Mass Digitization and the Garbage Dump: The Conflicting Needs of Quantitative and Qualitative Methods

Paul Gooding

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

There has been widespread excitement in recent years about the emergence of Large-Scale Digital Initiatives such as Google Book Search. While many have become excited at the prospect of a digital recreation of the Library of Alexandria, there has also been great controversy surrounding these projects. This paper looks at one of these controversies: the suggestion that mass digitization is creating a virtual rubbish dump of our cultural heritage. It discusses some of the quantitative methods being used to analyze the big data that has been created, and two major concerns that have arisen as a result. First, there is the concern that quantitative analysis has inadvertently fed a culture that favours information ahead of traditional research methods~~a culture that looks to reject traditional research methods, in favour of new and unproven technologies~~. Second, little information exists about how LSDIs are used for any research other than quantitative methods. ~~This has~~ These problems have helped to fuel the idea that digitization is destroying the print medium, when in many respects it still closely remediates the bibliographic codes of the Gutenberg era. The paper concludes that more work must be done to understand what impact mass digitization has had on all researchers in the humanities, rather than just the ~~vocal~~ early adopters, and briefly mentions the work that the author is undertaking in this area.

Introduction

Large-scale digital initiatives (LSDIs) such as Google Book Search (GBS) have received a huge amount of attention in recent years. Their unprecedented scale has led to them being labelled 'Million Book Projects' (Crane 2006), and they make large quantities of digital and print content available for full text searching and reading online. This has created widespread excitement at the prospect of a digital recreation of the Library of Alexandria (Battles 2004, p. 214) where the world's books are available online for public access. Yet there has also been huge controversy surrounding these projects. This paper will look at one of these controversies; the suggestion that mass digitization is creating a digital version of the Library of Babel (Borges 1964), a virtual rubbish dump of our cultural heritage. It will look at the quantitative methods that are being utilized to analyse the big data that has been created: Also known as 'distant reading' (Moretti 2007), quantitative analysis has massive potential as a research tool. But there are two facets of the technique that could exacerbate the problem of low quality digitization. First~~ly~~, there is concern that quantitative analysis has inadvertently fed a culture that ~~looks to reject traditional research methods, in favour of new and unproven technologies~~favours abstract information over traditional research methods. Second~~ly~~, little evidence exists about how LSDIs are being used for any research other than quantitative methods. This has helped to fuel the idea that digitization is destroying the print medium, when in fact it still closely remediates the bibliographic codes of the Gutenberg era. There is a risk therefore that focusing on the ability of early adopters to use large corpora with a low quality threshold could encourage a policy of digitizing quickly at the expense of quality.

Quantity versus Quality

> This whole dump is full of twinkling stars, reflections and fragments of culture.
>
> (Kabakov 2006, p. 36)

It appears impossible to reconcile the conflicting demands of quality and quantity using existing digitization technologies. Small digital collections have, in general, ~~been heavily curated~~ relied heavily upon human intervention to ensure quality, and have therefore used time-intensive methods to maintain standards. Frequently, these projects also utilize intellectually intensive methods to ensure that digitized content is presented in its most suitable form: these include harnessing the expertise of the academic community (British Library 2010), the work done to represent texts in digital form that has been undertaken by the Text Encoding Initiative (TEI) (Text Encoding Initiative),and modelling complex research processes in the humanities (Terras 2005, Crane et al 2006). The time-consuming nature of these methods renders them unsuitable for cost-effective digitization on a large scale (Holley 2009), meaning that large-scale digitization must ~~LSDIs, on the other hand,~~ rely upon scalable technologies such as page scanning and Optical Character Recognition to produce searchable text and retrieve metadata (Coyle 2009). In the case of GBS, this led to a number of quality concerns that were expressed after its launch: problems with incorrect or incomplete metadata (Nunberg 2009; Coyle 2009), particularly relating to name authority (Jackson 2008, p.167); the poor quality of some page scans (Jones 2010, p.55; Duguid 2007); unreliable OCR that produces a high proportion of errors in the machine-readable text (R. James 2010); and the proliferation of editions that results from scanning multiple copies of an individual work (Duguid 2007). These errors become magnified in such a large collection, creating noise that must be filtered out by the user. Kabakov's glittering gems of culture are buried somewhere, but the bigger a collection grows the harder it becomes for users to find them.

This drive, in some quarters, to digitize first and worry about quality later comes partly from a particular attitude to the corpora being produced. Google, for instance, has openly stated that the primary purpose of GBS is to create a searchable database of book, a 'giant electronic card catalogue' (Schmidt 2005) rather than a readable archive. Bates characterized this as treating text 'like a kind of soup that "content providers" scoop out of pots and dump wholesale into information systems' (Bates 2002). He understands that when content is treated as digital information a flattening of the structures of knowledge occurs which means that 'thirteen hundred words of gibberish and the Declaration of Independence are digitally equivalent' (Brown & Duguid 2002, p. xiii). While Lanier explicitly compares this attitude to sifting through a rubbish dump (2011, p. 131), it is unfair to suggest that those using quantitative analysis are engaged in exactly this. Instead, it is evident that a digitization strategy that prioritises information over users risks damaging the utility and usability of massive digital resources. ~~a belief in sections of the academic community that quantitative tools render traditional methods obsolete is potentially dangerous if given precedence.~~

Examples of Quantitative Ccultural Aanalysis

Quantitative analysis allows researchers to utilise corpora that exhibit poor quality at a reading level: 'as the size of a collection grows, you can begin to extract information and knowledge from it in ways that are impossible with small collections, even if the quality of individual documents in that giant corpus is relatively poor' (Cohen 2006). This partly explains why we have already seen such varied use of corpora. Franco Moretti, for instance, uses quantitative analysis to study the wider knowledge networks that surround literary texts. He is concerned primarily with discovering models that make sense of literary history: graphs of the growth of the novel in various countries; maps that show the nature of space in narrative; and trees that demonstrate the taxonomy of novelistic

genres (Moretti 2007). Matthew Jockers has used quantitative analysis to detect differences in writing styles between 19th Century Irish and English novelists (Jockers 2011). Other projects use quantitative analysis to automatically classify documents, a technique known as 'document classification' in Computer Science (Cohen 2006). Cohen ~~has~~ created a tool ~~that allows~~for users to locate university syllabi on any topic through a Web search[1] (Cohen 2011). Similarly, Docuscope has created a tool that identifies the genre of literary texts (Allison et al. 2011).

Due to its close links to Google, the Culturomics project has received a great amount of press coverage. This method of cultural research uses the Google Books word corpus to look for usage patterns of specific words or sentences (Michel & Shen 2010). Its novelty lies in its technological capabilities: 'of course there is a jump in scale, not just in the size of the corpus but also in the staggering processing power that the researchers can throw at it' (Nunberg 2010). The hyperbole that surrounded the paper's release (Bohannon 2011; Ruppert 2011) disguised some issues that other researchers in the field identified: difficulties with poor OCR and metadata; problems with legitimate but irrelevant data (Jockers 2010); changes in typography, such as the demise of the long-S, rendering results innacurate (Sullivan 2010); the problem of decontextualized data being stripped of meaning, and the difficulty this creates in drawing reliable conclusions (Nunberg 2010). Researchers in the field, though, are clearly aware of these limitations (Culturomics 2010; Moretti 2007, p.9), and it is likely that such flaws will be improved upon in time. Instead, the problem lies with a cultural movement that appears to have uncritically adopted quantitative analysis, and the risk that a glut of mediocre copycat work could begin to emerge.

[1] The Syllabus Finder was available until 2009, when changes in Google's API service caused it to stop working. The corpus that was created is available for download from Dan Cohen's website (Cohen, 2011)

Culturomics operates at a distance from the humanities, identifying itself more closely with the scientific processes of the human genome mapping project: 'these approaches tend to rely on (i) hundreds or thousands of people in massive, multi-institutional and multi-national consortia, (ii) novel technologies enabling the assembly of vast datasets containing a very specific type of data, and (iii) the deployment of sophisticated computational and quantitative methods in order to interpret the resulting data' (Culturomics 2010). As a result, a number of commentators have used this to justify the marginalization of close reading in the humanities. As Lanier points out, they 'care about the abstraction of the network more than the real people who are networked, even though the network by itself is meaningless' (2011, p.17).

Data vs Knowledge

> Since you had no past, you're going in for a magnificent, compensatory future
>
> (H. James 1987, p. 66).

In 1996, Geoffrey Nunberg noted that the rhetoric of technological determinism suggests that any successful innovation must inevitably destroy its predecessors (Nunberg 1996). The reaction to mass digitization certainly suggests that little has changed. Chris Anderson, for instance, claims that big data will effectively mean the end of theoretical research: 'the new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance without coherent models, unified theories, or really any mechanistic explanation at all' (Anderson 2008). The dismissal of the old is total: 'there's no reason to cling to our old ways. It's time to ask: what can science learn from Google' (Anderson 2008). Such powerful faith in quantitative data is almost religious in its fervour, but such hyperbolic rhetoric does the community a disservice,

because it suggests that ‘there will be no space left empty, no gaps or breaches to worry about’ (Duguid 1996). The ~~scientific~~ genomic methodology with which Culturomics aligns itself cannot be transferred in its entirety to humanities research, because it cannot operate without a qualitative framework. Luckily, academics in the field (Cohen 2010; Michel & Shen 2010) acknowledge that there are limits to what big data can tell us: ‘Quantitative data can tell us when Britain produced one new novel per month, or week, or day, or hour for that matter, but where the significant turning point lies along the continuum – and why – is something that must be decided on a different basis’ (Moretti 2007, p.9). In other words, the humanities will always have to rely on human analysis to some degree. Lanier claims that the alternative, the act of decontextualizing a text through user interfaces that obscure content and authorship, will result in only one book (Lanier 2011, p.46). This does~~n’~~ not go far enough. When texts are deconstructed to the extreme of granularity, and interactions become mediated by automated tools, there is no book. Instead, there is merely a massive corpus of words which carry no great epistemological significance. The thematic and structural elements of text are stripped away, to be replaced by a modern construction of the meaning of information.

The Changing Meaning of Information

Information was traditionally defined in a particularistic sense, imparting facts about specific events. In separating information from context, it becomes instead an abstract entity with human characteristics and a distinct personality: ‘people treat information as a self-contained substance. It is something that people can pick up, possess, pass around, put in a database, lose, find, write down, accumulate, count, compare, and so forth’ (Brown & Duguid 2002, p.120). These words foreshadow a contemporary climate where books exist as carriers of word-level information, demanding new methods of interaction. One of these demands is that all information must be equally valuable, and therefore must

merely be freely available; the 'egalitarianism of information dispersal', as Schmidt (2005) describes. There are obvious similarities to the idea of the inter-text: a work that exists as part of an intertextual network, its meaning mediated through links, citations, influences drawn from other texts, and the knowledge of each reader (Barthes 1977). But where the inter-text significantly undermines the author's influence by reducing them to a cipher for wider cultural ideas, the abstraction of information moves the meaning away from the text as a whole. Meaning instead resides in the words, which then become analogous to computer data. As data, information does~~n'~~ not necessarily carry epistemological significance because there is no need for it to directly provide knowledge (Graham 1999), or to prove its authenticity. The quantification of massive corpora then makes sense precisely because it appears to exist in isolation from the texts from which it is drawn.

Yet there is a contradiction at the heart of the method, because it ~~is~~ relies entirely on existing information. Analysing existing knowledge in this manner can, and does, produce new knowledge, but it can also act in a merely confirmatory capacity: 'striking as these results were, did we think they had produced new knowledge? The answer, of course, was no: Docuscope had corroborated what literary scholars already knew' (Allison et al. 2011, p.8). The novelty of a methodology can, as Duguid (1996) noted, hide the true impact of its findings. The power of quantitative methods will become evident when they are able to fully leave behind derivative forms of expression and assert their own original knowledge paradigm.

The Identity of Digital Technologies

This is problematic in itself, because it is clear that mass digitization is yet to exert its own unique identity. There is a tension between the reality of digital media and the

methods of interaction that they encourage. Schmidt's 'electronic card catalogue' (Schmidt, 2005) embodies this tension; while Google Books provides an unprecedented searchable database of textual information, the sources that it links to are indebted to Gutenberg. As Bolter and Grusin noted (1996, pp.356-357), new technologies commonly reference previous media in their developmental stage. Such close remediation is driven by an assumption that older media remain relevant, and so digitization still closely remediates the bibliographic codes of print that are so familiar to users. This is particularly evident in the marketing of eBooks to the public: 'reading on an iPad is just like reading a book. You hold iPad like a book and flip the pages like a book. And you do it all with your hands – just like a book' (Apple 2011). Large-scale digitization has the power to assert its own cultural paradigm, but its continuing reliance on the Gutenberg era demonstrates this is not yet the case. While mass digitization has provided information on a scale that unlocks new research methods, the limits of technology and copyright law force readers to locate texts in print or download digital copies that rely on the digital codes of print for structure and meaning. Digital media, for the majority, therefore still operates '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).

Formatted: Indent: Before: 0 cm

Conclusion

Where we're going we'll still need readers (Nunberg 2010).

We have seen above that mass digitization ~~quantitative analysis~~ opens up ~~powerful new research possibilities~~the possibility of increasingly large scale data-driven research methods, and this paper does not intend to dismiss ~~its~~ their validity. Rather, humanities research must continue to value the close reading that allows us to understand the outputs of quantitative analysis. More work must therefore be done on the users of LSDIs, in

order to establish how the research community is incorporating this new technology into their own work, and how future digitization projects can facilitate these behaviours. Little is known about how, or why, humanities researchers are using digitized texts as part of their research. The existing literature draws us towards Lanier's evocative image of quantitative researchers mining the past 'like salvagers picking over a garbage dump' (2011, p.131). This alarmist image gains power because the immaturity of mass digitization as a cultural paradigm continues to ensure a lack of evidence surrounding its true impact. My own research is studying how LSDIs are being used in the real world by combining quantitative and qualitative techniques in a case study approach. These methods will include web analytics and web log analysis, alongside a qualitative program of interviews, survey work and user observation, in order to discover more about the ways in which large-scale resources are being used by researchers. Specifically, these techniques will look for answers to a number of important research questions: what impact are LSDIs having on researchers; who is using LSDIs in their work, and how are they being used; what benefits and drawbacks do the large-scale digitisation of text resources create; is there any noticeable difference in outcomes between commercial and publicly funded resources; and how can we use this knowledge to ensure that mass digitization can develop to benefit the widest community possible? ~~This is~~These questions are necessary in order to teach us more about the impact of digitization, and to move the debate beyond a polarized argument between noisy early-adopters and concerned adherents of print technology. Kabakov (2006) talked of a rubbish dump where the fragments of our culture were hidden below the surface waiting to be found. More work must be done to ensure these fragments remain accessible, and that mass digitization continues to develop as a tool for all of the humanities.

References

Allison, S. et al. (2011). *Quantitative Formalism: An Experiment*, http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf ([Aa~~A~~ccessed 7February ~~7,~~ 2011)~~]~~.

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).

Apple (2011). *iBooks: a Novel Way to Buy and Read Books*, ~~*Apple.*~~ http://www.apple.com/ipad/built-in-apps/ibooks.html (accessed 8 December 2011).

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

Bates, M. (2002). After the Dot-Bomb: Getting Web Information Retrieval Right This Time. *First Monday*, **7**(7). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/971/ (a~~A~~ccessed 23 March, 2011).

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

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.

The Greek Manuscripts Digitisation Project (2010). *British Library Digitised Manuscripts*, http://www.bl.uk/manuscripts/About.aspx (accessed 6 May 2011).

Brown, J.S. & Duguid, P. (2002). *The Social Life of Information*. Boston, Massachusetts: Harvard Business School Press.

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).

Cohen, D. (2010). Initial Thoughts on the Google Books Ngram Viewer and Datasets. *Dan Cohen's Digital Humanities Blog*, http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-ngram-viewer-and-datasets/ (accessed 7 November 2011).

Cohen, D. (2011). A Million Syllabi. *Dan Cohen's Digital Humanities Blog*, http://www.dancohen.org/2011/03/30/a-million-syllabi/ (accessed 30 March 2012).

Formatted: Font: Italic

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. et al. (2005). Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. *Lecture Notes in Computer Science*, **4172**: 353-366.

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).

Duguid, P. (2007). Inheritence 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).

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.

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

Graham, G. (1999). *The Internet: a Philosophical Enquiry*. London: Routledge.

Holley, R. (2009). How Good Can it Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, http://hdl.handle.net/10760/12908 (accessed 30 March 2012).

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

James, H. (1987). *The American Scene*. London: Granville.

James, R. (2010). An Assessment of the Legibility of Google Books. *Journal of Access Services*, **7**(4): 223 - 228.

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/ab-115.xml;query=;brand=default (accessed 14 November 2011).

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

Kabakov, I. (2006). The Man Who Never Threw Anything Away. In *The Archive*. London: Whitechapel, pp. 32-37.

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

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

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

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

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

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).

Ruppert, E. (2011). Culturomics: a New Digital Method? *Centre for Research on Socio-Cultural Change*, http://www.cresc.ac.uk/news/blog/culturomics-a-new-digital-method (accessed 15 September 2011).

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).

Terras, M. (2005). Reading the Readers: Modelling Complex Humanities Processes to Build Cognitive Systems. *Literary and Linguistic Computing*, **20**(1). pp. 41-59.

Text Encoding Initiative (n.d). TEI: Text Encoding Initiative. *Text Encoding Initiative*, http://www.tei-c.org/index.xml (accessed 27 July 2012).