Big digitisation: where next?

Andrew Green
Llyfrgell Genedlaethol Cymru / National Library of Wales

Paper delivered at the Digital Resources for the Humanities and Arts conference, Belfast, 8 September 2009

Introduction

Digitisation is an ugly word for a beautiful concept.

As more and more recorded knowledge is created, transmitted and stored electronically, the digital has become the kingdom in which many people live much of their intellectual and even social lives. Already existing knowledge in analogue form is retreating like a sea on an ebbing tide: it fades from consciousness and falls into disuse. Digitisation is one technique for rescuing that knowledge and giving it new life.

My theme is what I have termed 'big digitisation': the attempt to translate large quantities of analogue knowledge into digital form. I shall concentrate on knowledge originally in print form and I want to consider three questions: why this mass approach has emerged, what it has achieved and might achieve, and how different models for big digitisation compare.

Digitisation as a cottage industry

Until 2005 most digitisation was small-scale. It has been described in retrospect, a little dismissively, as 'cottage' or 'boutique' digitisation. But it would be wrong to be dismissive. True, some early examples could be described as 'trophy sites', little more than promotional extensions of their owner's websites, while others were simple miniature reproduction galleries of little interest.

But small scale, done well, has great virtues. Medieval manuscripts can be scanned at the highest of resolutions, enabling detail to be seen that was barely or not at all visible in the original. Digitised texts or images can be surrounded with complex skeins of interpretation: translation, commentary, essays and more, connected in ways impossible in a traditional printed edition. Original works or collections like the Dunhuang artefacts from eastern China, scattered by history to 20 different locations around the globe, can be reunited virtually. Scanned graphics, texts and moving images can be linked together to create powerful educational tools.

It seems to me that this kind of digitisation has an assured future – possibly not as end in itself but simply as a natural by-product of another process, such as an academic research project. But its serene progress has been disrupted by the advent of a large monster, big digitisation.

The importance of Google Books

In December 2004 Google announced that it intended to digitise not mere collections but whole libraries. It had signed contracts with five major research libraries, all but one of them in the US: Harvard University Library, New York Public Library, University of Michigan, Stanford University and the Bodleian Library, Oxford. The weavers' attics of cottage digitising were to be replaced by massive factories occupied by a new page-turning proletariat to create a huge ocean of words, each one retrievable in a second through Google's miraculous search machinery.

What was Google's aim? According to the company, it was to '…make it easier for people to find relevant books – specifically, books they wouldn't find any other way such as those that are out of print – while carefully respecting authors' and publishers' copyrights. Our ultimate goal is to work with publishers and libraries to create a comprehensive, searchable, virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers.' As we shall see, this statement needs careful scrutiny. Its emphasis on the catalogue – harmless metadata – rather than full text, and its implication of altruism and philanthropy are, at least in retrospect, doubtful.

Since 2004 Google's giant mills have been at work incessantly. The original seven libraries have been joined by thirteen others, many of them now outside the Anglophone world: Belgium, France, Germany, Japan, Spain and Switzerland are all now part of the Google family. This expansion into works in other languages was in response to criticism, notably by the then President of the Bibliothèque national de France, Jean Noël Jeanneney, that Google nursed the imperial ambition of building a digital library dominated by English language and culture.[1]

Whether Google has plans to extend the number of major libraries, and therefore the proportion of world literature included in its programme, is unclear. There are of course many missing elements, some large and obvious, like Russian, Chinese and other Asian literatures, others less so, like the absence of Celtic and Australasian literature. Google's ambition to extend beyond the original restriction to books, though, is clear. In December 2008 the company announced that it intended to digitise magazines and include their contents in Google Books[2], and in September that year it announced modestly that it would digitise all the world's newspapers: 'Around the globe, we estimate that there are billions of news pages containing every story ever written. And it's our goal to help readers find all of them, from the smallest local weekly paper up to the largest national daily.'[3]

(The library project was just part of a wider ambition to give digital access to the contents of a wider printed universe: Google planned to work with publishers as well as libraries – and not only on out-of-copyright titles. Its 'Partner Program' works with publishers to make available parts of in-print books.)

---

[1] Jean Noël Jeanneney and Teresa Lavender Fagan, *Google and the myth of universal knowledge: a view from Europe*, Chicago: University of Chicago Pres, 2007. See the commentary by David Bearman, 'Jean Noël Jeanneney's critique of Google: private sector book digitization and digital library policy', *D-Lib Magazine*, vol.12, no.12, December 2006: http://www.dlib.org/dlib/december06/bearman/12bearman.html#14

[2] http://googleblog.blogspot.com/2008/12/search-and-find-magazines-on-google.html

[3] http://googleblog.blogspot.com/2008/09/bringing-history-online-one-newspaper.html

But it was the Library Project that roused both publishers and authors to take up arms against it by bringing a class action to the US courts in 2005. The belief of the Authors Guild and the Association of American Publishers was that Google was clearly infringing US law by scanning books (and therefore creating digital copies of them) without the prior permission of the copyright owners. Google denied the claim, explaining that its publication of 'snippets' from the works was justified under the principle of 'fair use'.

The backlash to Google Books

In October 2008, however, Google negotiated an out-of-court settlement with the authors, relevant to works within US copyright published before 5 January 2009.[4] This proposed to set up a Book Rights Registry, established through a contribution of $34.5m by Google, and in future by a contribution by Google of 63% of its profits from its Books service. The Registry would pay rights owners a portion of Google's profits arising from its scanning of their works and making portions of it available in full text, and from selling digital copies of the works. Those submitting a claim for retrospective digitisation by Google by 5 January 2010 would receive $60 per book; those wishing to object to the settlement or opt out of it should register their intention by 4 September 2009. As for income, as Google explained, 'for out-of-print Books and, if permitted by Rightsholders of in-print Books, Google will be able to sell access to individual Books and institutional subscriptions to the database, place advertisements on any page dedicated to a Book, and make other commercial uses of Books.'[5] There is a distinction between out-of-print in-copyright books, which Google will scan unless the rights holder objects, and in-print books, which Google will scan only with the express permission of the rights holder.[6]

It is not certain that this settlement will be accepted. A formidable army of enemies is arrayed against it, including Microsoft, Yahoo and Amazon. The US Department of Justice is investigating the deal in case it is anti-competitive. In Europe the French and German governments have come out firmly against the deal, and the European Commission is considering its response, proposing a harmonisation of the copyright laws of member countries to facilitate mass digitisation of European material so as not to be 'left behind' by the United States.

A federal court verdict on its legality is due in a Fairness Hearing on 7 October 2009. A Google victory will both legitimise the digitisation the company has already completed and allow it to continue digitising pre-2009 books in future, except where rights holders have opted out. It will have effects not just in the US but also abroad, wherever a 'US copyright interest' is involved (the effects of this are hotly debated).

There are already signs that Google intends to intensify its efforts outside the United States if it succeeds in the US court. There are rumours about possible deals with the Bibliothèque nationale de France, despite that library's previous hostility to Google,

[4] The best summary of the terms of the settlement is Jonathan Band, 'A guide for the perplexed: libraries and the Google Library Project settlement': http://wo.ala.org/gbs/wp-content/uploads/2008/12/a-guide-for-the-perplexed.pdf

[5] http://www.googlebooksettlement.com/intl/en/Final-Summary-Notice-of-Class-Action-Settlement.pdf

[6] For a critique of the settlement see Robert Darnton, 'Google and the future of books', *New York Review of Books*, vol. 56, no.2, 12 February 2009: http://www.nybooks.com/articles/22281

and the Italian Ministry of Culture has announced that it intends to signed a 'pre-agreement' with Google to digitise works from as many as 47 libraries throughout Italy. The European Commissioner for the Information Society and Media, Viviane Reding, has very recently conceded that collaboration with Google may be the only way for European libraries to make available substantial quantities of digitised books.[7]

Google Books: an assessment

No one could deny that within less than five years Google has transformed online access to older books and set the benchmark against which any other mass print digitisation project will be judged. In October 2008 the company announced that 7m titles were available on its Books website: perhaps only 1.5m of these will offer full texts, but it has been estimated that the company has already digitised 30m books and that at least 50 million full text books will result from the digitising the collections of the current pool of participating libraries. This will be a formidable corpus of knowledge, available for all manner of purposes.

But doubts and suspicions remain about the Google project and the settlement, even among some of the participating libraries (since the settlement Harvard University Library has withdrawn its own participation in in-copyright scanning).

First, 'fully participating' libraries (but no others) are provided by Google with a free digital file of the books digitised by Google from their own collections. They are empowered to take copies for preservation purposes. But use restrictions are severe and the copyright in the digitised texts may remain with Google in some or all cases: it is unclear from the published agreements between Google and participating libraries whether a library, if it wished to make its collection available online free, against Google's wishes, would be able to do so.

Second, through the terms of the settlement Google is moving away from a free-only model for access to a mixed economy: free for out of copyright material but pay per item for the full text of other books (here Google would compete with Amazon). The company will also charge libraries or other institutions for access to the complete online collection of all titles on behalf of their members. In effect this places Google in the same league as Elsevier and other monopoly sellers of digital content: the price of access could be high when no alternative source of supply exists.[8] It is also possible that Google could gain effective control of so-called 'orphan works' (works in copyright but whose rights holders can no longer be traced), being protected from post-digitisation copyright lawsuits; these may make up between 5% and 10% of all library holdings.[9]

---

[7] http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=5181

[8] The draft Settlement does make provision for one public access terminal in each US public library, which would offer free access to the collection.

[9] *In from the cold: an assessment of the scope of 'orphan works' and its impact on the delivery of services to the public*, JISC, April 2009: http://www.jisc.ac.uk/publications/documents/infromthecold.aspx#downloads
The UK government intends to include in a forthcoming 'Digital Economy Bill' a measure on orphan works: 'To pave the way for a more effective framework to deal with orphan works, the Government proposes to introduce legislation to enable commercial schemes for dealing with orphan works to be set

Third, the settlement, along with its commitment to expanding the size of Google Books, leaves Google as by far the dominant digitiser, at least in the area of books. A private monopoly is a dangerous situation in most circumstances, and libraries may feel more nervous than ever about entrusting their contents to Google, when the company's long-term intentions cannot be relied on. Some also have fears about privacy of knowledge about the reading of users of the Google service.

And finally, despite its size and current predominance Google is neither a public nor a permanent institution. It does not necessarily adhere to the standards such an institution would insist on: the quality of Google's scanning, and therefore of its quality control, is often deficient[10], and damage has been reported to original books during the scanning process in the University of Michigan[11]. More seriously, there can be no certainty that, in the same way as Google swept aside other search engine-powered companies, it in its turn will not be superseded by others. It can be no accident that of the national libraries of the world, charged with collecting, preserving and giving public access to their countries' publications in perpetuity, only one so far has decided to throw in its lot with the Google Libraries project.

Alternatives to Google

If Google is by far the largest but not the whole answer to big digitisation, what are the other solutions, and which economic models are available?

Google's project began as a partnership between the company and academic institutions. Likewise the Open Content Alliance, launched in 2004 in part as a response to Google Books, is a consortium of nonprofit organizations and Yahoo, and is administered by the not-for-profit Internet Archive. But it works in co-operation with copyright owners, using the 'opt-in' rather than 'opt-out' approach to rights management, and mounts digitised texts, mostly from North America, on the web for all to see without payment, often using Creative Commons licences.

For a time Microsoft was a contributing member of the Open Content Alliance, and financed digitisation projects, most notably by digitising about 100,000 out-of-copyright books in the British Library[12]. But in 2008 it announced its withdrawal from the field. This means that the OCA now mainly acts as a repository or aggregator of digital content donated by its contributors, but does not itself finance digitisation.

Another example of the public alliance is Carnegie Mellon University's 'Universal Digital Library', the initial aim of which was to digitize one million books in cooperation with partners in Egypt, India and China.[13]

---

up on a regulated basis':
http://www.culture.gov.uk/images/publications/DB_ImplementationPlanv6_Aug09.pdf
[10] Robert B. Townsend, 'Google Books: is it good for history?', *Perspectives*, September 2007:
http://www.historians.org/Perspectives/issues/2007/0709/0709vie1.cfm
[11] http://www.bl.uk/aboutus/stratpolprog/ccare/pubs/2007/libervol17thinktank.pdf)
[12] http://www.bl.uk/news/2005/pressrelease20051104.html
[13] http://www.ulib.org/index.html

Some national archives and libraries decided on a rather different strategy, entering into partnerships with large companies on a purely commercial basis and placing less emphasis on free access or even on online access at all. The best-known example is The National Archives in England and Wales, which employed commercial firms to undertake the digitisation of historic census returns, the expense being recouped through charging users for downloading extracts from the returns. The British Library is currently planning to contract with a company that will undertake, at its own risk, the digitisation of a large number of out-of-copyright newspapers. In return the company will enjoy the right to exploit the digitised collection by selling access to it worldwide for a specific period. The only free public access available during that period may be within the two physical British Library locations.

Those who advocate this approach point to its potential to make rapid progress, and to provide sustainability to a continuing digitisation programme – assuming, of course, that the commercial model itself holds water. The major disadvantage is the loss, at least for a considerable period, of either the wide public access that lies at the heart of the mission and indeed justification of national public institutions, or the free use of information that is also one of their central characteristics – or both. Libraries then become indistinguishable in their functions from publishers.

An alternative is to view both the creation and provision of digitised knowledge as a public good. Public libraries were established and still flourish as a means of ensuring that all members of society, irrespective of their circumstances, could have access to all published knowledge. Traditionally their method of achieving this goal was to offer a place where any citizen could read and usually borrow any printed publication. The 21st century equivalent is surely to offer citizens free access, where possible online, to publications in digital form.

It follows that, in the case of those publications already in print, libraries – in particular, I would suggest, national libraries, that have a duty to preserve and give access to their countries' published output – should do their best to arrange for their online public accessibility.

During the last few years several national libraries, especially in smaller countries of little interest to global giants such as Google, have developed plans to digitise their national collections. They include the national libraries of France, New Zealand, Slovakia, Australia, Norway, Finland and the Netherlands. The National Library of Wales has an ambition to give free online public access to as much as possible of the printed heritage of Wales, from the sixteenth century to the present day. This is such a large ambition that it will inevitable take many years and many millions of pounds to realise. However, a start has been made, with a project funded by JISC, the Joint Information Systems Committee of the UK higher education funding bodies, to digitise the contents of fifty of the leading periodicals of Wales, in both Welsh and English[14]. Two features of this project, now in its final stages, are worth noting: first, the titles all date from the twentieth century, and so involve numerous copyright complexities, and secondly a small part of the funding came direct from the Welsh Culture Ministry: a token of the importance of the project to our own government. The Welsh Assembly Government is paying the lion's share of a second, larger

[14] http://welshjournals.llgc.org.uk/

project, worth £3m and now under way, to digitise newspapers and periodicals from the nineteenth century, most of the money being derived from a central capital fund intended for programmes of strategic importance across all government departments[15].

Like Google each of these national initiatives must either avoid or tackle the question of in-copyright material: in the second case, by arriving at agreements with rights holders, either on a case-by-case basis (as happened with the National Library of Wales's 20th century periodicals project), or through a nationally negotiated agreement, as has happened recently in Norway[16]. Similarly, public digitisers have an interest in legislation (rather than a private Google deal) on orphan works that safeguards the public interest[17].

Even though a public programme of this kind is inevitably subject to uncertainties of funding and sustainability[18], its outcomes are surer and more substantial than those of a scheme funded through Google or other commercial sources. The quality of the digitisation process and product are under public control. The digitised texts are publicly owned: they are not 'given away' to private concerns, temporarily or permanently, ignoring the public investment in caring for their analogue originals. They have a secure home in a permanent public institution and are capable of continued preservation. And they are available freely to all citizens, in some cases for reuse and repackaging.

These are huge benefits, and should be fought for by all of us who care about public access to knowledge.

Exploiting digitised texts

I should like to end by considering the future benefits of having digitised a substantial proportion of a national published literature.

There are several gains from having scanned a large part of an entire universe, rather than a very small, themed selection. Google's motivation in adopting such an approach, expensive though it is, is now obvious from the terms of its proposed Settlement: a larger, heterogeneous collection can appeal to much wider audiences, both popular and specialist: a substantial interest that may be monetised through advertising, direct sales and other means. Comprehensiveness also carries inherent advantages for content analysis, for example by providing linguists and

---

15 http://www.llgc.org.uk/index.php?id=1514&no_cache=1&tx_ttnews[tt_news]=2277&tx_ttnews[backPid]=160&cHash=5379f44c5d

16 Marianne Takle, 'The Norwegian National Digital Library', *Ariadne*, issue 60, July 2009: http://www.ariadne.ac.uk/issue60/takle/; Vigdis Moe Skarstein, 'Strategies for a digital national library', paper delivered at the World Library and Information Congress, Milan, August 2009: http://www.ifla.org/files/hq/papers/ifla75/190-skarstein-en.pdf

17 British Library, *Orphan works and mass digitisation*, London: British Library, [n.d]: http://www.bl.uk/ip/pdf/orphanworks.pdf

18 For a recent comparison of sustainability models for digitisations see Nancy L. Maron, K. Kirby Smith and Matthew Loy, *Sustaining digital resources: an on-the-ground view of projects today*, London: JISC, 2009: http://sca.jiscinvolve.org/files/2009/07/sca_ithaka_sustainingdigitalresources_with_casestudies_sm.pdf

lexicographers with massive corpuses of raw material. In future, as techniques for searching, text mining, document recognition and automated translation become more sophisticated, as the semantic web develops and as 'cyberscholarship' becomes more common, these great reservoirs of text will yield new knowledge, and perhaps even generate new research fields.

As interest in this kind of specialist analysis grows it will ask complex questions about the digitised material itself: how, for example, might it be possible to extract consistent and structured information about places, people or dates out of a large mass of unstructured and heterogeneous data; how, in the absence of detailed structured metadata, might relevant items be recalled or grouped successfully; how could OCR techniques might be improved or supplemented to achieve greater accuracy[19].

Big digitisation collections can be made even more amenable to analysis if joined with cognate material from other sources in aggregation sites such as the Open Content Alliance, Carnegie Mellon's 'Universal Digital Library' or the European Union's 'Europeana'[20].

But perhaps the most intriguing uses of big digitisation arise from what users themselves do with the texts that they find. The National Library of Australia found that if it presented online the raw OCR'd versions of scanned historic newspapers there would be no shortage of volunteers willing to correct large quantities of inaccurately interpreted text – and compete against one another for the title of champion corrector.[21] (The term for this activity is 'distributed proofreading') It is not difficult to imagine how groups of users might respond, given the opportunities and the tools, to the presence of huge quantities of text in digital form: by annotating, translating, citing, discussing, analysing, reusing and repackaging.

Many of these developments we can only speculate about today. Without doubt the uses to which people in the future will put the fruits of big digitisation will be very different from today's uses. All the more reason why we should do our very best to plan today in a way that safeguards the interests of the researchers of tomorrow.

[19] See Simon Tanner, Trevor Muñoz and Pich Hemy Ros, 'Measuring mass text digitization quality and usefulness', *D-Lib Magazine*, vol.15, nos 7-8, July/August 2009: http://www.dlib.org/dlib/july09/munoz/07munoz.html

[20] http://www.europeana.eu/portal/

[21] Pam Gatenby, 'The Australian newspapers service and user interaction through text correction', paper presented at the World Library and Information Congress, Milan, August 2009: http://www.ifla.org/files/hq/papers/ifla75/99-gatenby-en.pdf