Writing History in a Paperless World

Archives of the Future

Kaur, Ravinder

ARCHIVES AND SOURCES
Writing History in a Paperless World: Archives of the Future
by Ravinder Kaur

The centrality and certainty of paper has for long been a given in our imagination of archives. The carefully preserved documents – books, files, notebooks, private letters, identity cards, memos, index cards, charters, declarations, petitions – all neatly numbered and catalogued – are intricately tied to the idea of archives. Paper as the primary medium of documentation, record and knowledge circulation has recently been the subject of extensive scholarly attention. Consider Lisa Gitelman's rich historical account of the life of the document in paper and near-paper form, or 'paper knowledge' as she calls it, and how it offers a detailed insight into the inextricablility of materiality of paper from the art of documentation.[1] The well-established association of archives with dust[2] on paper documents and government custody[3] of those has been rigorously explored. Paperwork – its very nature, its contradictions and unpredictability and how it underpins the edifice of bureaucracy – is the subject of two compelling pieces of historiography: Ben Kafka's account of eighteenth-century France during the Revolution[4] and Matthew Hull's history of government in urban Pakistan in mid twentieth century.[5] Just as we begin to understand the work of paper in organizing our lives, the spaces we inhabit, and the ways in which we interact with public authorities,[6] we seem to be entering a paperless world increasingly defined by biometric identification,[7] digital documents, instant messages, and a new form of public sphere via social media performed on the Internet. It seems the usual paper trail that the present leaves behind for the historians might be thinning out, or at least jostling for attention and space in competition with its digital form.

The question I want to pose here concerns the form of archives that will be available to the historians of the early twenty-first century. Or put differently – what will be left behind of the contemporary present in lieu of paper for the future historians? The larger question relates to the project of history writing, or how we might rethink the notion of the past itself in an accelerated digital era of fast-moving social media. To be sure, concerns about the archives of the future have been raised in different forms before. Already more than a decade ago, the historian Roy Rosenzweig warned about the problems of 'preserving our digital cultural heritage' given that ways to archive the digital present were still at the development stage.[8]

University of Copenhagen   *rkaur@hum.ku.dk*

He articulated this around the dual problem posed by rapid digitalization – the overload of information on one hand, and the scarcity of archival records on the other. In 2003 when Rosenzweig pointed out the problems of future archives to his fellow historians, the digital world was still gathering steam. His worries about the 'rapid accumulation of data' at that point, for example, stemmed from the increasing presence of Google search engine, and how it indexed and ordered websites and returned large amounts of information to search queries. When Google was started in 1998, it received about 10,000 search queries each day; by 2004 that figure had grown to 200,000. In the intervening decade, however, the speed and extent of Google search engine has multiplied at an unimaginable scale. In 2014, Google processed an average of 40,000 queries per second, 3.5 billion queries per day, and 1.2 trillion queries per year.[9] The size of the Internet itself has grown exponentially in the meanwhile – from a single website in 1991 to about a billion in 2014.[10] Of these, more than 670 million websites are currently live, and about 103 million were added to the Internet in 2013 alone. The number of live webpages in 2013 was estimated to be 14.3 trillion and 328 million domain names were added that year.[11] The number of Internet users has grown from 16 million in 1995 to 2,937 million in March 2014.[12] In short, the digital world has expanded on a scale that Rosenzweig could barely have anticipated just a decade ago.

What appeared to be 'information overload' then, now turns out to be just a small fraction of the seemingly limitless digital universe that is yet unfolding. If at all, the spectre of 'information overload' or 'information deluge' or 'infoxication' is more a signal of being overwhelmed, of our inability to keep up with new technologies and the accelerated pace of data generation, accumulation and circulation.[13] In modern work culture, information overload is now even classified as a serious problem, since the immediate availability of too much data is equated with superficial knowledge, and thereby with diminished human ability to take proper decisions.[14] As we grapple with ever increasing levels of information difficult to keep pace with and fully control, we need reminding that technology shift is in fact a constant in human history. Each dramatic shift – say from mass print culture to audio/visual, and more recently from emails to social media – leaves behind a specific trail and footprint that requires constant transformation of archival tools and practices. The digital era is indeed vastly different from the print era, and that's not just in terms of materiality but the very nature and scale of data production and consumption. Probably what has altered in the decade since Rosenzweig's first astute warning to historians is the seemingly endless expansion of the digital universe. As the scale of information generation, duplication and circulation continues to grow at ever accelerating speed, we might ask if technology is the only solution to the problem of recording and storing the data for future use. Rosenzweig's response to the archive problem was primarily to think in terms of technological solutions. Indeed while technology can be harnessed to shape

appropriate archival practices for mining the past, the notion of time and shifting temporality that constitutes the past has itself been accelerated in the high-speed present. The question of archives of the future, I suggest, is thus intricately connected to how the past is imagined and constructed in the accelerated contemporary. In what follows, I lay out a twofold account of the nature of data generation and consumption, the challenges of archiving the past in an ever-accelerating digital era, and two approaches to the future's past.

FORMS OF PAPERLESS DATA

In 2007, the sixtieth anniversary of India's Independence was marked by a high-profile publicity campaign called 'Lead India' that invited the citizens to lead the nation to its long-awaited glorious future. The campaign was not initiated by the Indian state but sponsored by the Times group, India's largest and most profitable media corporation. It featured extensively in print in *Times of India* in a series of political advertisements, news items and accompanying editorials that drew attention to the fragile state of the nation. The theme was also turned into a nationally televised reality show that ran over eight weeks to choose a leader for an increasingly impatient nation. The entire campaign acquired its own digital home on a domain called www.lead.timesofindia.com where print advertisements and audio/visual material were put together. This website was also a place where citizens could interact on discussion forums and contribute commentaries or their motivational stories of fighting the corrupt government. The website was one of the most popular destinations for disaffected citizens who wanted to mobilize change in the system. Some time in 2009, the website disappeared in its entirety. From then on visitors found a 'service unavailable' message on an otherwise blank page. A few bits and pieces survived in the shape of youtube videos, screenshots posted on blogs, or other websites discussing the Lead India campaign. The reasons for the removal of the website have never been clear. The only brief explanation from an employee of Times of India stated that the campaign did not fit the corporate strategy of the company any more.[15] (See fig. 1.)

The disappearance of the digital world of Lead India is by no means the first time a website or a domain name has disappeared. In fact, the Internet is littered with dormant sites or the tell-tale signs of websites no longer in existence. Yet Lead India was no ordinary site. It was not only a repository of print and audio-visual material, but also the place where popular mobilization took shape through citizen participation. For the historians of the future, the lost website signals the loss of a highly important source through which to capture the currents of Indian politics/media in the early twenty-first century. It also helps us understand the nature of data – lost and available – and the possible shape of the archives to come.

Data generation in digital form is characterized by two mutually constitutive features – duplication and excess. In the case of Lead India the core
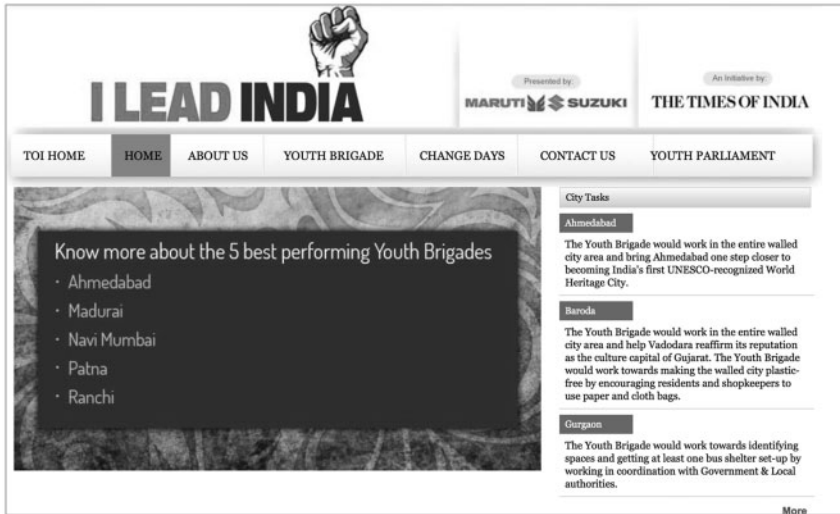
Fig. 1. A current 'Lead India' screenshot from its 2014 version. The original website from 2007 and its revision in 2009 have disappeared. This version has added 'I'. http://timesofindia.india-times.com/ileadindia.cms, accessed 12 Sept. 2014.

data, campaign advertisements, was generated simultaneously in both print and digital form. This dual form, in fact, characterizes the vast majority of newspapers, magazines and books that seek to reinvent themselves in the digital age. The digital format allows for endless duplication and circulation at almost no cost and a quicker pace even as the traditional paper circulation and modes of photocopying remain intact. This means a part of the digital material remains available in traditional modes and archived as such. Yet websites also produce data that is in excess of the paper form, first in discussions and commentaries, 'web exclusives' and supplements to the print editions, and second as the content on popular social media sites such as Facebook, Twitter, Instagram and Tumblr. Unlike in the traditional media, in social media the data is entirely user/reader generated. This digital platform is used for interaction with friends/acquaintances/ strangers not only on mundane aspects of everyday life – sharing images and commentaries on daily occurrences – but also on spectacular events that usher in dramatic upheaval. Examples range from from images of dinners served at a family table, portraits of children (and more recently, even scanned ultrasound images of unborn children), or places visited, to mobilizing political opinions, sharing information about political protests, and communicating with prospective voters and consumers. In the recent political upheavals in the Middle East – from the Green Revolution in Iran to the Arab Spring and beyond – social media was a key tool of mobilization among the increasingly techno-friendly youth population. Similarly, the 2014 general election in India was contested as much on the digital interface as on the ground, as free phone applications like 'Whatsapp' gained popularity across the urban/

rural divide.[16] Governments around the world now use digital platforms as a 'good governance' practice to make the workings of the state more accessible and transparent. In many parts of the world, ministries, government agencies, politicians, public and private organizations now strive to have a 'digital presence' and make use of social media to create a semblance of accessibility and information distribution in the public sphere.

The question that has still to be successfully addressed is how to collect and record the ever-increasing data generated via various digital platforms. In this regard, a particular feature of this new form of data becomes significant: namely the possibility of quantifying the impact and reach achieved by each bit of information. In this 'economy of approval', the weblinks, photos and places visited, along with messages and even responses to the original posts, are opened up for scrutiny. If Facebook users are encouraged to use the 'like' and 'share' feature to demonstrate their engagement, Twitter users can likewise 'retweet' or 'favourite' the messages. This adds a secondary, and an important, layer of data to the original bits of information, that tells how they were received in the public sphere. Thus, what is of significance here is also the constantly changing order in which the information
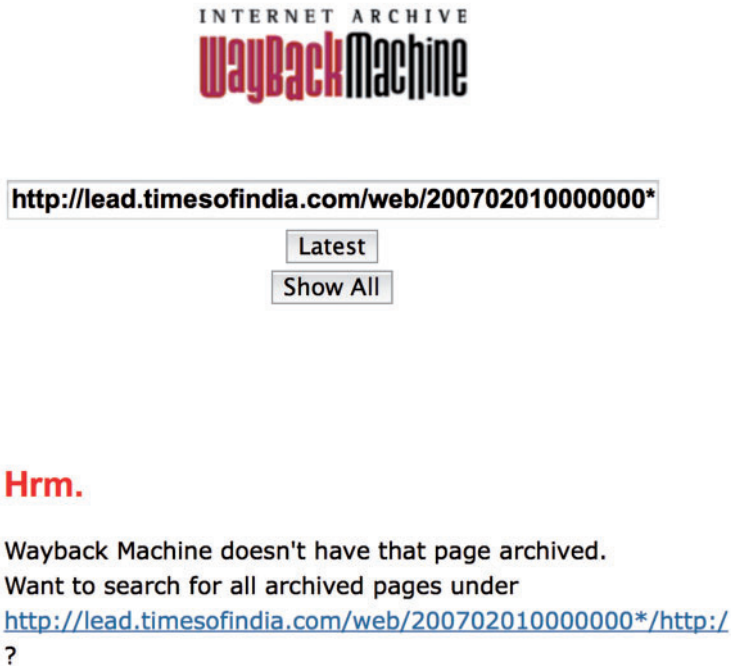

INTERNET ARCHIVE
WayBackMachine

http://lead.timesofindia.com/web/200702010000000*

Latest

Show All

Hrm.

Wayback Machine doesn't have that page archived.
Want to search for all archived pages under
http://lead.timesofindia.com/web/200702010000000*/http:/
?

Fig. 2. Message about the defunct 'Lead India page', on Wayback Machine (www.archive.org) accessed 6 Sept. 2014.

appears for the readers. It is adjusted and tailored according to user behaviours. The news-feed algorithms derived from user data – likes, the time of the day, and the number of times specific stories are accessed – mimic individual patterns of consumption. This not only includes prioritizing certain stories in the news feed and highlighting 'trending' stories, but also adding user-specific advertisements and sponsored stories according to prior use.[17] Data on social media is therefore not identical or standardized on any two given screens. This aspect makes social media very different from sources like newspapers, magazines and books that are available to the reading publics in a standard form. The account of nations and nationalisms as 'imagined communities' offered so persuasively by Benedict Anderson just a few decades ago is now open for scrutiny once again. The digital universe allows for a variety of specific and specialized reading publics that are not constrained or shaped by national boundaries but rather by connectivity to the Internet.[18]

A question raised by this specific digital form concerns ownership of the data and the contemporary debates on user privacy. Unlike the traditional archives, that are maintained and mediated by government bodies, the digital data especially where user-generated is more likely to be controlled and mediated by private corporations. The data at stake here is not only what is available on the digital surface – posts, news-stories, photos – but also the 'deep data', the analytics and algorithms of user behaviour, patterns, locations that ground the information in real-time context. The information that used to be collected via censuses and surveys by the nineteenth and the twentieth-century states as a form of governmentality is now mined by corporations that track search words on Google in order to pitch advertisements for matching products. This aspect of digital data generation has been dealt with in the public discourse as 'death of privacy' that inevitably follows the act of sharing personal information on the social media.[19] Despite heated public debate on privacy and ethics, the lines between corporate ownership, copyrights and the privacy of users still remain blurred.

THE PAST IN THE AGE OF ACCELERATION

If a prime characteristic of paperless data is the accelerated speed at which it is manufactured, duplicated and circulated, then its customized, multi-layered and deep quality makes it a particularly complex resource to handle. The challenge for future historians will be to capture data of this nature in not only its full extent but also its depth. If the digital form allows for more and more detailed information to be revealed, that information is also more difficult to hold on to. The deletion of websites, posts, tweets, blogs and the accompanying layers of data are all too frequent events in the digital world. One approach, thus far, has been to harness technology to assemble the digital archives. Since the early 1990s several attempts have been made to preserve electronic records on the Internet. These include the Pitt Project (1993–6), the non-profit Internet Archive (www.archive.org, started in 1996), and Alexa Internet (www.alexa.com), its for-profit

branch that was later bought by Amazon. Methods of preserving include sending crawlers to the Internet to amass the data, and taking periodic snapshots. Yet the drawback of even an extraordinary service like the Internet Archive's Wayback Machine is that snapshots are in some instances far too few and may well miss important events. Thus they do not always capture the full scope of whatever one sets out to explore. And where the websites are no longer active, there is little the Wayback Machine can do to retrieve the lost material (Fig. 2).

Another notable initiative is 'The Net Archive' www.netarkivet.dk, launched by the Royal Danish Library in 2005, which aims to preserve the Danish cultural heritage in the digital age. Rather than focus on the entire web, it targets only the Danish-language domains. Three strategies are deployed to harvest the present for the future: first, collecting all domains in the Danish language four times annually; second, selective harvesting of 80–100 domains on a daily basis; and third, harvesting material pertaining to specific events. While this systematic collection makes available a rich and extraordinary archive for the future, it still does not capture the full depth and breadth of the Internet in the Danish language. On a global scale, such initiatives to preserve the Internet are still far too few and cannot keep up with the rapidly increasing size and scale of the Internet. This means that much knowledge, especially in the non-Western world, risks being lost as quickly as it is produced. This is especially evident in a country like India where the digital presence of the government, public institutions and the public has become almost ubiquitous in the past decade. Yet hardly any attempt has been made to systematically preserve this material.

The fragile nature of the information made available on the Internet became evident to me as I set out to write the history of India's transformation after the economic reforms of the 1990s. This period overlaps with the era of burgeoning digital information, when the availability of information on government websites was fast becoming a norm (Fig. 3).

It was not just actual documents that were being posted in digital format, but wholly original material, produced continually for the consumption of web readers. One organization that I had been following closely was the India Brand Equity Foundation (www.ibef.org), responsible for creating a 'smart look' for India to promote it as an 'attractive investment destination' in the global political economy. IBEF was available to its publics as a web portal, on Facebook, Twitter, Youtube and Google Plus. Despite its extensive presence on various media, IBEF turned out to be unstable as far as archives were concerned. And this aspect was revealed to me only because I had been following the organization on different media and could see the changes taking place. Not only were documents and materials arbitrarily removed, the very goals and aims of the organization were revised frequently. It was as if the digital format permitted a circumvention of the committees and bureaucratic rituals that would otherwise govern the work of such a semi-government body. Thus, revisions and alterations that on
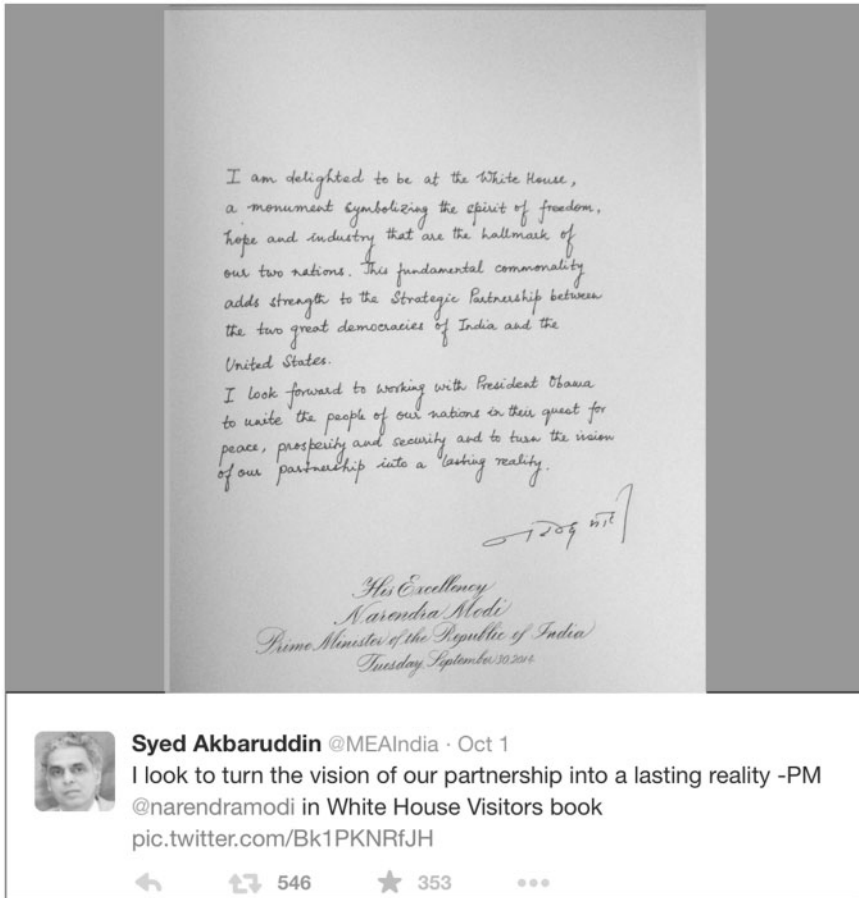
Fig. 3. Indian Prime Minister Narendra Modi's letter of thanks to President Obama, after his visit to the White House, was distributed widely on Twitter via the official handler of the Ministry of External Affairs, Government of India.

paper would take months or years were now being performed digitally in a matter of days, and with the significant difference for readers, or future historians, that there was no trace of what was before. While paper material is governed by bureaucratic practices, or rather contemporary bureaucratic practices are shaped by paper, the digital format has so far no protocol for preservation of information distributed in the name of the state. IBEF maintains almost no archive of the information distributed on the Internet in its formative years. In this case, the domain name was never removed or erased (as with the Lead India website), but its contents continued to change constantly – the shell remained while the inside was altered.

As I became more aware of the shifting nature of the digital material, I learnt to take screenshots, to save 'favourite' or 'like' material in order to create my own archive. This meant that I was now 'friends' with India, the nation, just as I was friends with or follower of various ministries, relevant organizations and key persons. The work of the historian is no longer that of someone who visits the archives: it now includes that of an archivist too. This is an important development as far as the project of history-writing in the twenty-first century is concerned. In response, new digital products have become available over the past few years for those interested in digging into and preserving the past while systematic archiving still remains a distant goal. These not only include easy online tools to take screenshots, Google alert services to get customized information and news, and large storage space on Google drive or Dropbox, but also analytic services to measure the reach and extent of specific websites, posts, tweets and most digital material already available. While some of these services are free, many others, especially those that offer detailed analytics for business improvement purposes, are professional. Google Analytics, Tweetreach.com, Sharedcount.com offer basic free services and also advanced services for a fee. Yet even these tools are not completely reliable: many of the services do not maintain records for more than a few months at a time.

As long as the technology and will to preserve the data lags behind the technology to generate and distribute data easily, the digital present in its entirety can probably never be fully preserved. Much of this indeed also has to do with funding required to preserve the digital material. In Denmark and Norway, for example, full public funding for preserving the digital information is available and this means that extensive digital archives are now stored for future history-writing. In places where such means are not available, or such investments are not made in good time, the rich present in the digital format may not be available in its full extent to future historians. This leaves the project of history writing in the Global South especially vulnerable as few attempts have been made to preserve the digital present. In India, where the past is a particularly contested domain, lost digital data in the early twenty-first century represents a considerable challenge for future historians. Perhaps the problem of preserving the present requires more than a

technical solution, as I learnt in my attempts to write the history of 'new India' (as the post-economic reform period is popularly called).

This calls for us to consider a different approach, probably a more radical move that requires a larger discussion among historians. The point is simple: how do we separate time or rather delineate the past in an age of acceleration. How do we engage with the past that is rapidly buried under fresh layers of information and news almost every minute and every second on multiple media? Do we not need to revise our notion of the past itself if we are to get a full sense of the contemporary before it is deleted or lost? What I am suggesting is that historians also need to begin engaging with the past just as it unfolds from the present. The nature of the future's past is as fragile as it is plentiful and seemingly excessive. The old notions of speed, acceleration, novelty, time/space compression and their implications for modernity have long been articulated, sometimes in awe and sometimes in anxiety.[20] Yet unlike the previous moments of modern *neuzeit* that have recurred endlessly in history and left behind their own range of archives, the current moment is leaving a deep trail in digital format that is both extensive and frail. The past, as Koselleck told us long ago, is no longer static, but it is accumulating and perishing at an accelerated rate that he could have barely imagined. The implications for the project of history-writing in the digital era are obvious – historians will not only have to become archivists but engage with the contemporary present if the past of the future is to be written.

Ravinder Kaur is Associate Professor of Modern South Asian Studies at the Department of Cross-Cultural and Regional Studies, University of Copenhagen. She is also Visting Professor at Centre for Indian Studies in Africa, Witswatersrand University in Johannesburg. Her current research focuses on the history of India's transformation from a postcolony to 'emerging market' in the global political economy.

NOTES AND REFERENCES

    1 Lisa Gitelman, *Paper Knowledge: Toward a Media History of Documents*, Durham, 2014.
    2 Carolyn Steedman, *Dust: the Archive and Cultural History*, New Brunswick, 2002.
    3 Ann Laura Stoler, *Along the Archival Grain: Epistemic Anxieties and Colonial Common Sense*, Princeton, 2009.
    4 Ben Kafka, *The Demon of Writing: Powers and Failures of Paperwork*, Cambridge, 2012.
    5 Matthew S. Hull, *Government of Paper: the Materiality of Bureaucracy in Urban Pakistan*, Berkeley, 2012.
    6 See for example, Kamal Sadiq, *Paper Citizens: How Illegal Migrants acquire Citizenship in Developing Countries*, New York, 2010, on acquisition of false paper documents as a migrant strategy to avoid state scrutiny. Hull in *Government of Paper* describes the centrality of paper in organizing urban spaces.
    7 Consider the *Adhaar* programme in India. With over 700 million registrations thus far, and ambitions to reach a billion users by 2015, it is the largest biometric identification programme in the world. See www.uidai.gov.in (accessed 10 Sept. 2014).

8 Roy Rosenzweig, 'Scarcity or Abundance? Preserving the Past in the Digital Era', *American Historical Review* 108: 3, 2003. Available in digital form on http://chnm.gmu.edu/digitalhistory/links/pdf/introduction/0.6b.pdf (accessed 10 Sept. 2014).

9 See 'Google Search Statistics' for a live update: http://www.internetlivestats.com/google-search-statistics/ (accessed 11 Sept. 2014).

10 See 'Total Number of Websites', on http://www.internetlivestats.com/total-number-of-websites/ (accessed 11 Sept. 2014).

11 The number of websites continues to fluctuate to account for dormant sites as well as new sites that are added on a running basis. See 'Size of the Internet as of 2013', http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html (accessed 11 Sept. 2014).

12 'History and Growth of the Internet from 1995 till Today', http://www.internetworld-stats.com/emarketing.htm (accessed 11 Sept. 2014).

13 Concerns about 'information overload' have emerged with each moment of technological advancement – from paper to emails to the digital era. See James Gleick, *The Information: a History, a Theory, a Flood*, New York, 2012.

14 As if to reaffirm the fear of too much information, the search query 'information overload' on Google generated 12.5 million results in 0.43 seconds.

15 Interview with *Times of India* employee at the branding department of the company, 9 Nov. 2013, Delhi (name withheld by request). A new campaign that followed shortly, called 'I Lead India', mimicked the logo and imagery of the original campaign to some extent, but was markedly different. See http://timesofindia.indiatimes.com/ileadindia.cms (accessed 12 Sept. 2014).

16 The surge in numbers of mobile phone users in India has been remarkable in the past decade. India now has more than a billion mobile phone users – in both urban and urban areas – and thereby Internet as well.

17 See 'News Feed Algorithm Change Improving Timelines of Posts', http://www.inside-facebook.com/2014/09/18/news-feed-algorithm-change-improving-timeliness-of-posts/ (accessed 15 Sept. 2014)

18 Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, London, 1991.

19 Lori Andrews, *I Know Who You Are, and I Know What You Did: Social Networks and the Death of Privacy*, New York, 2012.

20 See for example the works of Reinhardt Koselleck, *Futures Past: On the Semantics of Historical Time*, New York, 2004; Hartmut Rosa, *Social Acceleration: a New Theory of Modernity*, New York, 2013: Jonathan Crary, *24/7*, London, 2013.