# The Myth of the New: Mass Digitization, Distant Reading, and the Future of the Book

**Authors:**

**Paul Gooding**

UCL Centre for Digital Humanities, University College London, London UK

**Melissa Terras**

UCL Centre for Digital Humanities, University College London, London UK

**Claire Warwick**

UCL Centre for Digital Humanities, University College London, London UK

**Correspondence**: Paul Gooding, UCL Department of Information Studies, University College London, Gower Street, London, WC1E 6BT

**Email**: paul.gooding.10@ucl.ac.uk

## Abstract

This paper presents the theoretical background to a wider project that is attempting to increase our understanding of the impact and uses of large-scale digitization, being undertaken by the first author at University College London with the working title “What is the impact of large-scale digitization upon researchers and the information sector?” It discusses the controversy surrounding the emergence of mass digitization: the creation and collection of huge resources containing millions of pages of textual cultural content. It demonstrates that the polarized nature of the literature about this technological development is far from unprecedented, and in fact can be traced through the theory of a number of varied fields: the debate surrounding mechanization and digital technologies; our understanding of the role of the sublime in modern representations of technology; the similarities between the sociology of city life and digital information overload; and the way in which innovations are diffused throughout society. It proposes that these theories explain why debates around technological innovation often become so hyperbolic, creating an almost mythological view of technological determination. It concludes that, as a result of the processes outlined in this theory, mass digitization has become stuck between two conflicting rhetorical movements, and that it is therefore necessary to begin working to increase our understanding of this technology and to move the debate onwards using evidence from the real world.

## Introduction

In recent years, companies such as Google have generated huge interest in their attempts to digitize millions of books. Described by Crane as 'vast libraries of digital books' (Crane 2006), they provide a powerful future vision where books exist as part of a universal digital library. Google Book Search (GBS) is just one example of an explosion in digitization activity, encompassing the printed word in all its forms and including books, newspapers, court records, journals and personal correspondence. There is an active, and increasingly nuanced, discussion about the potential impact of this large-scale digitization of our cultural heritage (Lanier 2011; Nunberg 1996; Darnton 2009; Jeanneney 2007; Coyle 2006), but the popular contemporary debate is punctuated by hyperbolic claims that digital texts must inevitably destroy the print paradigm. The focus is on the novelty of digitized content, the impact of the digital upon our existing intellectual structures, and the possibilities of emerging research techniques. These methods, labelled 'distant reading by Moretti (2007), involve researchers undertaking quantitative analysis of the massive literary corpora that have been created. Important work on these corpora has occurred within the Digital Humanities, including but certainly not limited to research by Jockers (2011), Cohen (Cohen 2006) and the Culturomics research group (Michel & Shen 2010). Continued research is being enabled by the Digging into Data[1] funding stream provided by NSF[2], NEH[3], SSHRC[4] and JISC[5]. Mainstream tools such as the Google Ngrams Viewer[6] have been created as a result of this work, allowing others without advanced technological knowledge to apply quantitative methods in their own work. Despite the debate and research that is occurring, there exists

[1] http://www.diggingintodata.org
[2] http://www.nsf.gov/
[3] http://www.neh.gov/grants/guidelines/diggingintodata.html
[4] http://www.sshrc-crsh.gc.ca/home-accueil-eng.aspx
[5] http://www.jisc.ac.uk/whatwedo/programmes/digitisation/diggingintodata.aspx
[6] http://books.google.com/ngrams

very little work to define the true impact of large-scale digitization on the wider research community.

This paper presents a preliminary theoretical background to a wider research project that looks to address this knowledge gap. This wider project is being undertaken at University College by the first author, with the working title: "What is the impact of large-scale digitization upon researchers and the information sector?" There are a number of important questions that must be addressed: What impact are Large-Scale Digitized Collections (LSDCs) having on researchers and the information profession? Who is using LSDCs for research? How are they being used, for what, and how does this differ from existing research methods? And how can we apply this knowledge to ensuring that large-scale digitization develops to benefit the entire research community. These questions will be addressed in future work using a mixed methods case study approach to learn about the use and users of two large-scale digitized collections of newspapers: British Library Nineteenth Century Newspapers,[7] and the National Library of Wales Historic Newspapers and Journals Project[8]. It will collect and analyse data from a number of sources: web analytics; citation analysis; interviews with information professionals; surveys of users; and task-based workshops and user-reported behaviour. Applying an evidence-based approach will focus the debate more closely on the realities of large-scale digitization, rather than the hyperbolized debate that this paper will argue currently exists. It is important, however, to first understand the nature of this debate before we can develop an empirical framework for better understanding the realities of large-scale digitization.

This paper focusses on the theoretical background for the research described above. It proposes that we can find precedents for the hyperbolic contemporary discussion in the

[7] newspapers11.bl.uk/blcs/
[8] http://www.llgc.org.uk/index.php?id=4723

literature of many related fields: the debate surrounding mechanization and digital technologies (Benjamin 2007b [1936]; McLuhan 1962; Baudrillard 1994; Barthes 1977b; Lanier 2011); our understanding of the role of the sublime (Burke 1998) in modern representations of technology (Mosco 2004); the similarities between the sociology of city life and digital information overload (Deutsch 1961); and the way in which innovations are diffused throughout society (Rogers 2003). It will argue that we can draw on these varied bodies of work in order to understand why large-scale digitization has fostered such impassioned debate, and therefore underline the importance of adding to the literature by beginning to uncover how LSDCs are being used by contemporary researchers.

## Defining Large-Scale Digitization

While there has been a general conflation of the concepts of large-scale and mass digitization and an accompanying blurring of the two, it is certainly the case that each exhibits distinct characteristics. Large-scale digitization has increasingly come to the attention of the public in the last decade, with many public and academic institutions undertaking projects to digitize their own collections. Mass digitization, on the other hand, has been limited to the relatively few companies with the expertise, funding and scalability to digitize textual content on a massive scale. This author therefore adopts Coyle's distinction between mass digitization and large-scale digitization. She defines large-scale projects as "more discriminating that mass-digitization projects. Although they create a lot of scanned pages, they are concerned about the creation of collections and about reproducing complete sets of documents" (Coyle 2006). GBS, then, is clearly a mass digitization project, with is stated aim of creating "one giant electronic card catalog that makes all the world's books discoverable… by anyone, anywhere, anytime" (Schmidt 2005). The case studies that will form the basis for further research,

though, are both still operating at a collection level, and must therefore be considered as LSDCs.

## The nature of the debate

Despite this distinction, mass digitization is the issue that has provoked the greatest debate, and we shall see that it has been remarkably varied in both subject and tone. There is no doubt that digitization has led to a massive increase in the amount of cultural content available online (Crane 2006). This has led in turn to claims that we can now begin to consider the concept of the universal digital library, the "prospect of having an electronic wonder in the form of a virtual library of Alexandria" (Keegan 2005). This perceived scale is partly responsible for much of the excitement surrounding mass digitization. There is a sense of the novelty of the approach, an enthusiasm for this possible digital future emerging from an "attitude… that putting things on dead trees was obsolete and getting it all into a searchable, digital format was a quest that had to be accomplished someday" (Toobin 2007).

While many commenters have made realistic assessments of the benefits of large-scale digitization on scholarly activity (Mussell 2012; Hawkins & Gildart 2010; Jones 2010), others have made more grandiose claims for its impact. Anderson, for instance, claims that the era of big data renders the theoretical model obsolete: "correlation supersedes causation, and science can advance without coherent models, unified theories, or really any mechanistic explanation at all" (Anderson 2008). While he writes in relation to the sciences, a rhetorical shift towards scientific methods centred on big data is also evident in the humanities. The Culturomics research group, for example, directly compares their methods to emerging scientific fields which utilize large-scale data:

> Various fields with the suffix "-omics" (genomics, proteomics, transcriptomics, and a host of others) have emerged in recent years… These fields have created data resources and

> computational infrastructures that have energized biology. The effort to digitize and analyze the world's books has proceeded along these lines (Culturomics 2010).

This self-declared identification with new scientific methods gives the impression of a revolutionary new style of research emerging in the humanities. While most scholars have been careful to temper the more hyperbolic language, it is certainly true that this tendency has not been mirrored in some excitable media coverage (Bohannon 2011).

It is also true that there has been great criticism of large-scale digitization. A number of practical concerns have been raised relating to the technology and implementation of LSDCS: the low accuracy rates of Optical Character Recognition (OCR) software (Tanner et al. 2009; Martin 2008); inaccurate or misleading automatically generated metadata (Coyle 2009; Jackson 2008; Nunberg 2009b), and concerns over the metadata standards chosen (Nunberg 2009a; Duguid 2007); the poor quality of some page scans (Duguid 2007); and wider concerns about the continued cultural dominance of Anglo-American material (Jeanneney 2007; Grafton 2009; Hetcher 2006).

Theoretical objections have also been raised about the move towards digital access to cultural materials, criticizing the impact of the digital medium. Birkets, writing in 1994, related this shift to a concept called 'deep time': the time spent considering a text, inhabiting its words and concentrating solely on its content in order to understand it on a deeper level (Birkets 1994). Birkets claims that exposure to large quantities of digital content will destroy our attention spans, our ability to read deeply, our willingness to engage with a text for extended periods, the survival of our literary and historical narratives, and even our ability to read as individuals:

> On bad days I think… that it is inevitable that generation by generation all independence and idiosyncracy and depth will be worn away; that we will move ever more surely in lockstep,

> turning ourselves into creatures of the hive, living some sort of diluted universal dream in a perpetual present (Birkets 1994, p.32).

Birkets reaches this conclusion by presenting the physical and intellectual act of reading as a natural activity undermined by digital technologies. Wolf reaches a similarly negative conclusion through a contradictory approach. She disagrees that reading is a natural process of the human brain, and argues that it is therefore only possible because the brain effectively rewires itself to adapt to new inputs. Like Birkets, though, she assumes the overwhelming negativity of any such effects:

> Will the present generation become so accustomed to immediate access to on-screen information that the range of attentional, inferentional, and reflective capacities in the present reading brain will become less developed (Wolf 2008, p.214).

This dystopian rhetoric continues in Lanier's work. He presents a similarly damaging picture of the outcome of this shift in reading and attention, arguing that words "will be scanned, rehashed, and misrepresented by crowds of quick and sloppy readers into wikis and automatically aggregated wireless text message streams (Lanier 2011, p.xiii). The emotional power of these rhetorical arguments is evident, but they are speculative at best. Thus three writers from three different decades present similarly dystopian arguments, perhaps because there is so little evidence to substantiate or refute such claims. What evidence that does exist is often contentious or anecdotal (Carr 2008). Too little is known about the impact of digitization, and digital technologies more widely, to prove such extravagant claims for their impact, whether positive or negative.

## The Technological Sublime

This inflammatory discourse has been a constant for more than twenty years, and is one of the drivers of the controversy surrounding LSDCs. In the absence of evidence that would

feed informed theory, we remain in a transitional phase where technology develops more quickly than our understanding of its impact. Deegan and Sutherland point out that:

> The representation structures of any and all technologies… have implications for the formulation of knowledge and the reformulation of our critical engagements: that means of storage and reproduction are related; that the medium is, after all, the message (Deegan & Sutherland 2009, p.5).

Technologies often follow a recognizable transitional process in the public consciousness as language and understanding adapt to innovation (Williams 1977, p.54). Our theoretical understanding of digitization as both technology and social force is inherently incomplete, and therefore subject to the same historical forces that have influenced the adoption of other innovations.

A surprising explanation for this process exists in a modern adaptation of the 18th Century theory of the sublime:

> Whatever is fitted in any sort to excite the ideas of pain, and danger, that is to say, whatever is in any sort terrible, or is conversant about terrible objects, or operates in a manner analogous to terror, is a source of the sublime; that is, productive of the strongest emotion which the mind is capable of feeling (Burke 1998, p.36).

The work of Vincent Mosco establishes that this definition of the sublime is still relevant to the disruptive role of social discourse in technological adoption. Digital technology, he argues, can only be fully understood by recognizing some of the myths with which it is associated. This mythological status is common to the biased nature of the discourse surrounding innovations. The result is a powerful narrative of ruptures and dramatic changes, which is rarely reflected in reality (Mosco 2004, p.20). Indeed, Nunberg notes that "the past can seem an unbroken stream of proclamations that man is living in an epochal moment

(Nunberg 1996, p.10). The true status of an innovation is thus clouded by this social discourse: recognizing this tendency shows us that it is vital to engage in critical appraisal through evidence-based approaches.

The idea of large-scale digitization is particularly fraught because, without any such evidence, it lends itself easily to the creation of a corresponding mythology. It appears to go a long way towards satisfying the dreams of philosophers "haunted by the myths of knowledge and wholeness that books spawn when massed in their millions" (Battles 2004, p.214). Where understanding lags behind innovation, the rhetoric of technological determinism can fill the void, building on an illusory ideal of instant access to the world's knowledge with minimal, time, energy, and space concerns. Technological debate thus operates in a postmodern world, where "history moves by abrupt and sweeping discontinuities" (Nunberg 1996, p.10). As Duguid (1996) points out, this can make us believe the future to be more complex than the past, and foster a belief that complex new technologies must inevitably supersede their more simplistic predecessors. It is often the case that once a medium becomes commonplace, a sense of familiarity obscures our awareness of the mediation that inevitably occurs:

> Electricity and radio are, of course, still powerful forces in the world. But the Age of Electricity, like the Age of Radio, is over. Both electricity and radio have passed into powerful banality (Mosco 2004, p.20).

This banality can manifest itself, as we saw from Birkets' argument, in a mistaken belief that familiar technologies are somehow manifestations of natural human traits, therefore creating a strong emotional response towards innovations. This modern manifestation of sublimity, though, is in direct contradiction of the existing theory surrounding how technological diffusion actually works as a social process.

## Diffusion of Innovations

In this context, diffusion is understood to mean "the process in which an innovation is communicated through certain channels over time among members of a social system" (Rogers 2003, p.3). An established body of work explains how diffusion can shape the period between the invention of a technology and its widespread acceptance (Ling 2010; Rogers 2003; Wejnert 2002; Moore 1998). The success of a particular technology relies on the popular contemporary debate, taking place as it does during the period when understanding is still developing:

> Technologies do not have a momentum of their own at the outset that allows them… to pass through a neutral social medium. Rather, they are subject to contingency as they pass from figurative hand to hand, and so are shaped and reshaped (Hutchby 2001, p.441).

More specifically, human factors profoundly influence the adoption process, which resembles an S-shaped curve of adoption that relies both on the utility of a given technology and the influence of "opinion leaders" and "change agents" (Rogers 2003, p.27). Rogers describes opinion leaders as individuals who play a vital role in the diffusion process, earning and maintaining a central position in their communication system through a combination of "technical competence, social accessibility, and conformity to the system's norms" (Rogers 2003, p.33). They are responsible for influencing those within their social network, and therefore the rate of uptake among their peers. In the case of LSDCs, these opinion leaders are drawn from many overlapping sectors: academia, publishing, librarianship and the technology industry, spanning both sides of the debate. The role of the change agent differs, because they look to influence the decision-making process in a manner that benefits their own agency, whereas opinion leaders operate independently (Rogers 2003, p.366). It is therefore essential to consider the differing roles of participants in any debate; for instance,

the print publisher writing a blog against digitization, or a Google employee promoting the features of their own book search technology.

## The Power of the Medium

The arguments we see in this debate tacitly accept that the medium of representation plays a role in giving meaning to cultural expression, and that new media must therefore act as transformative forces. As Benjamin points out, the transformation of a cultural object into a new medium is vital in defining its relationship to the audience:

> The film actor lacks the opportunity of the stage actor to adjust to the audience during his performance, since he does not present his performance to the audience in person. This permits that audience to take the position of a critic, without experiencing any personal contact with the actor. The audience's identification with the actor is really an identification with the camera (Benjamin 2007b [1936], p.80).

Rather than providing us with an authentic cultural experience, each technological transformation acts as a mediation of the original, changing its meaning both subtly and profoundly. Indeed, Williams theorizes that "our culture, being materially formed, is subject to change through the changing technologies which always constitute its fundamental processes" (R. Williams 1977, p.54). Knowledge has long been mediated in this way, from the predominantly oral tradition of pre-literate societies, to the scribe-written texts of the pre-print era, followed by the transition towards the solitary consumption of a work by one defined authorial figure (Deegan & Sutherland 2009, p.29), and so specific technologies can come to be associated with a particular form of artistic expression. In this way, the concerns surrounding mechanical reproduction, and its impact upon the authenticity of cultural artefacts, foreshadowed many of the concerns that have been raised about mass digitization:

> Even the most perfect reproduction of a work of art is lacking in one element: its presence in time and space, its unique existence at the place where it happens to be. This unique existence of the work of art determined the history to which it was subject throughout the time of its existence (Benjamin 2007b [1936], p.220).

When a cultural artefact is removed from its original form, there is a danger that it will be stripped of its context, its history, and thus its authenticity. The physical artefact therefore gains importance as an authentic historical record at the same time that its ritualistic aura is eroded by reproduction. Yet this process is vital in facilitating the presentation of objects, and texts, for mass public viewing, and comes with corresponding benefits for improved education and cultural awareness. As a result, the masses are frequently exposed to a representation rather than the authentic object, and thus find their understanding mediated through this representation. The medium, as McLuhan famously claimed, is indeed the message (McLuhan 1964).

In McLuhan's narrative, print forced humanity into a state of conformity and consumption which the technologies of the Twentieth Century had the power to reverse, leading towards a technology-driven recreation of the oral tradition (Deegan & Sutherland 2009, p.9). He saw the digital realm's dismantling of physical space and time as a way to lead humanity towards a more natural mode of communication:

> We are back in an acoustic space. We have begun again to structure the primordial feeling, the tribal emotion from which a few centuries of literacy divorced us (McLuhan 1967, p.63).

But at the heart of McLuhan's narrative is a flawed conception of what this digital space would come to resemble. His conception of a global village, "a single constricted space resonant with tribal drums" (McLuhan 1962, p.63), is at odds with our experience of the digital space in its contemporary form. Whereas the village operates in a semi-closed system

with clear boundaries, the internet is unprecedented in its scale and open format. It more accurately resembles the loose structure of a modern city, a distinction that makes a big difference to our understanding of digitized content.

## The City and Information Overload

City life has provided definite cultural benefits, including massively increasing the effectiveness of human interactions. Deutsch describes this effectiveness as the probability of carrying out a specific interaction, regardless of cost; in other words, the increased proximity of other citizens means that any desired human interaction is more likely to be possible in a city than a small village (Deutsch 1961, p.101). Not only is each interaction easier, but inhabitants of a city experience a very particular freedom:

> If freedom is the opportunity to choose, then the metropolis is, in so far as it is an engine for facilitating change, is also one of choice. This liberation may be physical, in terms of the visits, the meetings, the sights now possible, or psychological and vicarious, in terms of the choices and experiments which can be made in the imagination (Deutsch 1961, p.101).

The internet shares this achievement with cities; the online world is free, in the sense that it provides people with a greater range of available interactions. This comes at a cost, though, due to the increase in exposure to external stimuli that has been described as information overload: “an inability to process inputs from the environment because there are too many inputs for the system to cope with” (Milgram 1970, p.1462).

Milgram argues that information overload can impact on daily life in a number of ways, damaging work performance and the evolution of social norms (Milgram 1970, p.1462). Deutsch further notes that when a person is confronted by increased choice, the

opportunity cost of his chosen course increases when no effective mechanism exists for filtering excessive information:

> Whatever he does will necessarily imply forgoing something else that has also appeared relevant and in a sense attractive… we are quite likely to have also increased his vague but nagging sense of self-doubt and misgiving as to whether he has made the best choice (Deutsch 1961, p.102).

Unlike McLuhan's global village, we must recognize that the internet and the city provide both increased opportunity and increased opportunity cost to individuals. The behavioural traits of internet users thus begin to resemble those of city residents. These citizens have a diminished capacity for reacting to new stimuli with the same energy as they once had: "they became desensitized as they sought to shield themselves from excessive stimuli in the form of media, ideas, communications, and so forth" (Palfrey & Gasser 2008, p.189). Milgram identifies a number of practical results of this opportunity cost to individuals, suggesting "the allocation of less time to a piece of information, the use of filtering devices, and the creation of specially designed institutions to absorb inputs" (Palfrey & Gasser 2008, p.194). Researchers have discovered similar traits among web users, with their behaviour characterized as a promiscuous and diverse reading style that bounces horizontally between various sources while spending less time concentrating on each one (Nicholas et al. 2004).

In a sense, then, the manner in which researchers interrogate the massive datasets produced by LSDCs echoes behaviour that has been recorded in other areas where information overload has occurred. There is a recognition that it is impossible to read everything that is already online (Borgman 2007, p.216), so the digital format provides readers with technological solutions to this problem. This abstraction is at the heart of quantitative cultural analysis: when textual information is provided at such a large scale, it no longer needs to be read in order to be interrogated for some kinds of meaning. But the

meaning produced by these methods differs to that of close textual analysis, relying instead on networks (Moretti 2007), trends (Michel & Shen 2010) and pattern analysis (Jockers 2012).

## Reality and Remediation

There has been a remarkable amount of resistance to quantitative methods in the humanities. There are certainly still practical issues to be solved (Jockers 2010; Sullivan 2010; Nunberg 2010), but there is also a theoretical resistance to the method. It is perhaps unsurprising that this is the case when we consider precursors in the literature. Baudrillard, for instance, warned of his concern that the process of reducing cultural objects to components in a digital network threatened the cultural boundary between truth and falsehood. Where current commentators have related this threat to the growing popularity of big data (Lanier 2011), Baudrillard saw a damaging form of abstraction originating in the technology itself:

> The real is produced from miniaturized cells, matrices, and memory banks, models of control – and it can be reproduced an indefinite number of times from these. It no longer needs to be rational, because it no longer measures itself against either an ideal or negative instance. It is no longer anything but operational (Baudrillard 1994, p.2).

It is hard to endorse the radical scepticism of a critic who claims that electronic media are not real in any meaningful sense, existing only as "a hyperreal, produced from a radiating synthesis of combinatory models in a hyperspace without atmosphere" (Baudrillard 1994, p.2). Yet he is certainly correct in one sense: a literary corpus is indeed purely operational, impossible for humans to engage with unless parse through automated tools or reconstituted into the original text. It is more accurate to suggest that, rather than representing the literary corpus as a disconnected entity can be more usefully compared to a poor translation:

> What does a literary work 'say'? What does it communicate? It 'tells' very little to those who understand it. Its essential function is not statement or the imparting of information. Yet any translation which intends to perform a transmitting function cannot transmit anything but information, hence, something inessential. This is the hallmark of bad translation (Benjamin 2007a [1923], p.69).

Thus the abstract literary corpus is a pure manifestation of what we have discussed previously. Not only is it responsible for the creation of the filtering and interrogation tools that Palfrey and Gasser argue are needed to deal with information overload (2008), but it actually relies entirely upon them for the creation of meaning. This goes further than the concept of the intertext; the work that exists as part of an intertextual network, its meaning and significance reliant on an intricate web of influences and associations (Barthes 1977a). The intertext reduces the author to a cipher for cultural ideas and focusses attention on the text, but the corpus as entity shifts meaning away from the text and towards the network. Meaning resides in the words, which then become both literally and figuratively a form of computer data. While the network is still vital to quantitative methods, the text itself is not.

This is problematic because it is increasingly clear that LSDCs must perform multiple functions, operating as both vehicles for creating these corpora and as authoritative digital archives of textual sources that meet the needs of researchers in many fields fields (Mussell 2012; Hawkins & Gildart 2010; Towheed 2010; Deegan & Sutherland 2009). In keeping with Bolter and Grusin's dismissal of the notion that digital remediations are somehow less real (1996, p.346), it is likely that most uses of digitized content, for the majority of users, operate "not as a radical break but as a process of reformulating, recycling, returning and even remembering other media" (Garde-Hansen et al. 2009, p.14). We are ultimately left with a technology that is stuck between two theoretical movements: one that prioritizes fluid, freely available corpora of digital information (Anderson 2008); and one that clings to the

representational codes of print because it considers these existing bibliographic codes to be vital to the intellectual process (Birkets 1994).

## Conclusion

This paper has demonstrated that the hyperbolic terms of the popular debate surrounding mass digitization are far from unprecedented. It traces a number of theories from different fields that, when combined, allow us to understand the direction and the nature of this extremely contemporary controversy: the concept of the digital sublime which explains how technology can inflame passion and encourage mythologizing tendencies in commentators; the diffusion of innovations which demonstrates that adoption of technologies is influenced by both this hyperbolic argument and the actions of those with a vested interest in shaping the debate; the importance of the medium as a source of meaning for our culture, and of understanding this medium in the terms it actually exists; the similarities between the city and the online world, and the way in which the similarities between city life and the online world lend themselves towards particular strategies and methods for filtering the information glut; and the way in which the literary corpus remediates existing works in new ways, while existing as part of a wider context of digital users who are likely to still be engaged in using new technology to complete old tasks (Garde-Hansen et al. 2009, p.14).

This theoretical background is important precisely because it allows us to make sense of the debate that has been outlined above. It demonstrates that many of the most strident claims regarding mass digitization emerge from a field where there is a vacuum in the space where detailed factual information would normally exist. This lack of evidence allows claims to go unchecked for long periods of time, gain traction in the literature and remain in circulation for years and decades. We should therefore, rather than engaging in the debate on its own terms, move forward from a position of rhetorical stagnation and entrenchment by

gathering evidence of real-world usage in order to fill the knowledge void that exists. Rather than framing the debate in oppositional terms, projects such as the one which this paper is part of can help to build our understanding of this emerging and exciting medium.

## References


**Anderson, C.** (2008). The End of Theory: The Data Deluge That Makes the Scientific Method Obsolete. *Wired*, published 23 July 2008: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 31 May 2011).

**Barthes, R.** (1977a). From Work to Text. In *Image-Music-Text*. London: Fontana Press.

**Barthes, R.** (1977b). The Death of the Author. In *Image-Music-Text*. London: Fontana Press.

**Battles, M.** (2004). *Library: An Unquiet History*, London: Vintage.

**Baudrillard, J.** (1994). *Simulacra and Simulation*, University of Michigan Press.

**Benjamin, W.** (2007a [1923]). The Task of the Translator: An Introduction to the Translation of Baudelaire's Tableaux Parisiens. In *Illuminations*. New York: Schocken Books.

**Benjamin, W.** (2007b [1936]). The Work of Art in the Age of Mechanical Reproduction. In *Illuminations*. New York: Schocken Books.

**Birkets, S.** (1994). *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. New York: Ballentine Books.

**Bohannon, J.** (2011). Google Books, Wikipedia, and the Future of Culturomics. *Science Magazine*, **331**. www.sciencemag.org/content/331/6014/135 (accessed 7 February 2011).

**Bolter, J.D. & Grusin, R.A.** (1996). Remediation. *Configurations*, **4**(3): 311-358.

**Borgman, C.L.** (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge Mass.: MIT Press.

**Burke, E.** (1998). *A Philosophical Enquiry into the Origin of our Ideas of the Sublime and Beautiful*. Oxford: Oxford University Press.

**Carr, N.** (2008). Is Google Making Us Stupid? What the Internet is Doing to our Brains. *The Atlantic*. http://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/ (accessed October 29, 2012).

**Cohen, D.** (2006). From Babel to Knowledge: Data Mining Large Digital Collections. *D-Lib Magazine*, **12**(3). http://www.dlib.org/dlib/march06/cohen/03cohen.html (accessed 7 February 2011).

**Coyle, K.** (2006). Mass Digitization of Books. *Journal of Academic Librarianship*, **32**(6).http://www.kcoyle.net/jal-32-6.html (accessed January 7, 2011).

**Coyle, K.** (2009). Google Books Metadata and Library Functions. *Coyles InFormation*, http://kcoyle.blogspot.com/2009/09/google-books-metadata-and-library.html (accessed 7 January 2011).

**Crane, G.** (2006). What Do You Do With a Million Books? *D-Lib Magazine*, **12**(3). http://www.dlib.org/dlib/march06/crane/03crane.html (accessed 7 January 2011).

**Culturomics** (2010). *FAQ – Culturomics*, http://www.culturomics.org/Resources/faq (accessed 15 September, 2011).

**Darnton, R.** (2009). Google & the Future of Books. *The New York Review of Books*. http://www.nybooks.com/articles/archives/2009/feb/12/google-the-future-of-books/ (accessed November 25, 2010).

**Deegan, M. & Sutherland, K.** (2009). *Transferred Illusions: Digital Technology and the Forms of Print,* Ashgate Publishing.

**Deutsch, K.W.** (1961). On Social Communication and the Metropolis. *Daedalus*, **90**(1): 99–110.

**Duguid, P.** (1996). Material Matters: the Past and the Futurology of the Book. In G. Nunberg, ed. *The Future of the book*. Berkely and Los Angeles: University of California Press.

**Duguid, P.** (2007). Inheritance and loss? A Brief Survey of Google Books. *First Monday*, **12**(8). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1972/1847 (accessed 29 November 2010).

**Garde-Hansen, J., Hoskins, A. & Reading, A. eds.** (2009). *Save As....* Basingstoke: Palgrave MacMillan.

**Grafton, A.** (2009). Apocalypse in the stacks? The research library in the age of Google. *Daedalus,* **138**(1): 87–98.

**Hawkins, R.A. & Gildart, K.** (2010). *Promoting the Digital Literacy of Historians at the University of Wolverhampton Using Nineteenth Century British Library Newspapers Online*`. http://www.heacademy.ac.uk/assets/documents/subjects/history/cs_hawkins_digitalliteracy_20100426.pdf (accessed May 9, 2012).

**Hetcher, S.** (2006). The Half-Fairness of Google's Plan to Make the World's Collection of Books Searchable. *Michigan Telecommunications and Technology Law Review*, **13**(1).

**Hutchby, I. (**2001). Technology, Texts and Affordances. *Sociology*, **45**(1): 441–456.

**Jackson, M.** (2008). Using Metadata to Discover the Buried Treasure in Google Book Search. *Journal of Library Administration*, **47**(1): 165-173.

**Jeanneney, J.-N.** (2007). *Google and the Myth of Universal Knowledge*. London: Chicago University Press.

**Jockers, M.** (2010). Unigrams, and Bigrams, and Trigrams, Oh My. *Matthew L. Jockers*, http://www.stanford.edu/~mjockers/cgi-bin/drupal/node/53 (accessed 28 September 2011).

**Jockers, M.** (2011). Detecting and Characterizing National Style in the 19th Century Novel. *Digital Humanities 2011, Proceedings.* University of Stanford, Palo Alto, July 2011. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab115.xml;query=;brand=default (accessed 14 November 2011).

**Jockers, M.** (2012). Computing and Visualizing the 19th-Century Literary Genome. *Digital Humanities 2012 Proceedings*. University of Hamburg, Hamburg, 2012.

**Jones, E.** (2010). Google Books as a General Research Collection. *Library Resources And Technical Services*, **54**(2): 77-89.

**Keegan, V.** (2005). A Bookworm's Delight. *The Guardian*. http://www.guardian.co.uk/technology/2005/oct/21/comment.bookscomment (accessed November 30, 2010).

**Lanier, J.** (2011). *You are not a Gadget*, London: Penguin.

**Ling, R.** (2010). From the Telegraph and Telephone to the Negroponte Switch. In W. R. Neumann, ed. *Media, Technology, and Society: Theories of Media Evolution*. Michigan: University of Michigan Press.

**Martin, S.** (2008). To Google or not to Google, That is the Question: Supplementing Google Book Search to Make it More Useful for Scholarship. *Journal of Library Administration*, **47**(1): 141–150.

**McLuhan, M.** (1962). *The Gutenberg Galaxy: the Making of the Typographic*. Toronto: University of Toronto Press.

**McLuhan, M.** (1964). *Understanding Media: The Extensions of Man*. London: Routledge and Kegan Paul.

**McLuhan, M.** (1967). *The Medium is the Massage*. Harmondsworth: Penguin.

**Michel, J.-B. & Shen, Y.K.** (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science Magazine,* **331**(6014): 176-182.

**Milgram, S.** (1970). The Experience of Living in Cities. *Science*, **167**: 1461–1468.

**Moore, G.A.** (1998). *Crossing the Chasm: Marketing and Selling Technology Products to Mainstream Consumers.* Second ed. Chichester: HarperCollins.

**Moretti, F.** (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London and New York: Verso.

**Mosco, V.** (2004). *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge Mass.: MIT Press.

**Mussell, J.** (2012). *The Nineteenth-Century Press in the Digital Age*. Basingstoke: Palgrave MacMillan.

**Nicholas, D. et al.** (2004). Re-Appraising Information Seeking Behaviour in a Digital Environment: Bouncers, Checkers, Returnees and the Like. *Journal of Documentation*, **60**(1): 24–39.

**Nunberg, G. ed.** (1996). *The Future of the Book*, Berkely and Los Angeles: University of California Press.

**Nunberg, G.** (2009a). Google Books: A Metadata Train Wreck. *Language Log*, http://languagelog.ldc.upenn.edu/nll/?p=1701 (accessed 6 January 2011).

**Nunberg, G.** (2009b). Google's Book Search: A Disaster for scholars. *The Chronicle of Higher Education*. http://chronicle.com/article/Googles-Book-Search-A/48245/ (accessed November 29, 2010).

**Nunberg, G.** (2010). Counting on Google Books. *The Chronicle of Higher Education*, http://chronicle.com/article/Counting-on-Google-Books/125735 (Accessed 11 September 2011).

**Palfrey, J. & Gasser, U. (**2008). *Born Digital: Understanding the First Generation of Digital Natives*, New York: Basic Books.

**Rogers, E.M.** (2003). *Diffusion of Innovations.* Fifth ed. New York: Simon & Schuster.

**Schmidt, E.** (2005). The Point of Google Print. *Official Google Blog*, http://googleblog.blogspot.com/2005/10/point-of-google-print.html (accessed 22 November 2010).

**Sullivan, D.** (2010). When OCR Goes Bad: Google's Ngram Viewer and the F-word. *Search Engine Land*, http://searchengineland.com/when-ocr-goes-bad-googles-ngram-viewer-the-f-word-59181 (accessed 7 February 2011).

**Tanner, S., Munoz, T. & Ros, P.H.** (2009). Measuring Mass Text Digitisation Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. *D-Lib Magazine*, **15**(7/8). http://www.dlib.org/dlib/july09/munoz/07munoz.html (accessed May 6, 2011).

**Toobin, J.** (2007). Google's Moon Shot: The Quest for the Universal Library. *The New Yorker*. http://www.newyorker.com/reporting/2007/02/05/070205fa_fact_toobin?currentPage=all (accessed November 25, 2010).

**Towheed, S.** (2010). Reading in the Digital Archive. *Journal of Victorian Culture*, **15**(1): 129–143.

**Wejnert, B.** (2002). Integrating Models of Diffusion of Innovations: A Conceptual Framework. *Annual Review of Sociology*, **28**: 297–326.

**Williams, R.** (1977). *Marxism and Literature*. Oxford: Oxford University Press.

**Wolf, M.** (2008). *Proust and the Squid: The Story and Science of the Reading Brain*. Cambridge: Icon Books Ltd.