How Data-Driven Decision-Making is Benefiting Independent Trade Publishers

Sophie Cox

Abstract

Historically, publishing houses have relied on gut feel, intuition, and experience to determine which titles are picked up for publication. However, with the advent of the internet a vast source of data has become available to them that by analysis can lead to identification of trends and the discovery of gaps and niches in the market. This data is especially useful for smaller publishing houses that can exploit these market opportunities.

Key Words

Data; Technology; Trade Books; Independent Publishers; Niches; Trends

Introduction

Traditionally, trade publishing has been a 'gut business', driven by intuition and experience. However, in the last twenty years, publishing has evolved alongside technology into a business that combines human intuition with scientific data in order to inform decision making. The data are nothing particularly new; however, we are only just beginning to see trade publishers take full advantage of it. With recent technological developments, data are becoming increasingly available. Between 2013 and 2014, 40% more data were available to businesses due to increased internet usage (Brookes Open Online Courses, 2018). Armed with this vast source of data, trade publishing houses can begin to move towards a data-

driven decision-making business model, where data analysis informs decisions throughout the company, from editorial, to marketing, to design. This analytical approach reduces the risks that are inherent in traditional publishing methods, meaning that more smaller companies could prosper.

A Revision of Darnton's Communication Circuit

Since the invention of the internet in the 1990s, publishing has moved away from more traditional routes of publication, based on intuition and judgement, to more technologically informed routes, based on conclusions drawn from sets of data. Book historian Robert Darnton's communication circuit (Figure 1) demonstrates the traditional publishing route, with Darnton himself claiming in 1982 that all print books share this lifecycle (Darnton, 2006).

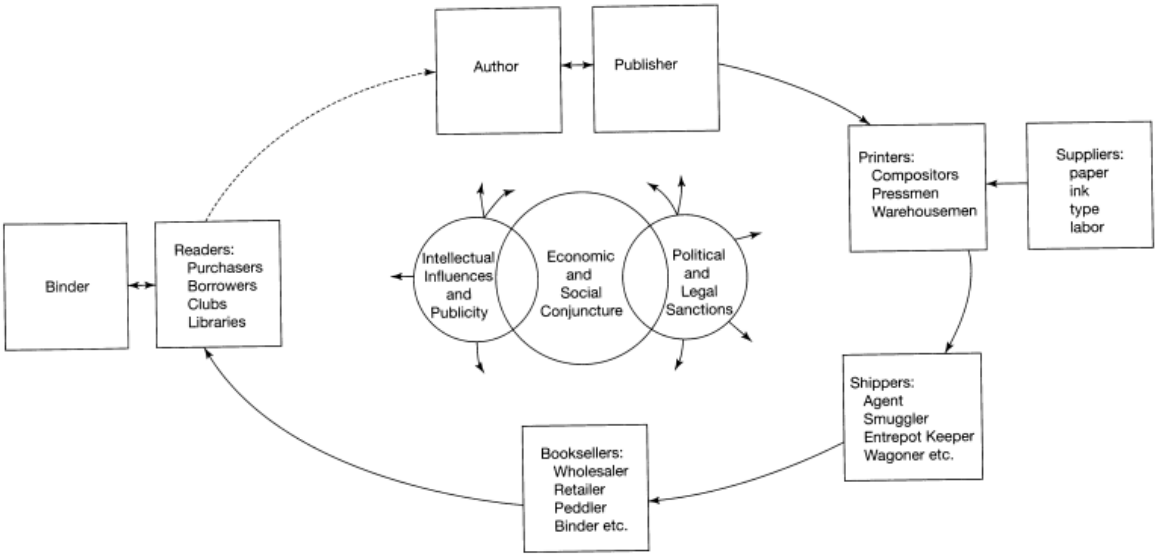

Figure 1: Robert Darnton's communication circuit (Darnton, 2006)

Darnton's circuit shows the different stages that a print book goes through in its life, and the relationship between these stages. He claims that with 'minor adjustments, it should apply to all periods *in the history* of the printed book' (Darnton, 2006 - own italics).

Indeed, the communications circuit may be applicable to all print books pre-1982, but it requires some major adjustments before it represents a contemporary print book's lifecycle. However, Darnton's communications circuit's weakness, that it often requires adjustments, is also its strength as it forms a strong basis for looking at print book lifecycles because it identifies some immutable stages in book production. In order to adapt the communication circuit to apply to today's publishing environment, technology needs to be at the centre, influencing each stage of a book's lifecycle.

The invention of the internet changed how we communicate and brought with it new technologies that altered how publishers interact with consumers. Online and digital platforms and sites provide the publisher with data that can be analysed in order to extract useful information and insights that can aid publishing decision making. Digital strategist Mark Holden calls this the extraction phase (Holden, 2017). He theorizes that there are five stages of the digital age and that we are currently in the third stage: having already found the data, and organized the data in the form of websites, search engines and OS programs, we are now learning to extract useful insights from the mass of available data in order to maximize a company's efficiency and effectiveness (Holden, 2017). As Ken Michaels, COO of Macmillan Science and Education, states, the goal of collecting and analysing data in a publishing house is to be able to 'chart better strategic business objectives, improve the effectiveness and efficiency in all parts of the business, including developing better products and audience outreach' (Lichtenberg, 2014). Darnton's communication circuit is missing this technological influence. However, Darnton recognizes the importance of the information that a reader can provide, such as 'who reads what, in what conditions, at what time, and with what effect' (Darnton, 2006). Whilst Darnton is thinking of these data as being collected from sociological studies, data collected from technological applications can provide more information that is more accurate as it is collected directly from readers' actions. In the past, focus groups and surveys have provided publishers with consumer insights; however, this type of data may be biased as a consumer may only say what they think the person asking the questions wants to hear. Technological data, collected from

social media, search engines, and similar, avoid this issue as the consumer is almost unaware that their online activities are sending data back to companies.

Darnton claims that reader data are most important to an author (Darnton, 2006). Whilst data may be useful in informing an author's writing process, it is more meaningful to a publisher as the data can help them respond to reader needs and desires by informing their decisions on what to publish and when. A publishing house also has more capacity to analyze data and extract meaningful information from it than a single author. The publisher may decide to share data insights with an author in order to encourage them to write their text in a more commercially-viable way, but it is therefore the publisher rather than the reader informing the author. In light of all of this, Darnton's circuit can be rearranged as follows (Figure 2) to show the central influence of technology on modern publishing practices and how reader data inform the publisher more than they do the author.

As Mark Holden says, publishers already have a collection of organized data waiting to be analysed (Holden, 2017). This is often referred to as 'big data'. Publishers are beginning to extract useful insights from these data that are collected from online social media platforms, websites, and ebooks. Publishers are now analysing these data in combination with more traditional consumer insight information collected from sales figures and focus groups, such as demographic factors, in order to find patterns in book trends, buying behaviour and reading communities. This analysis is then used within the publishing house to inform a range of decisions across multiple departments. By combining traditional data analysis with new analysis of technological data, publishers are enhancing their ability to determine if and how a book can be a success. Rather than simply relying on experience to make a decision, data analysis can add an extra layer of knowledge to help inform an editor, marketer, or designer how to successfully create a commercially viable book. Therefore, rather than replacing intuition and judgement, data are enhancing publishers' abilities to determine what to publish.

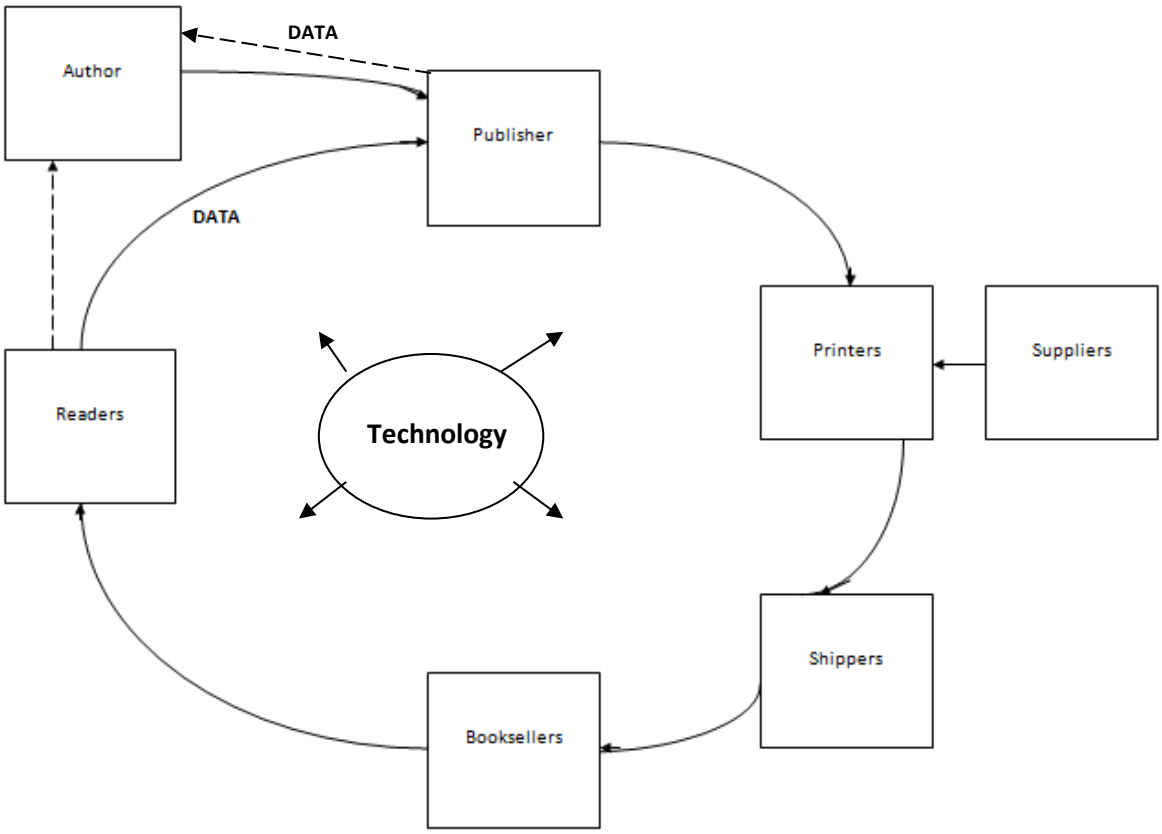

Figure 2: Revised communications circuit to include technological influence and reader data informing publisher

The goal of using big data is therefore to minimize the risks inherent in traditional publishing, in particular the financial risk. In the past, editors have been using their intuition to determine which books they publish, trying to anticipate what consumers want. However, a data analysis approach to acquisition reduces the risk that what is published is not what is wanted by the reading public. By basing acquisition decisions on trend data and audience feedback, editors can produce more commercially viable lists that cater more precisely to reader needs. This in turn reduces the financial risk of publishing a book as one can be assured that there is a market for the product and that the units will sell. A more accurate print run can also be determined based on estimated audience numbers, which will reduce printing costs and the likelihood of leftover stock.

Similarly, analysis of ebook data can inform marketing and design decisions. Ebook data is a particularly large source of consumer data. Developments in ebook technology

over the last decade has meant that publishers can obtain engagement data, meaning that they can understand how a book is used post-purchase. These data include open rates, completion rates, reading speed, identification of popular passages, and frequency and intensity of reading. By analysing these data, publishers can identify, for example, books that are viable for a marketing push. Ebook retailers, such as Kobo, analyse their reader data and produce graphs such as the following, which looks at units sold versus completion rate (Kobo, 2014):

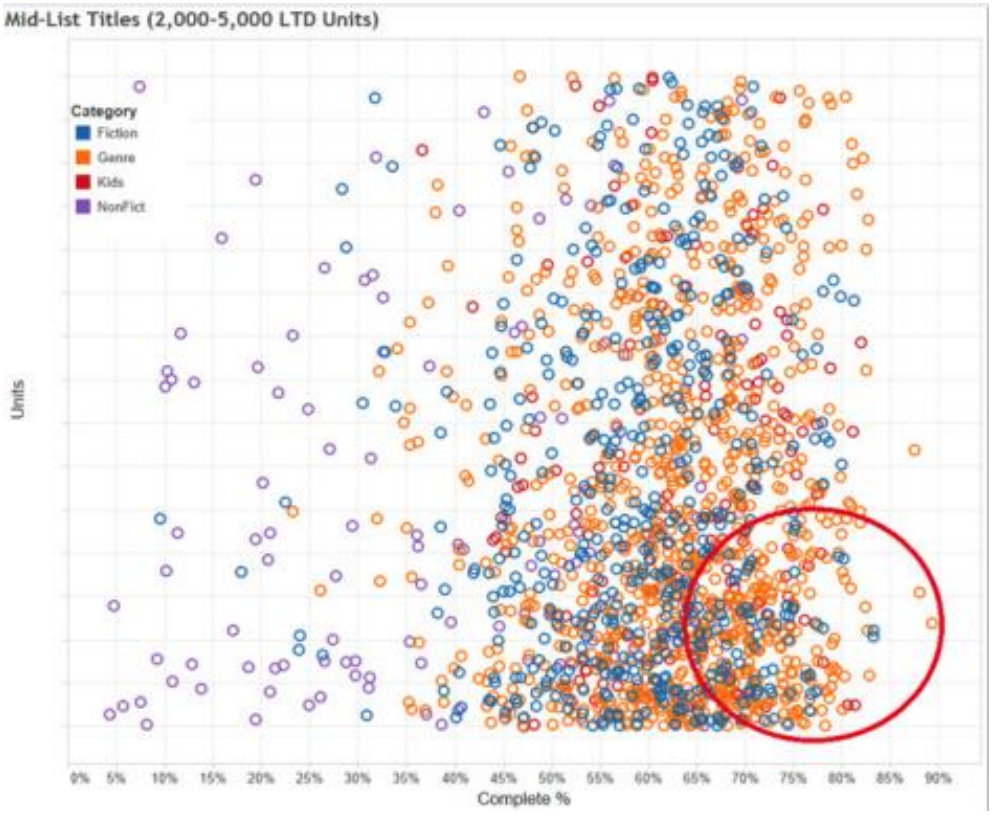

Figure 3: Graph produced by Kobo showing the unit sales and completion rate of a range of mid-list titles (Kobo, 2014)

The titles in the red circle, which have high completion rates (suggesting high reader engagement levels) but low sales, could be viable for a marketing push or a new marketing campaign. Publishers may also ask why so few units have been sold. Is the cover not eye-catching enough? Is it a niche market? Does the author need to be rebranded? This single extraction from a set of data can help publishers adjust their practices to better utilize their

budget and make the most of their products. Ebook data can also inform decisions on whether or not to publish an author's second book by looking at completion rates, and influence cover design decisions by looking at open rates. Data can therefore be seen to influence decision-making across a range of departments in a publishing house and leads to a more fluid relationship between reader and publisher. Darnton's communication circuit can therefore be modified further (Figure 4), breaking down individual publishing roles and looking at how reader data can influence decision making.

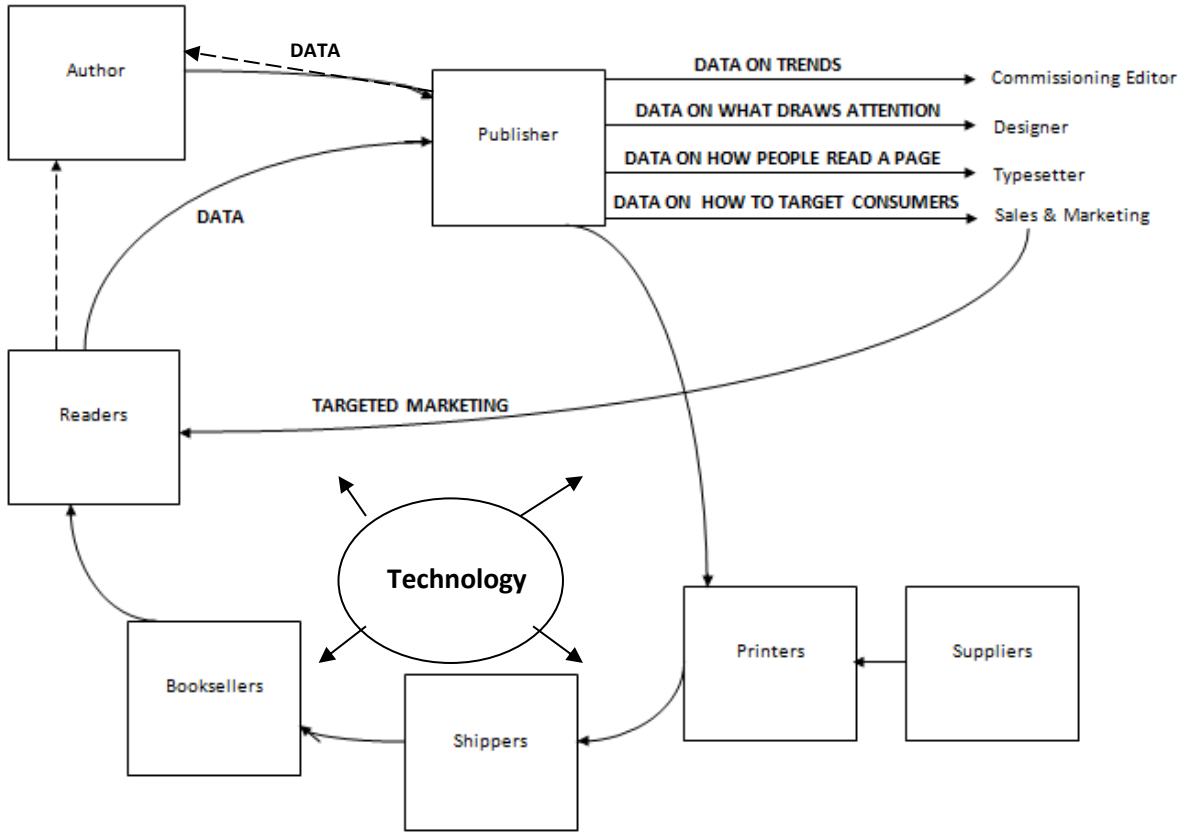

Figure 4: Revised communications circuit to show how data informs different publishing roles

Independent Publishers and Data Analysis

This data-informed business model is particularly useful for smaller publishing houses who have smaller budgets. Data analysis can assure these smaller companies that they are producing the right products to cater for consumer needs, that the marketing material is

reaching consumers in a meaningful and personalized way, and that the product design will attract consumers in bookshops. A study by McKinsey & Company showed that over a five-year period, publishing companies that focused on data analysis in order to aid decision making increased their marketing return on investment by 15-20% (Brookes Open Online Courses, 2018). This suggests how important an analytical approach to publishing could be to some smaller companies.

Analysing trend data is also a potential method for small publishing houses to identify niche gaps in the market that they can occupy. Analysing communities on platforms such as Facebook and Twitter can also help publishers identify these niche reading communities. By focusing on catering to niches, independent publishers can avoid competing directly with larger companies with bigger budgets. This could help to diversify the trade book market as more niches are being catered for. Data-driven decision making may therefore lead to a growth in independent publishing houses as these small companies can now better utilise their finances to focus on cultivating a strong commercial list.

Indeed, we are already seeing new imprints being created that are founded on data-driven decision making. Wattpad, an online library of public domain ebooks, has announced that they are setting up a new imprint called Wattpad Books. This new venture is different to traditional publishing companies in that it is using machine learning technology to curate their list of books. Wattpad Books will use an algorithm, called story DNA machine learning technology (Wattpad Books Website, 2019), to identify potential bestsellers from the 565 million stories published on Wattpad. The algorithm breaks the stories down into their elemental features (e.g. sentence structure, word use, grammar use) and combines this with Wattpad audience data to determine which stories have the potential to become bestsellers. With 70 million users, Wattpad provides the algorithm with billions of reader data points, including reading time, commenting number and sentiment, and number of reads.

Wattpad claims that this data-driven approach to publishing makes financial sense as they know that there is an audience for the books that they are publishing and they can make informed decisions on print runs so as to minimize leftover stock, similar to how other

publishers are analysing and applying ebook data to their publishing practices. Wattpad also claim that this approach can help to correct some of the systematic issues within publishing, namely the issue where 'editors who have similar backgrounds in some of the biggest cities in the world decide what the entire book buying public should be reading' (Liptak, 2019). Ashleigh Gardner, head of Wattpad Books, argues that a data-driven approach nurtures diversity by removing human bias from editorial decision making (Liptak, 2019). Indeed, with a growing concern about the lack of diversity within the publishing industry as a whole, machine learning may hold the key to how we can begin to more widely publish books written by authors from a range of cultures, races, genders, and socio-economic backgrounds. By circumventing human bias and preference, books that may not have otherwise been published via traditional publishing routes are being brought to the forefront and presented to an increasingly diverse reading public due to data analysis.

Potential Problems with Data-Driven Approaches

Whilst Wattpad's algorithm may nurture diversity among the authors of the books published, it ultimately focuses on producing commercially viable bestsellers. There is therefore the risk that the books published by a data-driven company will lack diverse content as they are chosen based on trends and patterns. Algorithms that determine whether a book is good or not based on sentence structure, grammar, and word use will pick out texts that have similar structures and are easy for readers to consume. If this approach to editorial acquisition was adopted by all publishing houses, authors such as James Joyce, Virginia Woolf, and Charles Dickens would probably never have been published due to their complex sentence structures and innovative styles of writing. These types of authors are the people who move literature forward, developing new concepts of the novel and pushing literature to the extreme. If publishing adopts algorithmic filtering of manuscripts, the next innovator may go unpublished. Furthermore, if this is the future of publishing, authors may deliberately write to conform to the standards expected by algorithms, thus muting author creativity.

Pierre Bourdieu claims that when it comes to publishing literature, publishers either produce books purely to cater to market demand, accruing market value, or to produce 'pure art', accruing symbolic value (Bourdieu, 1996). He states that these are two separate types of business and that you cannot produce a book that builds both types of value. Bourdieu states that the better publishers are 'oriented to the accumulation of symbolic capital' (Bourdieu, 1996). These publishers would have been the ones who published people like Joyce, Woolf, and Dickens, as they published based on literary merit and were not catering to a pre-existing demand. Literary titles worth publishing, according to Bourdieu, instead generate their own demand after publication. Bourdieu claims that 'An enterprise moves closer to the "commercial" pole the more directly or completely the products it offers on the market respond to a *pre-existing demand*, and *in pre-established forms*' (Bourdieu, 1996). Publishing houses and imprints whose business model centres around data-driven decision-making and data analysis are closer to what Bourdieu calls the 'commercial pole' (Bourdieu, 1996). The data analysis approach to publishing tends to be focused on catering to market trends, as the data inform editors on what readers currently want and, therefore, what type of book has a pre-existing market and will sell well. Bourdieu here also identifies the problem discussed above about books only being produced in 'pre-established forms'. He was worried that this commercial approach to publishing would lead to a lack of innovation, eventually leading to a market flooded by books that are regurgitations of one another.

The main benefit of data-driven decision-making in the publishing industry is that it minimizes the financial risk of producing a book. This is of benefit to smaller publishing houses, as discussed earlier, because they have limited finances to risk. Therefore, if they can be certain that a book will sell and be in profit, they can publish more books with their limited budget. However, Bourdieu claims that this risk-shy approach will only produce non-literary works worth little merit, simply accumulating market value (profit) rather than symbolic value. He condemns the commercialization of the book as a commodity, claiming that catering to popular demand will produce books of little consequence. He states that 'one finds... enterprises... aiming to minimize risks by an advance adjustment to predictable

demand and benefitting from commercial networks and procedures for marketing... designed to ensure the accelerated return of profits by a rapid circulation of products which are *fated to rapid obsolescence*' (Bourdieu, 1996 - own italics). If this is the case and indeed data-driven publishing companies are catering to 'predictable demand', this approach to publication may actually increase competition between publishers. Whilst at the moment smaller publishing houses are using data to identify niche gaps in the market to successfully exploit, if every company adopts this approach then more companies will spot the same trends and gaps and try to cater to them. This will not only lead to a more competitive publishing environment, but also a more limited range of content as publishers will focus on publishing books that fit in the current trends and gaps. Bourdieu's claim that these types of book, aimed at providing quick profit by catering to a mass market, are 'fated to rapid obsolescence' (Bourdieu, 1996) may indeed play out if the market suddenly becomes flooded by similar titles. Although Bourdieu is more concerned with the fact that he deems mass market books to be low-brow literature, this critique of fading into obsolescence could be the result of too many titles being published into one niche gap in the market. Over-publication into a single niche will result in none of the books catering to those readers doing particularly well as the few readers in that market gap will have too many options. This may result in competitive pricing, each publisher pricing lower than the other in order to sell units, but this will end in reduced profits despite more sales.

Instead, Bourdieu claims that publishing companies should be 'founded on the acceptance of the risk inherent in cultural investments' (Bourdieu, 1996). The books worth publishing, he claims, are ones that not only generate their own demand, but also are risky, thus pushing the boundaries of literature, and focused on long-term profits rather than immediate financial gain due to the product needing to gain traction before creating demand for itself by becoming popular (Bourdieu, 1996). However, not all publishing houses can function in this way, in particular the independent and smaller publishers. If, for example, an imprint has limited finances, they cannot afford to wait for a book to make profit as they will need to recoup their financial outlay for the book in order to continue business and publish the next title. By using data-driven decision making to publish a

handful of books, therefore, smaller publishing companies can make some quick profit in order to go on to publish some more literary titles that cater to a niche market and may be more fitting of Bourdieu's idea of a book worth publishing. If this is the case, then Bourdieu's claim that a publishing house cannot accrue both market and symbolic capital is flawed. If, in order to publish the literary and gain symbolic capital, a publisher needs to publish some mass market successes to gain economic capital, they are actually accumulating both types of value identified by Bourdieu. If the aim of a publisher is indeed to eventually publish 'literary books' in the sense that Bourdieu would think of the literary, then data analysis may prove to be invaluable as it can aid a publisher in producing books that can build their finances and enable them to publish and cater to the '"intellectual" public' (Bourdieu, 1996).

Conclusion

Data are already changing how we publish books. Data-driven decision making is beginning to become an everyday activity for a range of roles within a publishing house, changing the way we view manuscripts and how publishers shape them ready for the market. Data are positively affecting publishing practices and making processes more efficient and the outcomes more financially rewarding as publishers are catering to consumer needs rather than predicting trends and missing the mark. Data are particularly important for the smaller and independent publishing houses as they reduce the risk of producing a book by informing publishers what it is that consumers want. By reducing the risks inherent in traditional publishing practices, data are enabling these smaller companies to develop and grow, building their brand and creating commercially viable lists. However, the publishing industry needs to retain some level of human judgement. If all publishing houses move towards a data-driven business model, few risks will be taken and great books may miss out on being published. Therefore, whilst data provide a great way for new businesses and smaller companies to prosper, larger publishing companies and conglomerates who can afford to produce the books that may not currently have a market need to be urged to continue to take risks in order for literature to move forward and develop. Without risk-takers, the trade

market will stagnate and be consumed by bestseller wannabes. In conclusion, data analysis is a good thing and publishers should ignore it at their own risk. There is a place for data-informed publishing in every company, but human intuition and gut feelings will continue to guide and grow the publishing industry.

References

Barlaskar, U. 2018. Why you should be data-informed and not data-driven. *Hackernoon* [ONLINE] Available at: <https://hackernoon.com/why-you-should-be-data-informed-and-not-data-driven-76079d187989> [accessed 13.3.19]

Bourdieu, P. 1996. The Market for Symbolic Goods. In: trans. Susan Emanuel. *The Rules of Art: Genesis and Structure of the Literary Field*, pp.141-173. Cambridge: Polity.

Brookes Open Online Courses. 2018. Big data and its uses in publishing. *Moodle* [ONLINE] Available at: <www.moodle.openbrookes.net/mod/book/tool/print/index.php?id=1184> [accessed 14.3.19]

Darnton, R. 2006. What is the History of Books? In: ed. Finkelstein, D., and McCleery, A. *The Book History Reader*, pp.9-26. London: Routledge.

Flatt, M. 2019. Algorithms, trends and partnerships: the inside story of Wattpad books. *The Bookseller*. [ONLINE] Available at: <https://www.thebookseller.com/futurebook/algorithms-trends-and-partnerships-inside-story-wattpad-books-942486> [accessed 14.3.19]

Holden, M. 2017. *Merge: The Closing Gap Between Technology and Us*. London: Merge.

Hume, K. 2019. Wattpad Launches Wattpad Books, a New Publishing Division to Bring Diverse, Data-Backed Stories to Book-Lovers Everywhere. *Business Wire*. [ONLINE] Available at: <https://www.businesswire.com/news/home/20190124005084/en/Wattpad-Launches-Wattpad-Books-New-Publishing-Division> [accessed 13.3.19]

Kobo. 2014. Publishing in the Era of Big Data. *Kobo*. [ONLINE] Available at: <https://news.kobo.com/_ir/159/20149/Publishing%20in%20the%20Era%20of%20Big%20Data%20-%20Kobo%20Whitepaper%20Fall%202014.pdf> [accessed 14.3.19]

Lichtenberg, J. 2014. Big Data: Should Publishers Even Care?. *Publishers Weekly* [ONLINE] Available at: <https://www.publishersweekly.com/pw/by-topic/industry-news/publisher-news/article/62114-big-data-should-publishers-even-care.html> [accessed 13.3.19]

Liptak, A. 2019. Wattpad is launching a publishing imprint called Wattpad Books. *The Verge*. [ONLINE] Available at: <https://www.theverge.com/2019/1/24/18195753/wattpad-books-launching-publishing-imprint-self> [accessed 13.3.19]

Michel, J., and Aiden, E.L. 2011. What we learned from 5 million books. *TED*. [ONLINE] Available at: <https://www.ted.com/talks/what_we_learned_from_5_million_books/transcript?language=en#t829672> [accessed 14.3.19]

Schivas, J. 2019. Don't rage against the machine! How scholarly publishers can embrace AI. *The Bookseller*. [ONLINE] Available at: <https://www.thebookseller.com/futurebook/don-t-rage-against-machine-how-scholarly-publishers-can-embrace-ai-960256> [accessed 13.3.19]

Speranza, A. 2016. Big Data and Book Publishing: Data Mining of Literary Taste?. *Medium*. [ONLINE] Available at: <https://medium.com/@alishope/big-data-and-book-publishing-data-mining-of-literary-taste-e1bb9dde3015> [accessed 14.3.19]

Walsh, M. 2019. Netflix and the art of data-driven decision making. *The Bookseller*. [ONLINE] Available at: <https://www.thebookseller.com/futurebook/netflix-and-art-data-driven-decision-making-960261> [accessed 13.3.19]

Wattpad Books Website. 2019. [ONLINE] Available at: <https://books.wattpad.com/> [accessed 13.3.19]