# A BRIEF SURVEY OF ONLINE LIBRARIES

*by Chris Phillips*[1]

**ABSTRACT**

This article is an attempt to survey briefly some of the projects that are making digital versions of published works available through the Internet.

*Foundations (2008) 2 (6): 443-450* *© Copyright FMG and the author*

## 1. Introduction

In December 2004, the Internet search engine Google announced the launch of a major project to scan some 10 million books or more, to digitize their contents, to make the resulting text fully searchable and to provide free access to all the volumes that are in the public domain. In response to Google's announcement, other Internet corporations declared an interest in the digitization of books, and Microsoft in particular initiated a significant book scanning programme.

As a field of study, medieval history - and perhaps medieval genealogy in particular - is unusual in that so much of the primary source material was published in the 19th and early 20th centuries. Even with prominent families, it's often the case that crucial evidence has been in print for a century or more (or even, in some cases, for several centuries), but is awaiting rediscovery in a forgotten or unappreciated work. Clearly, if all this published material were freely available and searchable through the Internet, research into medieval genealogy would be transformed.

This article is an attempt to survey briefly what progress has been made towards this genealogical Utopia by Google and others over the last four years, and to gauge future prospects. It covers only collections that are freely available through the Internet (though there are a number of others that are accessible only by subscription, or through libraries or academic institutions).

Inevitably much of the detail below will very soon be out of date - and if it had been written only a month ago it would already have needed substantial revision - but I hope the essentials will remain accurate enough to make it useful for a year or two.

## 2. Online library projects

### 2.1 Project Gutenberg

Google's programme is the most ambitious project to digitize books, but was far from being the first. Project Gutenberg (www.gutenberg.org) was launched in 1971, long before the development of the Internet, with the aim of producing electronic versions of public-domain books by manual transcription, though now the donkey work is done by optical character recognition (OCR) software, and checked by volunteers. To date more than 25,000 electronic books have been produced, and they are all freely

---

[1] The author is the owner of the website “Some notes on medieval English genealogy” at http://medievalgenealogy.org.uk/
*Email contact:* cgp@medievalgenealogy.org.uk

available through the Project's website. Unlike some of Project Gutenberg's successors it does not offer images of the original pages, only plain text or HTML versions, but the text of all the volumes in the collection is searchable, through the advanced search form on the website.

## 2.2 Google Book Search

Google Book Search (books.google.com) draws text from two sources: (1) the "Partner Program", through which publishers and authors submit the text of books for inclusion in the index and (2) the "Library Project", through which Google is scanning and digitizing printed volumes, in partnership with a number of major academic libraries (initially five, in the USA and UK, but now nineteen, including several in Continental Europe and Japan).

The website offers an advanced search form, combining fields for bibliographical information (author, title, date of publication and so on), with options to search for words and phrases within the text of the books. Several types of results are presented. For books deemed by Google to be in public domain, a "Full view" is offered. For some other books submitted by authors and publishers a "Limited preview" may be available, allowing the user to read only part of the work - either sample chapters or a limited number of pages. For other books still in copyright, there may only be a "Snippet view", showing a line or two containing the search term. But for many other works, no preview at all is available, for reasons that aren't made clear. The search form also allows the user the option of showing results only for volumes for which a full view (or a limited preview) is available.

For public domain works, the "Full view" offers images of the scanned pages (viewed either as a vertical array of single pages or as pairs of pages presented in an imitation book format). There are normally also the options of downloading the page images for the whole book as a PDF file, and viewing plain text instead of page images. For navigation within the book, there is a hyperlinked table of contents and also a facility to search for keywords within the volume.

This is certainly the largest of the book digitization projects, though it's remarkably difficult to find hard figures for how many volumes have been scanned so far. Official pronouncements seem to be limited to "more than a million volumes". A search on Google's website for all books published up to 2008 returns nearly twice that many results, though it has been noted that the numbers of results displayed are often wildly inaccurate, and this total will include books submitted by publishers as well as those scanned in libraries. Such information as there is about the rate of scanning at the participating institutions suggests that many more than a million volumes must have been scanned by now. However, only a fraction of those are public domain works, available in their entirety.

This sounds like a researcher's dream, and it can't be denied that Google Book Search is already a very useful research resource. But there are problems with the service, which it's as well to be aware of.

One of these problems is rather unexpected, given Google's reputation as a world leader in Internet search technology. Even casual users will have noticed that the total number of results claimed for each search bears little relation to the number actually displayed. This is not just a cosmetic flaw, as closer investigation using the advanced search form reveals many other anomalies - such as general searches giving less results than more specific ones. So, unfortunately, it's not safe to assume that searches will find all the relevant material in the collection.

It must also be remembered that even if the search mechanism worked perfectly, it would be dependent on Google having produced a complete and accurate electronic version of the text of every book scanned.

In reality, the quality of the scans can be poor, with some pages missed, some incompletely scanned, and some too faint to be legible even to the human eye. And Google's electronic text is written not by human beings, but by OCR software, which is never perfect, and which can produce particularly bad results when applied to poor-quality scans and older material whose typeface may be unfamiliar to the computer program.

So it should never be assumed either that a search for a keyword through the whole collection will find all the relevant volumes, or that a search within a particular volume will find all the occurrences of the keyword.

Searching for particular volumes is also made difficult by the fact that the quality of the bibliographical information can be very poor - to the extent that one often has to display the title page of the book to be sure what one is looking at (and of course for many volumes in its collection Google makes it impossible to do this). This is a particular problem for volumes that are part of a series, for which the series can (usually) be identified, but the volume often can't. It's to be hoped that this defect, at least, will be remedied in the future, as Google has recently reached an agreement allowing it access to the WorldCat catalogue records of the Online Computer Library Center (see below).

An even greater handicap for most of the world is Google's policy on copyright. Essentially this seems to be that, for those in the USA, books published in the USA before 1923, and books published elsewhere before 1908, are deemed to be in the public domain. But for those outside the USA, Google applies much more stringent criteria. For example, it seems that for those from the UK only books published before 1865 are available in full. Although Google has claimed to have procedures for checking whether works are in the public domain outside the USA, and even that it is working hard to make this process "faster and more efficient", I have yet to see any evidence of this actually happening. It should be added that even for visitors in the USA, many books which should be in the public domain according to Google's criteria are, inexplicably, available only as "snippets", or not at all.

To some extent, the access restrictions imposed on those outside the USA can be circumvented by using a proxy server such as Proxify (proxify.com) to disguise the visitor's location. This allows access to the basic functions of the site, though some features, including the option to download whole volumes, are not available.

## 2.3 The Open Content Alliance

The Open Content Alliance (www.opencontentalliance.org) was announced in 2005 by the Internet Archive (IA) and Yahoo!, and has been seen as an open, collaborative alternative to the closed, commercial Google project. Other contributors include - or have included - Microsoft and a number of major research libraries (including the UK National Archives). The Alliance's own website appears to be moribund, but the Internet Archive has brought together a library of more than 400,000 items, in the "Texts" section of its website (www.archive.org/details/texts), many of them digitized under the auspices of the Alliance. The collection also includes material contributed by other projects, such as most of the texts produced by Project Gutenberg and some of those from the Universal Digital Library (see below).

The collection can currently be searched by using the site's advanced search form, and by choosing "texts" (not "text"!) as the "Mediatype". Unlike Google, the site offers no facility to search within the digitized text of the volumes, though this is planned for the future. Only bibliographical information, such as keywords for author, title and date of publication, can be specified, in order to search for particular volumes.

Each volume is normally made available for download as a plain text file and as page images in various formats, including PDF, DjVu (a proprietary format requiring the installation of free viewing software) and the "flipbook" (the Internet Archive's own imitation book format). But for some works, such as those provided by Project Gutenberg, no page images are available - only a transcript of the text.

To some extent the same concerns about the quality of scanning and OCR, discussed above in relation to Google, also apply to the collections available through the IA. In addition, at present there seem to be fairly frequent server problems, affecting both searches and downloads. But the great advantage of this resource over Google's is that all the volumes in it can be read - it consists mainly of works in the public domain, together with some copyright material included with the permission of the owners. So while Google's collection is far larger, so many of its books are available only as "snippets" or not at all, that the amount of material fully available to read through the IA appears to be comparable with that at Google Books. In fact, some trial searches indicate that the Internet Archive may still have the edge in this respect.

But sadly the project has recently suffered a body blow, as Microsoft, which had been funding the scanning of a thousand books a day, announced abruptly in May 2008 that it would be terminating its involvement. The scans will not be lost, as Microsoft has said it will donate them to the IA and remove contractual restrictions on their use. But Microsoft's Live Search Books site, through which its scans were available - and which allowed full text searches, which the IA is not yet in a position to offer - has already been closed down. The Alliance's other main corporate backer, Yahoo!, does not appear to have any active involvement at present. Some funding from other sources (such as the Library of Congress) remains available, and Microsoft has donated to the project the book-scanning equipment that it had purchased, but this will clearly mean a drastic reduction in the rate at which material can be added to the collection, unless a substantial source of new funding can be found.

## 2.4 The Universal Digital Library

The Universal Digital Library (www.ulib.org), originally known as the Million Book Collection, is a project run by staff of Carnegie Mellon University, in collaboration with the Chinese government and academic institutions in China, India and Egypt. As the older name suggests, the aim was to digitize a million books and make them available through the Internet. This aim has been accomplished, and the collection now claims more than 1.5 million volumes.

Unfortunately for those in the West, nearly a million of these are in Chinese, and about 150,000 are in other oriental languages. Only about 370,000 are in English. The other snag is that like Google's collection, this collection includes many volumes that are still in copyright and, unless the owner has granted permission, only a sample of these works - not more than 10% - can be viewed. Unlike Google's policy, the copyright criterion does have the virtues of simplicity and clarity: works are deemed to be in the public domain if they were published before 1923, and are otherwise assumed to be still in copyright. According to the statistics on the site, only

just over 100,000 public-domain works are included in the collection. It's not clear how many of these are in English, but it's worth noting that 30,000 of the volumes are also available through the IA website.

The search form and the presentation of the volumes are similar to those provided by the IA - only bibliographical information can be specified. There is no facility to search the text of the volumes, though it is intended to provide one in the future. The works can be viewed either as text, or as page images in either TIFF or DjVu format, both of which require the installation of free viewing software. The disadvantages are that the books can be viewed only page by page - there is no way of downloading a whole book - and that the viewing software doesn't work properly on the current version (7) of Internet Explorer, the most popular web browser. It is intended also to provide books in PDF format in the future, and to allow whole volumes to be downloaded.

## 2.5 Gallica

Another major digitization project, predating Google's, is Gallica, the "bibliothèque numérique" of the French Bibliothèque nationale, established just over a decade ago. The original site (gallica.bnf.fr), available only in French, offers images of scanned pages, with whole volumes downloadable as PDF files, searchable by specifying bibliographical information only - there is no electronic text, and therefore no capability of searching for keywords within the text. In 2006, there were 90,000 volumes online. This figure is still displayed on the main page, though many more volumes have been added in the last two years. The catalogue also includes some records relating to digitized books available elsewhere.

The new site, Gallica2 (gallica2.bnf.fr), still under development, offers versions of the main pages in English and Spanish as well as French, and in some respects represents a move towards services similar to those offered by Google. An important development is the addition of volumes whose text is searchable. The advanced search form therefore allows keywords and phrases from the text to be specified, as well as bibliographical information. The new site has also followed the lead of Google and Amazon by including, for a trial period of a year, several thousand recent books submitted by publishers. Currently, Gallica2 includes about 68,000 volumes (about 85% of them in French), of which 64,000 are freely available and 12,000 have searchable text. Evidently, each of the Gallica sites includes works that are not available on the other, so at present it is necessary to search both.

It's worth noting that although there's naturally a preponderance of French-language works in Gallica's collection, it also contains much that is relevant to medieval English genealogy, including many volumes of the Rolls Series. Unfortunately the quality of some of the older scans can be very poor, though.

## 2.6 Amazon

The online bookseller, Amazon (amazon.com), offers a search facility covering the text of recent books submitted by publishers, similar to that provided by Google Book Search. In the advanced search form, the user can specify text keywords (or exact phrases in quotation marks), together with bibliographical information if desired. (Note that this facility doesn't currently appear to be available at some of Amazon's other sites, such as amazon.co.uk.) There is also the option to "search inside" particular books whose text is available. This search facility was originally advertised in 2003 as covering 120,000 books, and presumably now covers many more. Tests suggest the number of volumes included is comparable with the number of recent works indexed by Google.

More recently, Amazon has also included some scanned books in its database - some of them out-of-print volumes submitted by copyright holders, and others public domain material scanned at libraries. These are offered for sale by Amazon's "print on demand" division, and are currently available only as paper copies, not in digital formats. These books appear to constitute only 1% or so of the total number in the database, so that Amazon's involvement in book scanning is negligible in comparison with that of the other projects.

### 2.7 Other projects

The projects outlined above are the largest in progress, each covering at least 100,000 volumes, and some many more.

Of course, there are many smaller projects, and some of them are highly relevant to medieval genealogy. They range from substantial online libraries, often based on particular themes, and maintained by academic institutions and libraries, to small collections or even individual volumes relating to a particular place or family, transcribed and put online by enthusiastic amateurs.

Indeed, it's often the case that smaller collections are digitized with more care than the huge, industrial-scale libraries of Google and its peers, and that the quality of the text and bibliographical information is higher as a result. Some of the small collections also include significant amounts of copyright material, with the permission of the owners. For example, the British History Online website (www.british-history.ac.uk), provided by the Institute of Historical Research, contains a digital library that includes many recent volumes of the authoritative Victoria County History series.

Obviously, the disadvantage of the smaller collections is that of fragmentation. There is no way of searching all of them for a particular volume, still less of searching the text of all the volumes they contain. This problem is discussed in the next section.

## 3. Searching the online libraries

Hundreds of thousands of public-domain books are now fully available through the online libraries, and the text of an even larger number of books still protected by copyright can be searched for keywords and phrases. Despite the problems discussed above, this amounts almost to an embarrassment of riches for the researcher, but the resources are fragmented between half a dozen large collections and innumerable smaller ones.

How can the researcher best find relevant material within these online libraries? Specifically:

(1) how can she find out whether a particular book is available, and

(2) how can he search for keywords within the text of the online books?

### 3.1 Searching for a particular book

The bad news is that at present there is no alternative to performing a separate search in each of the major collections described above.

There is a useful unofficial online catalogue, the Online Books Page (onlinebooks.library.upenn.edu), which covers versions - text and/or page images - of published books in English that are freely available through the Internet. This

provides a basic search facility allowing words from the title and author's name to be specified. Currently it lists more than 30,000 volumes. This does include some items from the large collections - indeed, many of the recent additions come from the Internet Archive - but it is useful mainly as a finding aid for material in smaller online libraries.

On a much more modest scale, I try to maintain links to material relevant to medieval English genealogy on my own website (www.medievalgenealogy.org.uk). Details of links not yet included are always welcomed.

Some help is also provided by WorldCat (www.worldcat.org), a union catalogue of (non-digital) libraries, maintained by the Online Computer Library Center, advertised as containing 100 million bibliographical records. The basic information in this catalogue is now freely available online. Although WorldCat was originally intended as a means of locating paper publications, it does also include some links to electronic editions, in the "Get It" section of the catalogue record, and there are plans to extend this feature in the future. In principle, these include links to some of Google's digitised books, though these are displayed only if the user will have access to the full volume, so they are of limited help to those outside the USA. Worldcat also provides links to books in some subscription-based electronic libraries, and some digital books - such as those on the British History Online website - are listed in their own right, as "Internet Resources".

Another resource that may become very useful in the future is the Open Library Project (openlibrary.org). This is a an offshoot of the Internet Archive which, essentially, aims to be to the library catalogue what Wikipedia is to the encyclopaedia. That is, it aims to have a web page for every book ever published, which can be edited by any user, containing bibliographical data and information about library holdings, and linking to digital versions where available. It is intended to make the electronic texts of all these books searchable. Not only that, but it is hoped eventually to make the texts themselves editable, so that users can proofread them and correct errors of character recognition.

Obviously these aims are extremely ambitious. At present the website is described as a beta version, in the process of development, but it already includes nearly 13.5 million catalogue records, for more than 240,000 of which the full text is available (mostly from Internet Archive scans or Project Gutenberg transcripts). It is claimed that this text is already searchable, but the keyword search never returns more than a single page of results, so clearly it is not working properly yet. That being the case, since the site contains far fewer catalogue records than WorldCat, and far fewer full texts than the Internet Archive, it is currently useful only as a taster of things to come. Clearly, this project has the potential to evolve into an extremely useful resource - even a reasonably comprehensive catalogue of links to online books elsewhere on the Internet would be very helpful - though there must always be concerns about the reliability of the contents of websites which can be freely edited by their users.

## 3.2 Searching for a keyword

When it comes to searching the text of books online, for most of the major collections described above there is no alternative to performing separate text searches where this is possible - that is, for Google, for Project Gutenberg, for the new books indexed by Amazon and for a small part of Gallica's collection. For the Universal Digital Library there is not yet any way of searching the text.

The odd one out is the Internet Archive library. This doesn't yet offer a search of its own texts (though it intends to do so in the future through the Open Library project, just discussed). But the full text of each work is visible on the website, so in theory keywords and phrases can be found by performing a site-specific search using one of the general-purpose Internet search engines (for example, by including the qualifier site:archive.org in a Google web search).

Unfortunately, this doesn't currently work in practice, because the search engines do not index the whole of the Internet Archive website. A test using phrases from the text of ten volumes in the collection showed that Google was able to find only two of them, Microsoft Live Search only one, and Ask.com and Yahoo! none at all.

## 4. Conclusions

Perhaps unsurprisingly, the essential conclusion nearly four years on is that Google's book digitization project is by far the largest, and is certain to remain so for the foreseeable future. That project deserves two cheers (perhaps only one cheer from those outside the USA) for providing a very useful research resource, though one with many flaws and weaknesses. The biggest disappointment is that, whereas it seemed at one stage that Google's project would be imitated by several other giant Internet corporations, the competitors have now largely withdrawn, leaving Google in sole possession of the field. Google recently insisted that it was still as committed as ever to its large-scale digitization programme, but of course Microsoft's abrupt departure is a reminder of how ruthless commercial organisations can be in terminating unprofitable operations. And even if Google does have a continuing commitment to scanning millions more volumes, there's evidently little commercial motivation to remedy the long-standing problems with the service.

As a consequence, the Open Content Alliance, once seen as a rival potentially capable of competing with Google on equal terms, has failed to develop in this way, leaving the Internet Archive holding a significant collection of digitized books, but apparently without the prospect of multi-million-volume scanning continuing in the future. It seems possible that this grouping, and the related Open Library initiative, may end up competing with Google in a rather different sense - not only by coordinating smaller-scale scanning projects but also, potentially, by involving more Internet users in associated activities, such as cataloguing the wealth of material scattered across the Internet and improving the quality of the electronic texts.