LIBER QUARTERLY, ISSN 1435-5205

Igitur. Utrecht Publishing & Archiving Services

The Google Mass Digitisation Project at Oxford

by RONALD MILNE

ABSTRACT

For most of the 400 years of the Bodleian Library's existence, users have had to travel to Oxford to use its collections. In recent years, Oxford has undertaken a number of focused, 'boutique' digitisation projects. Now, as a partner in the Google Library Project, an immense range of scholarly and other 19th century out-of-copyright library materials from the Bodleian's collections will be digitised *en masse* and will be made freely available over the internet to anyone who has Web access. Millions of books and journals will be scanned in the course of the project and the author contends that digitisation on such a scale represents a revolution in the dissemination of information that parallels the impact of the invention of printing from moveable type in the 15th century.

INTRODUCTION

Sir Thomas Bodley, in founding his Library in Oxford in 1602, intended it to be not just for the University of Oxford, but also for what he called the 'Republic of Letters'. By this he meant that the library's collections should be open to all who had need of them. This ethos is one that has held good for four centuries, and of which Oxford is proud. The Bodleian's doors are open to the intellectually curious, whether they have an academic affiliation or not. Indeed, it is notable that 60% of the Bodleian's registered readers are not members of the University of Oxford.

From the start, Bodley intended his library to be one of the great libraries of the world. He encouraged his 'very great store of honourable friends' to give money and books to support his project. The Bodleian Library acquired the legal deposit privilege in 1610, the first library in the UK to do so, and today it is the oldest extant legal deposit library in the world. With a stock of over 8 million items, built up by legal deposit, purchase and gift, Oxford is fortunate in possessing in the Bodleian what is the richest array of manuscript and print collections of any university library in the world.

Over the centuries very many scholars have travelled to Oxford to use the Bodleian's collections, while those unable to make the journey, or unable to stay for long, have used microfilms and transparencies as surrogates to aid them in their studies. Recently, the advent of internet and the ability to digitise large quantities of text and images, and make them available over the Web, has transformed ways of working.

A BRIEF OUTLINE HISTORY OF DIGITISATION AT OXFORD

Oxford has been involved in digitisation projects since 1993, when through the sponsorship of Toyota City, the Bodleian undertook a project focusing on motor car ephemera in the library's John Johnson Collection, supplemented by images of other forms of transport. The images were produced by means of Kodak's Photo-CD technology: the original items were photographed onto 35mm slides, and scanned onto Photo-CD discs. These were then processed to convert the images to compressed JPEG and GIF images at various resolutions.

The technology has, of course, moved on since then and the University has undertaken a host of other digitization activities in the interim. These include the Early Manuscripts at Oxford University project funded through the joint higher education funding councils' Specialised Research Collections in the Humanities initiative. [1] It created high resolution digital images, which were scanned directly from the original manuscripts. Over eighty manuscripts dating from the 9th to 19th centuries were selected as major treasures from their respective Oxford libraries, to create access to material where the original might otherwise be too fragile to handle. Other projects include the Broadside Ballads project (also funded under the Specialised Research Collections in the Humanities initiative), the creation of archive quality digital images of political and satirical prints and trades and professions prints from the John Johnson collection under the JISC Image Digitisation Initiative (JIDI) [2] and, more recently, the digitization of the Shikshapatri, funded by the New Opportunities Fund (now the Big Lottery Fund). The Shikshapatri, a fragile Sanskrit manuscript, written by Shree Swaminarayan, founder of Swaminarayan Hinduism, outlines moral and spiritual codes for everyday life.

Six years ago, in 2000, recognizing the opportunities for enhancing access through digitisation, the Andrew WMellon Foundation gave generous help in establishing the Oxford Digital Library (ODL). The Foundation funded the essential infrastructure of the ODL and supported a range of digital projects based on core research material from Oxford's library holdings. It aimed to promote scholarly effort with relevance to research and teaching by digitising, delivering and enhancing the University's major library holdings A grant programme for projects was managed by the Oxford Digital Library in collaboration with an editorial board, comprising scholars and librarians from within and outside Oxford. Examples of the great variety of material digitized under the grant programme include political cartoons from the period of the French Revolution and Napoleonic wars from the Bodleian's Curzon Collection, and Sibthorpe and Smith's *Flora Graeca* (1806 - 1840) illustrated by Ferdinand Bauer. That project undertook the digitisation not only of the printed volumes, but also the original hand-coloured drawings from which the printed engravings were made.

Consistent with the policy of making its collections as accessible as possible, Oxford has also been a major participant in commercially funded projects such as Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO).

THE GOOGLE LIBRARY PROJECT

The projects described above, and others like them undertaken in Oxford and elsewhere, have generally focused on a particular subject area or genre. This has been for good reasons - on the one hand funding has been strictly limited and, on the other, it has seemed to make good sense to concentrate, in an academic environment, on projects that are of immediate interest to scholars. A lot of excellent work has been done, and if one had an observation to make it would merely be to remark that most of the work has been done piecemeal, without much reference to digitisation work going on elsewhere. Without wishing to detract from the digitisation projects that have been undertaken so far, they may largely be regarded as 'boutique' activities when compared with the Google Library Project.

The Google project, in contrast, involves digitisation on an industrial scale. Books are digitised *en masse* rather than cherry-picked. They are 'selected' only in as far as they should be in a fit state, in conservation terms, to undergo the non-invasive Google digitisation process and they should be of a format with which the Google scanning technology can currently cope.

Who is taking part in the Project? The 'Google Five', as they have become known are: the libraries of Harvard, Stanford, Michigan and Oxford universities and the New York Public Library. It seems likely that other libraries will become participants in due course. Millions of printed works will be digitised as a result of the project, which falls under the umbrella of the Google Book Search programme.

At Oxford we plan to digitise as much of our 19th century out-of-copyright material as possible, amounting to between 1 and 1.5 million volumes. The project is not limited to the Bodleian Library and we expect to draw, for example, on the collections of the Taylor Institution Library (Oxford's modern languages research library) and the Sackler Library (art history, archaeology, and classics (ancient history and literature). Together, the scope of Oxford's 19th century material is immense and caters for a very wide range of interests We intend to digitise printed material in every subject area and will cover all works, including books and journals which would have had no particular academic interest when acquired under legal deposit. Thus, alongside academic books and journals, we will be digitising recreational magazines, trade literature, post office directories, railway timetables, and much more.

Isn't there a risk of duplicating what is being digitised in other libraries? The answer is 'yes', but the incidence of duplication is not as great as one might expect. By analysing data from WorldCat, Lavoie, Connaway and Dempsey have shown that, among the 'Google Five', 56% of works are held uniquely by one of the Google Five libraries and that, when one compares two libraries out of the five, eight out of ten books are held uniquely (Lavoie, Connaway & Dempsey, 2005). Other interesting data that emerge from the analysis, particularly in light of the assertions that the Google Library Project is an example of Anglo-Saxon cultural domination, is that about 50% of the holdings of the five libraries are in languages other than English, and that over 430 different languages are represented in the libraries' joint holdings [3]

Between the signing of the agreement with Google and the scanning operation starting, we spent much time considering logistics and refining workflows. From the start, Oxford appointed a project manager to undertake day-to-day liaison with Google, and Google had a project manager on site in Oxford over six months before scanning commenced. Scanning in Oxford is underway and the operation is nearing full production. As part of the pre-digitisation process, books are appraised to determine their suitability for scanning, with Oxford conservators having the absolute right to say that a book may not be scanned for conservation reasons. Books are transferred by van from the library stacks to the digitisation centre nearby. The Bodleian, a non-circulating library, has not bar-coded its books in the past but, as it is an essential part of the pre-scanning workflow, this and associated bibliographic work is being undertaken as part of the production line process.

Results of surveys of Bodleian holdings undertaken as part of the preparatory work for the project have shown that as much as 20% of 19th century holdings are wholly or partially uncut. This was an unwelcome finding if not a complete surprise, given the amount of material acquired under legal deposit. We also found that around 3% of the Bodleian's stock from that period is uncatalogued. We are addressing these two matters.

The Google digitisation process captures the whole book, in as far as possible. Thus, boards are digitised as well as the text block. Frontispieces, plates and other illustrations are captured, too, although fold-out illustrations and diagrams fall outside the process at present. All pages are digitised in colour.

The agreement with Google is non-exclusive. Google retains a 'Google Digital Copy' and Oxford receives an 'Oxford Digital Copy'. The digitised book will be navigable, the entire text will be searchable, and there will be a link from the Oxford catalogue to the digital copy. As the books being digitised at Oxford are in the public domain, complete copies of individual works can be made freely available over the internet to anyone who has Web access. The costs of the infrastructure and the planning required are such that Google has agreed to host access to the Oxford copy for the time being. There are potentially many different ways in which we could make the Oxford Digital Copy available in due course. Unsurprisingly, we have already been approached by a number of individuals and organisations suggesting that we accord high priority to their area of activity!

CONCLUSION

What effect will mass, or large-scale, digitisation projects have on the information landscape? Given the millions of books and journals that will be scanned in the course of the Google Library Project, digitisation on such a scale surely represents a step-change in the dissemination of information that parallels the impact of the invention of printing from moveable type in the 15th century. In enhancing access to printed material that is not otherwise easily available, the Project has the potential to act as a transforming agent, in learning, teaching and research as well as many other activities. Will it mean the decline of traditional libraries? Anecdotal evidence would suggest that, while dissemination of the printed works will be greatly facilitated, and many who would not otherwise be able to access particular material will be satisfied with a digitised copy, others who perhaps had not intended to visit a library will be inspired to see the original. Scholars will regard their viewing on the Web of material held in the Bodleian and other libraries as valuable preparation and allow them to plan their research in Oxford and elsewhere more efficiently.

Would Sir Thomas Bodley have approved of books in his library being digitised, and on such a large scale? Without doubt he would have regarded the Google Library Project in Oxford as a natural extension of his desire to make the Bodleian's collections readily available to external readers, regardless of affiliation. The Bodleian ethos of facilitating access to all - the concept of the 'Republic of Letters' - has found expression in the Digital Age and of that Sir Thomas would be proud.

NOTES

1. See: Libraries Review: Specialised research collections in the humanities
http://www.hefce.ac.uk/research/initiats/Srcith/

2. The JIDI project enabled digitisation of copyright-cleared resources in a number of archival collections.

3. ibid.

REFERENCES

Lavoie, Brian, Lynn Connaway and Lorcan Dempsey: "Anatomy of Aggregate Collections: The Example of Google Print for Libraries". *D-Lib Magazine,* 11(2005)9. www.dlib.org/dlib/september05/lavoie/09lavoie.html

WEB SITES REFERRED TO IN THE TEXT

Big Lottery Fund. http://www.nof.org.uk/

Bodleian Library. http://www.bodley.ox.ac.uk/

EEBO - Early English Books Online. http://eebo.chadwyck.com/home

ECCO - Eighteenth Century Collections Online. http://www.gale.com/EighteenthCentury/

JIDI - JISC Image Digitisation Initiative. http://www.ilrt.bris.ac.uk/jidi/

ODL - Oxford Digital Library. http://www.odl.ox.ac.uk/