Optimising Publications for Google Users

Alan Dawson

Centre for Digital Library Research, Department of Computer and Information Sciences,
University of Strathclyde, Glasgow G1 1XH, Scotland

alan.dawson@strath.ac.uk

June 2005

Keywords: digital libraries, metadata, search engines, optimization, electronic publishing

Abstract

This article examines the responsibilities of libraries and librarians as Internet information publishers, in view of the popularity of Google amongst users. It argues that librarians should think explicitly about Google users whenever they publish on the web, and should be prepared to update their policies and procedures accordingly. Drawing on experience and practical examples of publishing ebooks and other collections within the Glasgow Digital Library, the article describes procedures that libraries can adopt to ensure that their publications are optimised for access by users of Google and other web search engines. The aim of these procedures is to enhance resource discovery and information retrieval, and to enhance the reputation of libraries as valued custodians of published information, as well as exemplars of good practice in information management.

The rise of Google

The development of Google from search engine to multinational corporation reached a new stage in June 2005, when it became the world's biggest media company, measured by stock market value. The Google brand name is now familiar to millions of people who have never even used the web. Yet as Google introduces further innovations and extensions to searching, such as Google Print and Google Scholar, there is increasing concern about the impact that habitual Google use amongst students has on education in general and on information literacy in particular (Devine & Egger-Sider, 2004; Markland, 2005).

In most discussion of the effects of Google on libraries, librarians are usually seen as consumers and information seekers, whether for themselves or on behalf of students and other library users. In contrast, this article focuses on the role of libraries as online information providers, in view of the fact that so many people now depend on Google as their primary means of access to digital information.

While most librarians probably use the web and Google routinely in their work, the impact of the ubiquitous use of Google on librarians is still unclear. It is tempting to conclude that the universal availability of fast and easy access to online information is devaluing, and perhaps undermining, the traditional role of librarians as intermediaries. Yet, as Becker (2003), Bundy (2004), Rumsey (2005), and others have shown, there is still a great need for librarians as educators, to help guide information seekers to relevant sources and to interpret search results. The main difficulty for librarians in promoting information literacy is comparable to Rumsfeld's (2002) perceptive observations about unknown unknowns; if users don't know there are things they don't know about Internet searching, then they may not perceive any need for librarians to help them. This principle may also be applied to librarians; if they don't know there are things they don't know about online publishing, they may not perceive any need for guidance. This article therefore aims to change some "unknown unknowns" about the publishing process to "known knowns".

The responsibility of librarians

Traditional values associated with librarians include precision, discipline, attention to detail, and helpfulness. They may be seen as curators or custodians of books and other sources of information, but librarians are not usually regarded as publishers of information. Yet almost all libraries do publish, on a small scale at least. For some libraries this may amount to little more than details of opening hours and

location posted on a website. For many others it will extend to making available their catalogue of holdings online, along with guidance notes, while an increasing number of libraries are digitising some of their collections and publishing them online.

Digital publishing involves responsibilities that are often overlooked. While most libraries will think about issues such as accuracy, design and accessibility while publishing information for users via their own websites, they may be less aware that they are also (potentially) publishing information via Google. As yet “publishing via Google” is not a common expression. Google is seen as a search engine, not a means of publication. Yet, as Dempsey (2004) points out, increasingly “on web” means “available via Google”. Online publishing can therefore be seen as a two-stage process: adding content to a website is merely the first stage of publishing. The second stage is getting that content successfully indexed by Google. This second stage is arguably more important, as it vastly increases the likelihood of the published material being discovered and used. However, for librarians it should not be simply a question of ensuring that their material is “published via Google”. Librarians should aim to set a good example as digital curators (who else is better placed?). This means taking the necessary steps to ensure that the indexing and retrieval process works as well as possible, given the inherent limitations of Google. The remainder of this paper aims to describe and explain these steps so that Google users are well served by librarians and by other responsible online publishers such as museums, universities and public institutions. In short, it shows how and why to optimise online publications for Google. As well as producing benefits for those carrying out the optimisation, by making their content easier to maintain and discover, in the long term the cumulative effects of more widespread optimisation will benefit everyone.

The publishing process

The principles for publishing specifically with Google in mind are largely the same as the principles for sound information management and publishing practice; there is no inherent conflict. The table below summarises these principles and the reasons why they matter for publishing via Google.

What	*Why*
Allocate concise identifiers to publications according to a consistent scheme.	In conjunction with the domain name, the file name (or other identifier) will form the URL of a publication that will uniquely identify it in on the web. Ideally this will not be changed, so the naming scheme should allow for internal reorganisation or policy changes. A combination of prefix, letters and numbers may be more robust than long, readable item names.
Ensure that every published document is owned by a specific person.	Specifying who is responsible for every published document makes it less likely that documents will be duplicated, overlooked or become obsolete. Fostering a culture of digital curation can have intrinsic benefits for any organisation, and will also enhance the currency and value of Google indexing and searching.
Keep publications current.	It is always good practice and helpful to delete or archive old versions of documents, just as old notices should be removed from noticeboards. It is particularly important that they are removed from publicly accessible web directories, so that they are not found and indexed by Google, thereby cluttering search results and confusing users.
Use XHTML rather than proprietary formats where possible.	Although Google can index PDF and Word documents, these formats are less desirable because they are proprietary rather than open, they require additional software or plugins, and the documents are usually larger and slower to access. Research has shown that Google does not index large documents in full (Price, 2004). Several other disadvantages of PDF have been outlined by Dawson & Wallis (2005).
Comply with accessibility standards.	Keeping document formats relatively simple makes them easier to maintain, more likely to comply with web accessibility guidelines (W3C, 2004), and more accessible to search engine robots. Relatively complex features of web pages, such as frames, forms, scripts, animations, logins

	and session identifiers, are deterrents to Google indexing as well as to people with disabilities, so should only be used if essential for a specific service or application. They are rarely required for routine publication. Use of meaningful ALT text in <img> tags will enhance indexing as well as promote accessibility (Calishain & Dornfest, 2003).
Use relative internal links, not absolute links.	To minimise the likelihood of broken links. Collections can be moved and all internal links will still work.
Use style sheets to control formatting.	Document design can easily be changed. Removing unnecessary markup, e.g. in font control tags, improves the content-to-markup ratio, making documents easier for the Google software robots to index and helping achieve a higher ranking in Google search results. Guidance on use of CSS is given by W3C (2005), while some examples of simple and sophisticated CSS usage are given by the Glasgow Digital Library ebook collection (`http://gdl.cdlr.strath.ac.uk/gdlebooks.html`) and the CSS Zen Garden (`http://www.csszengarden.com/`) respectively.

A simple example may serve to illustrate the consequences of not following the above guidance. Searching Google for "Journal of Internet Cataloging" (in June 2005) produced about 7200 hits. Top of the list of search results was the Journal of Internet Cataloging home page (`http://www.internetcataloging.com/`) and second was its table of contents. The very high relevance of these first two results to the search term illustrates the foundation on which Google has built its formidable reputation. Why would anyone wish to look beyond those top two results?

Following either link provides access to the table of contents of the journal from Volume 1 Number 1 in 1997 to Volume 5 Number 1 in 2001. There are no later issues available. This may easily lead the casual searcher to conclude that the journal has ceased publication, or that it no longer published tables of contents online. Neither is true. In fact the journal is alive and well, and accessible via the Haworth Press Online Catalog, which appears at numbers 3, 4 and 5 in the search results.

These results are far from ideal, but the fault does not lie with Google, which is doing a good job in ranking the most relevant results so highly. The appearance of obsolete and duplicate information in the search results is the responsibility of the publishers, who have failed to clear up after moving the content to a new location. Even in the new location, there appear to be two home pages for the journal, offering different information. Of course it is better to have some duplication than no information at all, but this type of duplication and retention of obsolete information, though very common on the web, would not be regarded as acceptable in a library catalogue, and goes against the well-established principles of precision in information management. Rather than criticise Google for its failings (in this case and most others it works rather well), it is surely the responsibility of librarians, cataloguers and publishers, large and small, to take more care in managing their own online publications. By doing so both publishers and users will benefit, in ensuring that searchers can more easily locate current publications.

These principles and guidelines for good publishing practice, as summarised and advocated above, are already well known and widely accepted, though often poorly implemented. However, relatively little has been published about optimising metadata, about republishing documents, or about the fine tuning of indexing by Google. These topics are addressed below.

Metadata optimisation

It is widely known that Google (along with most other search engines) does not make use of information held in HTML <meta> tags when indexing and ranking documents (de Groat, 2002; Sullivan, 2002). However, this does not mean that metadata is irrelevant to Google. On the contrary, it increases the significance of the one metadata element that Google does use: the HTML <title> tag.

Title tags are important for three reasons: firstly, because the Google search algorithms give them significant weight; secondly, because users see title tag contents highlighted in their search results; and thirdly because title tags become the default names for bookmarks in web browsers. It therefore follows

that anyone who wishes to encourage access to their publications should use the title tag carefully and consistently. When publishing online, librarians should make sure that every document and every web page contains a title tag that applies accurately and specifically to the contents. To pursue the above example of the Journal of Internet Cataloging, the main reason that Google search results were so relevant was because the publishers had helpfully ensured that the journal title appeared in the title tag (so they had done part of their job well). Some further guidance on use of the title tag to enhance Google searching is given by Dawson (2004).

One particularly important consideration for metadata optimisation is the way in which long documents are split into separate sections or pages for online access. Views and practices on this vary considerably. Many publishers make complete documents available as single large PDF files, whereas others aim to limit information to a single screen so that users do not have to scroll down a page to see all its content. For example, Wilson & Landoni (2002), in their guidelines for electronic textbook design, recommend the use of short pages as they can "increase users' intake of information". However, they were concerned only with the usability of ebooks (where the benefits of chunking content into very short pages are arguable), and did not consider the effects of chunking on resource discovery. As well as causing an unwieldy proliferation of pages, such fine chunking makes it less likely that pages will be located via Google (as each one will have less content and fewer links) and also makes it more difficult to create accurate and meaningful title tags for each page. A better strategy is therefore to organise publications so that web pages reflect the natural structure of the content. For example, a large ebook should be split into separate chapters, and perhaps into separate sections within chapters, but not divided further by paragraphs, and not organised by printed page, as paper pages usually represent an artifact of the printing process rather than inherent document structure. As well as enhancing usability and speed of access, chunking content by chapter or section allows title tags to be varied to include the titles of chapters or sections, along with the title of the overall work. This aspect of metadata optimisation is a significant aid to resource discovery via Google, as has been demonstrated by Dawson & Hamilton (2005). The process can be automated for large-scale publication of structured documents, so that it becomes highly cost-effective, although the mechanism for doing this varies according to context and is beyond the scope of this article.

The republication process

The well-meaning slogan "cool URLs don't change" has not permeated far into the ranks of webmasters and content managers. Reorganised websites and broken links seem as common as ever. Yet this is not surprising. There are often good reasons for moving documents and reorganising websites. Institutions change, people come and go, departments merge and split, and publications need to be updated to reflect new policies and realities. Most web publishing is by nature volatile. However, this is bad news for web searchers, as links to old locations persist in Google indexes long after the content has been moved or deleted. Solutions to the problem do exist but are often overlooked. A brief step-by-step guide to moving, renaming, or republishing document collections is therefore given below. This is intended to inform and encourage librarians and others to play their part in reducing the persistent problem of broken links, and thereby improving the value of search results. Note that this guidance applies *only if the documents are being renamed or restructured*. If the content is simply being updated, using the same file names and structure, then the problem of broken links does not arise and the guidance is not applicable.

Step 1. Publish the new collection to a new directory or folder on the web server, while leaving the old version in place, so that there is temporary duplication.

Step 2. Add links to the publications in their new location, and remove all links to the old location, but leave the old collection in place on the web server.

Step 3. Edit all the documents in the old folder to prevent them being indexed by Google, by inserting the relevant instructions to all web crawlers (not just Google) into the <head> element of each web page. This can be done by using any program that will do a global search-and-replace on all the files in a folder. For example, a simple macro could be used to locate

```
</head>
```

in every document and change it to:

```
<meta name="robots" content="noindex,nofollow">
</head>
```

Alternatively, if the documents are published via a content management system, then the relevant template needs to be changed to incorporate the new line, but it is important to ensure that the template only applies to the old documents.

This step will prevent all documents in the old folder from being indexed in future, but it does not remove references to them from the existing Google index. Searchers will therefore still be able to find the documents via Google, and links to them will still work. It is true that users will find the old versions, but *that is the only one available to Google users at this stage*, for second-stage publication of the new version has not yet occurred.

Step 4. Configure the web server (e.g. IIS or Apache) so that any http connections to the old folder are redirected to the new one (this may require liaison with a webmaster or system support personnel). It is possible to redirect all files within a folder to a single new location, or to redirect specific old files individually to specific new files. The latter option is more time-consuming but more helpful to users.

Step 5 (optional). In order to assess whether any individual redirection is required, use Google to find out which old pages are being linked to by external websites, by using the `link:` prefix in the search term, e.g. `link:gdl.cdlr.strath.ac.uk/redclyde`. The search results will display pages (internal and external) that link to the specified URL, and can be used to identify pages requiring individual redirection (although the results of using the `link:` prefix are not always comprehensive). An alternative strategy is to use a special-purpose program to automate individual redirection of old pages to new pages. This program will be server-dependent (for example, ISAPI_Rewrite for IIS or mod_rewrite for Apache) and will usually require the assistance of technical staff, but is worth considering if a large and heavily-used collection is being moved.

After carrying out step 4 or 5, users should be able to find the new documents via the old file locations, with no broken links. If implementation of redirects presents a problem, one solution is to use the PURL service (`http://purl.oclc.org/`) to maintain persistent identifiers and redirections.

Step 6. Search Google about once a week to see when the documents in the new location have been indexed, by typing in a relevant search term (e.g. a distinctive phrase from one of the documents) and noting the URLs displayed in search results. Once the indexing of new files has occurred then the old files can be removed, although the old folder and the redirections should be maintained for as long as links to old locations appear via Google.

Step 7. When it is clear that searching Google produces only links to the new publications, with none to the old versions, then the old folder, and the redirections, can be removed. However, this should only be done if there are no external links to the old collection. In practice, retaining long-term redirections should not cause any problems. The crucial issue is to maintain the link between the URL and the content, even if one or the other changes.

Although these procedures may sound complex, the whole process is conceptually fairly simple. This level of attention to detail should ensure that users do not get broken links when attempting to access the content, even after it has been moved. Rather than being frustrated by the inadequacies of Google and the preponderance of broken links, librarians can help the situation by taking greater responsibility for maintaining the effectiveness of Internet searching.

The above guidance draws on practical experience of reorganising, renaming, and republishing specific digital library collections, but it is not intended to be exhaustive, as there are different means of achieving the aims of persistent publishing. In practice, some types of republishing are simpler than others, and other issues not covered above may arise. A simple case study may serve to illustrate this. For example, owing to a change in policy as the Glasgow Digital Library expanded, it was decided to change the collection identifier prefix for the Springburn Virtual Museum from two characters (“sp”) to six characters (“spring”). Applying the change was trivial – the collection identifier field in the database was changed manually, a global change was made to all the item identifiers in the database, and all the relevant image files (including backup copies) were renamed using a bulk renaming program. All the relevant documents were then recreated automatically from the database, with the new six-character

prefix, and copied to the web server. However, in this case the new files were copied to the same folder as the old files, so that *the main URL for the collection would be unchanged*. This was a simple case of renaming files within a single folder, with no content restructuring. The other steps outlined above were then followed, and the changeover occurred smoothly. A more common, and more complex, scenario involves the restructuring of large collections. In such cases it is advisable to keep the old and new collections separate rather than mix old and new files in the same folder structure.

Controlling indexing by Google

The guidance above recommends using the `<meta name="robots" content="noindex,nofollow">` tag to prevent old versions of web pages from being indexed by search engine robots. This is a useful and effective tool for controlling publishing via Google, but its use is not limited to the republication process. Before describing how this tool may be used routinely in publishing, it is worth summarising the effects of the variations in syntax:

`<meta name="robots" content="index,follow">`	allows full robot access. This is the default, and is therefore redundant.
`<meta name="robots" content="noindex,follow">`	allows robots to follow links but not to index content. This allows precise control over indexing and can be very useful.
`<meta name="robots" content="index,nofollow">`	allows robots to index content but not follow links. This is less likely to be useful.
`<meta name="robots" content="noindex,nofollow">`	prevents robots from either indexing content or following links.

Use of these tags (especially the “noindex” option) is so simple and effective in controlling website indexing that the real difficulty is in deciding when to use them. The main value is in preventing unnecessary duplication, for if everything is indexed then duplication is likely, particularly where publishers try to be helpful by providing more than one route of access to the same content. However, it is not always obvious when to switch off indexing. The guidance offered below is therefore illustrative rather than prescriptive. The easiest way to illustrate the subtleties of controlling indexing is by using specific examples of handling different types of object within a digital library.

Example 1. Document text **Probable setting: index,follow**

Normally, publishers will want the substantive content of any publication to be indexed, so that it can be readily located. The only reason for preventing indexing would be to limit access.
Example page: `http://gdl.cdlr.strath.ac.uk/100men/gm66.htm`

Example 2. Title pages **Probable setting: index,follow**

If all the text on a title page is repeated on substantive pages, then indexing both title page and text pages will produce duplication of search results. Despite this, indexing of title pages is recommended, because they are useful signposts to users, and because they attract more external links than other pages, so are likely to be ranked relatively highly in search results.
Example page: `http://gdl.cdlr.strath.ac.uk/smihou/`

Example 3. Tables of contents **Probable setting: noindex,follow**

If all the text on a table of contents page also appears in the full text of a document (or in the abstracts of journal articles), then there is no need for the table of contents to be indexed (users would prefer to go directly to the relevant text). However, if neither full text nor abstract are available online, then the table of contents should be indexed.
Example page: `http://gdl.cdlr.strath.ac.uk/haynin/hayningontents.html`

Example 4. Combined title and contents pages **Probable setting: index,follow**

If the same page serves as the title page and the table of contents page, then there are conflicting arguments. In such cases indexing is advisable, as a little duplication is better than no information.
Example page: `http://gdl.cdlr.strath.ac.uk/minstr/`

Example 5. Chapter contents **Probable setting: index,follow**

As well as a table of contents, some large documents, such as ebooks, may have a contents page for each chapter or section. While the same argument against indexing can be applied as for the main table of contents, there is a subtle difference; the chapter itself may have a title. If the user's search term matches a word or phrase in the chapter title, then the best result is to display the chapter contents page, rather than any other occurrences of that term. It is also possible for the term to appear only in the chapter title but not in the text. Optimal results may not be possible in all cases, but it seems advisable to leave indexing on.
Example page: `http://gdl.cdlr.strath.ac.uk/keacam/keacam02.htm`

Example 6. Image wrapper pages **Probable setting: noindex,nofollow**

Wrapper pages are commonly used to add branding and navigation to a collection of images. While web browsers will happily display links to unwrapped jpegs or other image files, it is common practice to wrap such images inside a web page offering a familiar identity and interface. If indexing is left on, such wrapper pages can add immensely to the clutter and duplication in search results. The problem is so well-known and long-standing (and so few people bother to control indexing) that Google has taken measures to counteract it by suppressing duplication in its search results (hence the common message "*In order to show you the most relevant results, we have omitted some entries very similar to [those] already displayed*"). While Google's deduplication is indeed helpful, a more robust and satisfactory solution is for information providers to prevent the problem arising by suppressing indexing of wrapper pages that have no unique text content.
Example page: `http://gdl.cdlr.strath.ac.uk/aspect/aspect2003/sld/a03sldgba01a.html`

Example 7. Back-of-book indexes **Probable setting: noindex,nofollow**

Back-of-book indexes provide entry points and navigation for paper publications. They are not common in ebooks or other online publications, and it may be thought that the prevalence of searching has rendered such indexes redundant. Yet research has shown that, where indexes exists, users both value them (Wilson & Landoni, 2002) and can find information more quickly using them than via searching (Barnum et al, 2004). But should the indexes themselves be indexed? Doing so does undoubtedly create redundancy, as most terms appearing in an index also appear in the full text. Yet one of the arts of manual indexing is to use index terms that do not appear in the text itself. If a user search term matches such an index term, and the index is not indexed, then no match will be found, even though the content is relevant. But indexes are designed for browsing, not searching. There are clearly arguments either way. On balance, perhaps it is not worth indexing indexes unless they are known to contain a significant proportion of terms that are not found in the full text. Judgment is best made after testing some illustrative search terms in specific cases.
Example page: `http://gdl.cdlr.strath.ac.uk/stecit/stecitindextopic.html`

Example 8. Subject indexes **Probable setting: index,nofollow**

It is surprising that so few libraries or academic publishers provide access to their online publications via a controlled set of subject terms such as LCSH. Journal publishers routinely list issues in date order, and sometimes offer an author index, but rarely a subject index. Where a subject index does exist, the same considerations apply as to back-of-book indexes. However, the issue is not clear-cut. Preliminary evidence from the Glasgow Digital Library suggests that allowing subject indexes to be indexed by Google can increase the probability of relevant items being located via Google searches, even though the subject indexes are designed for browsing rather than searching. Again there is a balance to be struck between maximising resource discovery and minimising redundancy, and the best policy is not obvious.

As controlled subject terms are less likely to appear in the full text than back-of-book index terms, there are probably stronger arguments for indexing subject indexes than back-of-book indexes.
Example page: `http://gdl.cdlr.strath.ac.uk/subjects/gdlindexsubjects.html`

Example 9. Multiple document formats **Probable setting: index,nofollow**

If the same document is published in different formats, e.g. HTML and PDF or Word, it is hard to envisage any reason to index more than one version. However, the use of meta tags to suppress indexing will only work in HTML documents. An alternative method is therefore needed. One option is to include links to the alternative versions from the HTML version, along with the "nofollow" instruction to software robots. The disadvantage of this is that it will apply to all links, which may not be desirable. Another option is to store PDF and Word versions in a separate folder, and then prevent indexing of all documents in that folder by including the relevant instruction in a file called `robots.txt` at the top level of the web server. For example, indexing of everything in a folder called "pdfs" could be suppressed by adding the two lines

```
User-agent: *
Disallow: /pdfs/
```

to the `robots.txt` file. This would require suitable access permissions and possible liaison with technical staff. Further information is available from `http://www.robotstxt.org/`.

Principles of indexing

The above examples show that once the mechanism for controlling indexing is understood, it becomes fairly simple to control. The difficulty lies in deciding what works best for users. In order to help decide what to index, some basic principles of web indexing can be stated as follows:

- Any page that contains unique text content should be indexed.
- Duplication in search results should be avoided where possible.
- A little duplication is better than an empty result set.

In other words, while one match between search term and result might be optimal, two matches are better than none. These principles can then be applied to specific types of document as illustrated above. In a large collection it would be tedious to individually control indexing for every document, but this is not necessary. The main requirement is to be able to identify the distinct document types, and then to automate the process of index control. For example, if documents are automatically generated from a database or content management system, the program or template that controls output needs to be able to identify which type of document is being output, and to enable or suppress indexing in accordance with specific rules that reflect agreed indexing principles. This level of precision is perfectly achievable, but in practice requires some thought and testing to produce optimal results.

Google as a local search engine

The main point about "publishing via Google" is that it allows users to readily find relevant resources without having to know where to look (other than Google). However, having located a relevant website, users may wish to explore in more depth the resources available therein, by browsing or searching within a specific site. While this can be done from the Google home page, using the relevant syntax to restrict searching to a specific site, in practice few users do this. It therefore makes sense to consider using the facilities of Google to offer a local search service. Any website (or a section of a website) that is already indexed by Google can be made locally searchable by adding a search box and pre-limiting a Google search to a specific domain or folder. For example, adding the following markup to a web page would limit all searches to the domain cdlr.strath.ac.uk

```
<form method="get" action="http://www.google.com/search?">
<input type="text" name="q" size="30" maxlength="255" value="">
<input type="hidden" name="q" value="site:cdlr.strath.ac.uk">
<input type="submit" name="sa" value="Search">
</form>
```

Changing the third line to

```
<input type="hidden" name="q" value="inurl:cdlr.strath.ac.uk/pubs">
```

would restrict the search even further to the “pubs” folder of the same domain. (Much of the time, the `site:` and `inurl:` prefixes deliver the same result set, though Calishain & Dornfest (2003) assert that `inurl:` is generally more flexible.) This facility to restrict searches to specific folders is immensely useful, as it offers a simple means of making separate collections, or individual ebooks, independently searchable. It does however require that the organisation of folders (directories) corresponds to the structure of collections.

While most large collections already have their own search facilities, these can be expensive to set up and maintain. In contrast, Google is quick and easy to set up as a local search service, it is easy to customise the display of search results, it is familiar to users, searches are very fast, and it is free. Furthermore, although Google makes no use of metadata other than titles, its relevance ranking usually works remarkably well, and the summaries are often useful. For example, searching the University of Strathclyde websites locally via Google for the very common word “library” (in 2005) produces over 18,000 matches, with the main university library top of the list and the next three being major library-related resources. The relevance ranking is excellent, and the result ordering is just what a collection manager or user might hope for. Google is evidently capable of performing effectively at a local level, despite the lack of metadata, just as it does on a global scale.

On the negative side, because Google does not use metadata, searches can not be limited to author, subject or date fields (it is possible to limit searches to titles, but relatively few users do this). While this may be acceptable for a large and irregular collection of documents and departments, such as a university, it is far less acceptable for a tightly-focused and well-catalogued collection such as a journal archive. Search options are limited to those provided by Google, so useful features such as stemming are not supported. And perhaps more critically, the indexing of content is irregular, with a time lag (usually up to one month) between publication and searchability. Perhaps the biggest disadvantage is that collection managers do not have full control of the search algorithms, but this may be a price worth paying for a highly cost-effective service.

Some of the other limitations can be overcome using a combination of ingenuity and diligence. For example, field searching can be simulated by ensuring that the field name appears next to the content (e.g. Author: Hovis Presley), and then embedding the field name in the search form, in the same way that domain names can be embedded in forms to restrict searching to a specific site. The results are not as precise as genuine field searching but may well be sufficient for most purposes.

If Google is used to provide a local search service, the issues of republication, metadata optimisation, and indexing control become even more important, because duplication, inaccuracies, and inconsistencies are far more noticeable in a small result set. Furthermore, users are more likely to judge these as being the responsibility of the local institution rather than an inevitable consequence of searching across the whole of the web.

Conclusions

It is easy to criticise Google for not allowing the same precision as library catalogues and specialist databases, for returning too many results, with too much duplication, for including results that are out of date, irrelevant or superficial, and for failing to index the “invisible web” (Sherman & Price, 2001). While these criticisms are valid up to a point, they are not so much criticisms of Google itself, which offers a superb large-scale service, as criticisms of those who publish carelessly on the web, by leaving old or duplicate documents lying around for Google to find, by moving or deleting documents without leaving any redirection, or by discouraging Google software robots from indexing their content. Librarians and other information professionals must take their share of this criticism unless they follow sound procedures to optimise publication via Google. This article has outlined the issues and shown how the overall effectiveness of Google can be improved, not by making changes to search algorithms or metadata standards, but simply by behaving professionally and taking care to publish responsibly.

By improving their own practices in online publishing, and understanding how and why these affect the retrievability of publications via Google, librarians will eventually be in a better position to educate others to follow similar sound practices. They will therefore be developing a new role for themselves – that of

disseminating good practice in information provision as well as in information retrieval – and will be contributing to the broad view of information literacy that incorporates information management and technological fluency, as advocated by writers such as Bundy (2004) and Warnken (2004).

In the long term, the increased precision of resource description and indexing that arise from enhanced online publishing will improve the match between search terms and search results, so will ultimately improve information retrieval on the web. As more and more people become electronic publishers, there will be benefits for both information providers and information seekers in extending epublishing literacy.

References

Barnum, C., Henderson, E., Hood, A., Jordan, R. (2004). Index Versus Full-text Search: A Usability Study of User Preference and Performance, *Technical Communication*, Vol. 51 No. 2, pp. 185-206.

Becker, N. J. (2003). Google in perspective: understanding and enhancing student search skills, *New Review of Academic Librarianship*, Vol. 9, pp. 84-100.

Bundy, A. (2004). One essential direction: information literacy, information technology fluency. *Journal of eLiteracy*, Vol. 1, pp 7-22.

Calishain, T. and Dornfest, R. (2003). *Google Hacks*, O'Reilly, Sebastopol, CA.

Dawson, A. (2004). Creating metadata that works for digital libraries and Google. *Library Review*, Vol. 53 No. 7, pp. 347-350. `http://cdlr.strath.ac.uk/pubs/dawsona/ad200402.htm`

Dawson, A. and Hamilton, V. (2005). Optimising metadata to make high-value content more accessible to Google users. *Journal of Documentation*. In press.

Dawson, A. and Wallis, J. (2005). Twenty issues in e-book creation, *Against the Grain*, Vol. 17 No. 1, pp. 18-24. `http://cdlr.strath.ac.uk/pubs/dawsona/ad200501.htm`

De Groat, G. (2002). Perspectives on the Web and Google: Monika Henziger, Director of Research, Google, *Journal of Internet Cataloging*, Vol. 5 No. 1, pp. 17-28.

Dempsey. L. (2004). The three stages of library search. *Cilip Update*, November 2004. `http://www.cilip.org.uk/publications/updatemagazine/archive/archive2004/november/lorcan.htm`

Devine, J. and Egger-Sider, F. (2004). Beyond Google: The invisible web in the academic library, *The Journal of Academic Librarianship*, Vol. 30 No. 4, pp. 265-269.

Markland, M. (2005). Does the student's love of the search engine mean that high quality online academic resources are being missed? *Performance Measurement and Metrics*, Vol. 6 No. 1, pp. 19-31.

Price, G. (2004). A couple of comments About Google. `http://www.pandia.com/post/020-2.html`

Rumsey, S. (2005). Search interfaces for dummies? Paper presented at *LILAC 2005: Librarians Information Literacy Annual Conference*. `http://www.cilip.org.uk/groups/csg/csg%5Filg/Lilac05/Papers/rumsey.pdf`

Rumsfeld, D. (2002). Briefing for United States Department of Defense, 12 Feb 2002.

Sherman, C. and Price, G. (2001). *The invisible web: uncovering information sources search engines can't see*. CyberAge Books. `http://www.invisible-web.net/`

Sullivan, D. (2002). *Search Engine Features for Webmasters*. `http://searchenginewatch.com/webmasters/article.php/2167891`.

W3C (2004). Web Content Accessibility Guidelines 2.0. `http://www.w3.org/TR/WCAG20/`

W3C (2005). Cascading style sheets home page. `http://www.w3.org/Style/CSS/`

Warnken, P. (2004). The impact of technology on information literacy education in libraries. *The Journal of Academic Librarianship*, Vol. 30 No. 2, pp. 151-156.

Wilson, R. and Landoni, M. (2002). *Electronic textbook design guidelines*. `http://ebooks.strath.ac.uk/eboni/guidelines/`