't have access to Amazon's rich data and whizzy technology."[33] The scheme was pure rhetoric, and the company elected to instead keep most of its data private to build its own analytical services without sharing results with the producers.

Nonetheless, Amazon's valuable user-generated data sets exist in a semipublic fashion. Individuals' browsing histories are hidden, but recommendation engines show users how algorithms connect products for recommendations.[34] The company's services highlight some of the forms of data they are analyzing and offer a level of insight into trends in contemporary publishing. For example, the Amazon page for Neil Gaiman's *Norse Mythology* offers three data points: "frequently bought together," "customers who bought this item also bought," and "what other items do customers buy after viewing this item?" This granular recommendation engine exploits a multidimensional data set that accounts for time of purchase and confirmed or potential interest. The *Norse Mythology* recommendations demonstrate that users will likely buy bundles of Gaiman's books simultaneously, but if they do not purchase the book, they are still likely to buy similar titles by different authors as determined by Amazon's recommendation algorithm, including Graeme Macrae Burnet's *His Bloody Project* or Louise Doughty's *Apple Tree Yard* in February 2017. These recommendations constantly shift, prioritizing newer releases with high relevance as Stephen Fry's rendition of Greek mythology, *Mythos*, which appeared as the second-highest recommendation. Ed Finn concludes that "taken in aggregate, the recommendations offer a way to read Amazon's best guesses about literary desire at a given point in time, and as they are also potentially subject to influence from marketing campaigns, movie tie-ins, and the like, they operate in a feedback loop involving publishers and booksellers as well as consumers."[35] Nonetheless, the rhetorical framing of the recommendations engine is visible through the publicly available suggestions. Users can purchase more books by the same author

based on strong tie recommendations, while titles with weaker ties are noted through the looser affiliation of having been bought "after viewing this item." Amazon's focus on one-click purchasing and other methods to reduce the number of clicks to purchase is reflected only in the prepackaging of titles that have been bought together previously. The other recommendations require users to visit each page individually before adding the titles to their shopping basket, ensuring some users will not follow through with the joint purchase.

Ed Finn has documented recommendation networks for individual authors including Junot Díaz, Toni Morrison, and David Foster Wallace, but Amazon's recommendation engine at scale has largely eluded scholarly attention.[36] Julian McAuley et al. produced a data set featuring recommendations for over 600,000 Kindle titles produced in 2014 that demonstrates widespread exploitation of Amazon's clear recommendation categories for ebooks.[37] Three-quarters of the books feature at least one recommendation, with a third receiving over eighty. An asymmetry characterizes the links, however, as some books have thousands of recommendations, whereas most books are only recommended on one other product page. Most Kindle titles in the sample are loosely connected, but there remains a highly connected inner core of books linking to one another up to a thousand times, showing an instance of a group of titles frequently purchased together during 2014.

The uneven distribution pattern is typical of a network of this size.[38] The specific titles with the most recommendations (table 3.1) are more surprising, since the list does not correlate with *Publishers Weekly*'s best sellers for 2014, including a list dedicated to Kindle sales.[39] All the titles were published primarily for the Kindle, and many of the titles were self-published. Amazon is responsible for this insular view of recommendations, as only seventy-one references to ISBNs link to Kindle manuals and "how-to" guides, reinforcing the ebook marketplace as distinct from print. The algorithm is hard coded to ignore convergence between print and ebook purchases, thus favoring digital-only titles at the expense of legacy strongholds. Moreover, the frequent appearance of books produced by Mark and Aaron Shepard suggests a level of algorithmic gaming to increase their visibility in recommendations. By naming recommendations lists with their process of generation, Amazon allowed producers to focus on specific techniques to increase their visibility. A marketing campaign can request that users view pages together or download free copies of texts together to increase the likelihood the two titles will appear in mutual recommendations. The data can only reveal inherent truths about Amazon's data collection and analysis processes rather than trends within

Table 3.1 Frequently recommended titles

ASIN	Title	Inbound links
B005F9ZLD2	Mark Shepard, *Simple Sourdough* (Simple Productions, 2014)	6,831
B0057XK230	Jason Edwards, *Will Allen and the Great Monster Detective* (Rogue Bear Press, 2013)	4,212
B00CHTEMUG	Aaron Shepard, *The Legend of Lightning Larry* (Skyhook Press, 2013) [no longer available]	3,940
B00CYPKEN2	Aaron Shepard, *Pictures on Kindle* (Shepard Publications, 2014)	3,672
B005FG163Y	Aaron Shepard, *From Word to Kindle* (Shepard Publications, 2014)	3,220
B00BMHUDP2	J. S. Scott, *The Billionaire's Obsession* (Kindle Direct Publishing, 2013)	3,122
B00APM2K5Q	Earl Nightingale and Robert C. Worstell, *How to Completely Change Your Life in 30 Seconds* (Midwest Journal Press, 2012)	2,893
B00B56PP26	Steve Scott, *How to Write Great Blog Posts That Engage Better Readers* (Kindle Direct Publishing, 2014)	2,745
B00BSG4LXW	Steve Scott, *61 Ways to Sell More Nonfiction Kindle Books* (Kindle Direct Publishing, 2013)	2,651

the book trade itself. The company's dominance ensures some collapse between the two categories. Timothy Graham separates recommendations between "popular" in views and "best-selling" with the example of Veronica Roth's *Allegiant*, which averaged three stars from 8,241 reviews while sitting at the fourth best-selling Kindle spot.[40] The recommendations network in table 3.1 reflects how popularity can be gamed with the goal of improving sales, although this tactic favors short-term gain.

Amazon uses its "p13n" (personalization) data structure to generate information on elements such as titles "frequently bought together." These recommendations often come with restrictions: my account will only show items available in the United Kingdom and will prioritize products available directly through Amazon (tagged as A1F83G8C2ARO7P). Further conditions influence the display of recommendations: "frequently bought together" will only display a title if "has_seller_difference" and "has_diff[erent]_avail[ability]" are both false, restricting recommendations to books in Amazon's inventory rather than third parties. The focus on user-generated reviews demonstrates the company's commitment to producing

large volumes of data over traditionally authoritative sources. Bezos hired James Marcus and other reviewers to contribute expert opinions and suggested that the general philosophy of hiring reviewers and editors was "to seem smart and authoritative—to become not just a store but a *destination*."[41] The editors also curated the home page to highlight titles in a similar manner to the tables at the front door of a brick-and-mortar bookshop. This editorial team was later dismissed as Amazon began to focus instead on user-generated content. Bezos wanted reviews to reflect the honest opinions of customers rather than appear to be filtered, so he allowed both positive and negative reviews on the site. The current Amazon home page balances these two extremes by mixing personalized recommendations with generic advertisements. Editorial content still appears at the fringe of the front page through blogging insights such as Amazon Chart's Week in Books. Amazon's incursion into brick-and-mortar shops accelerated the trend through curating categories according to user-generated data, including "Most Wished For Cookbooks" and "Highly Rated: 4.8 Stars & Above."

Unfortunately, Amazon's popularity and the high stakes of getting good reviews on its websites led to an influx of fake or paid reviews, which Amazon is constantly battling to remove.[42] Amazon created mechanisms for user feedback to facilitate a greater curation of reviews. These include rating reviews if they are useful—which then affects a review's visibility—and the ability to respond to reviews. This has become a potent battleground in recent years, with readers using the reviews to attack authors' ideologies. For example, the release of Hillary Clinton's *What Happened* garnered a flood of 1,500 reviews within twenty-four hours of release, with a high proportion of one-star reviews coming from unverified purchases, which suggested the users were reviewing the book out of spite rather than having read the content.[43]

Daniel Allington argues that this plurality of opinions reflects "popular" rather than "literary" taste in titles like *The Inheritance of Loss*.[44] Users talked about the aesthetics of an edition or other tangents rather than focusing on the textual content. Occasionally the focus on the materiality of an edition, mixed with Amazon's enthusiasm for reusing content, can have an adverse effect on a book's product page, as criticisms of a poorly bound hardcover can translate to an additional poor score on the Kindle edition. Conversely, a poorly scanned text can have a negative impact on a luxury high-production print copy. Despite these limitations, Ann Steiner argues that the Amazon review systems "enhance [users'] feeling of belonging to a global community of readers with similar daily problems and desires."[45] This community spirit is not always apparent in review data. For instance,

only 9 of the 218 reviews of Mark Levinson's *The Box* received more than ten recommendations. Down-vote data are not publicly displayed but are evident from the position of the final two reviews, which are displayed outside the usual set of "verified purchases." User engagement is low, as only twenty comments were left for nine reviews. The most popular reviews were 745 percent longer than the mean review length and were posted early in the book's life cycle, authored by prolific reviewers in the "top 1,000" or otherwise mocked for stating "I though it would be about using Cargo Containers as homes but it was not I waisted money on this book." The prioritization of reviews by verified purchases and "Vine voices" outweighs user ranking of reviews, and the algorithm also promotes longer reviews. It is unclear whether the recommendation algorithm drives low user engagement or if the community issues emerge from knowledge that unless one is part of the several elite clubs, a review is unlikely to be promoted. Amazon reviews do not encourage extended social interactions but archive changes in taste postpublication, as early reviews are supplanted by more critical reappraisals of material later. Individuals more interested in substantial engagement with peers use websites such as LibraryThing or Goodreads that encourage more extended interactions rather than the superficial call-and-response system of Amazon's review service.

The review data therefore reveal the biases of Amazon's data collection and processing more than contemporary literary tastes. Although Amazon is the largest book retailer and clearly drives the culture, it still is insular, and much of the public-facing infrastructure reflects the company's practices rather than broader issues within publishing. The book trade exhibits an underlying distrust of the data-driven analysis of publishing trends emphasized in accounts such as Jodie Archer and Matthew Jockers's *The Bestseller Code* in doubling down on the concept of taste as a "gut instinct."[46] This gulf leads to exaggerated claims on both sides and conversations that are not productive for both parties, and as a result, the book trade is not geared for tackling the challenges presented by large technology companies. As Amazon has taken control of much of the bookish infrastructure and forced companies to adopt certain workflows, these new methods have created a degree of distrust. Mark Davis argues, "e-books can be understood as an important strategic site in a wider corporate struggle over the digital commons, and for information ownership and control, taking place among large digital corporations, and as a corporate tool in that battle."[47] Amazon has managed to increase the volume of data available for internal analysis while closing off the information from publishers, resulting in a divergent marketplace where ebooks designed for the Kindle Store succeed above those that have been developed for print distribution.

This is a section of doi:10.7551/mitpress/11985.001.0001

Four Shades of Gray

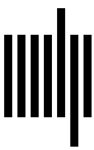
The Amazon Kindle Platform

By: Simon Peter Rowberry

Citation:

Four Shades of Gray: The Amazon Kindle Platform
By: Simon Peter Rowberry
DOI: 10.7551/mitpress/11985.001.0001
ISBN (electronic): 9780262369114
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Filosofia OT by Jen Jackowitz. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Rowberry, Simon Peter, author.
Title: Four shades of gray : the Amazon kindle platform / Simon Peter Rowberry.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Series: Platform studies | Includes bibliographical references and index.
Identifiers: LCCN 2021013279 | ISBN 9780262543507 (paperback)
Subjects: LCSH: Kindle (Electronic book reader) | Electronic book readers. | Electronic books.
Classification: LCC Z286.E43 R689 2022 | DDC 004.1675—dc23
LC record available at https://lccn.loc.gov/2021013279

10 9 8 7 6 5 4 3 2 1