The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0022-0418.htm

Optimising metadata to make high-value content more accessible to Google users

Alan Dawson and Val Hamilton
Centre for Digital Library Research, University of Strathclyde, Glasgow, UK

Received November 2004
Revised July 2005
Accepted July 2005

Abstract

Purpose – This paper aims to show how information in digital collections that have been catalogued using high-quality metadata can be retrieved more easily by users of search engines such as Google.

Design/methodology/approach – The research and proposals described arose from an investigation into the observed phenomenon that pages from the Glasgow Digital Library (gdl.cdlr.strath.ac.uk) were regularly appearing near the top of Google search results shortly after publication, without any deliberate effort to achieve this. The reasons for this phenomenon are now well understood and are described in the second part of the paper. The first part provides context with a review of the impact of Google and a summary of recent initiatives by commercial publishers to make their content more visible to search engines.

Findings – The literature research provides firm evidence of a trend amongst publishers to ensure that their online content is indexed by Google, in recognition of its popularity with internet users. The practical research demonstrates how search engine accessibility can be compatible with use of established collection management principles and high-quality metadata.

Originality/value – The concept of data shoogling is introduced, involving some simple techniques for metadata optimisation. Details of its practical application are given, to illustrate how those working in academic, cultural and public-sector organisations could make their digital collections more easily accessible via search engines, without compromising any existing standards and practices.

Keywords Digital libraries, Search engines, Optimization techniques

Paper type Research paper

Introduction

The mission of Google Inc. is "to organise the world's information and make it universally accessible and useful". The mission of the Dublin Core Metadata Initiative is "to make it easier to find resources using the internet". Despite the similarity of mission statements, the remarkable success of Google owes nothing to Dublin Core or any other metadata scheme. This paper proposes some simple and practical measures for bridging the gulf between the Google world and the metadata world in order to make both more effective for information retrieval. The proposals described are based on solid evidence of their success in practice, yet are also grounded in well-established principles of information organisation and resource description.

In some quarters, metadata has been seen as the panacea for the problems of finding information on the internet. During the late-1990s, in particular, around the time of the

The authors wish to thank Simon Bains at the National Library of Scotland for information supplied, and colleagues at the Centre for Digital Library Research for their helpful comments on earlier drafts.

Journal of Documentation
Vol. 62 No. 3, 2006
pp. 307-327

0022-0418
DOI 10.1108/00220410610666484

development of the Resource Description Framework and the Dublin Core, many articles were written claiming that the chaos of the internet would soon be tamed once web site developers started using such schemes. For example, Marchiori (1998) wrote:

> The only feasible way to radically improve the situation is to add to web objects a metadata classification, that is to say partially passing the task of classifying the content of web objects from search engines and repositories to the users who are building and maintaining such objects.

However, several years on, people do find information on the web, but not because of well-structured, semantically useful metadata. As Arms and Arms (2004) have observed:

> With the benefit of hindsight, we now see that the web search engines have developed new techniques and have adapted to a huge scale, while cross-domain metadata schemes have made less progress.

Given the great investment of time and effort expended on discussion and development of metadata schemes in recent years, it is not easy for the metadata community to acknowledge the enormous impact of an information retrieval tool that does not rely on metadata. There are many reasons for the success and public acceptance of the market leader Google, but metadata is not one of them.

Google is great

The achievements of Google are often either taken for granted or not given due acknowledgement in academic circles, so it is worth summarising them here. Google is extremely fast and reliable, it works on a massive scale, it produces useful results much of the time, it searches the full text of documents, it indexes multiple document types, including HTML, PDF and Word documents, it generates contextual summaries that are often useful, it is constantly updated, it has many advanced features for those who can be bothered to use them, it is very simple to use, and it is entirely free to users worldwide. It is an immensely valuable and widely used service that benefits from regular extension and innovation. The development of effective search engines such as Google has expanded easy access to information. Specialist training is no longer required to achieve results of some sort, and many users put a premium on speed over comprehensiveness and maybe even quality. As Wallis (2003) says:

> For sufferers of "information anxiety" the simplicity of the Google search interface must act as a calming tonic. It is not demanding of the information seeker in the formulation of search terms and almost always produces vast numbers of hits. It even helps out with your spelling.

Google's current pre-eminence as a search engine is well documented. The verb "to google" has entered the English language, and in an extensive survey (Brandchannel, 2004), Google was rated the world's number one brand name, above Apple, Mini, Coca-Cola, Samsung, Ikea and Nokia. When Google was unavailable for a few hours on 26 July 2004, a spate of newspaper articles ensued, written by shocked journalists. Dowling (2004), in *The Guardian* newspaper, described "A worrying glimpse over the lip of a murky abyss" while Mangold (2004) had his "first near-death experience".

Google is great but ...

While most Google users appear satisfied with its results (OCLC, 2003), information scientists are well aware of its limitations. Paul Miller, Common Information Environment Director for the UK Joint Information Systems Committee (JISC), gives credit where it is due but sums up the position:

> Google is great. Personally, I use it every day, and it is undeniably extremely good at finding stuff in the largely unstructured chaos that is the public web. However, like most tools, Google cannot do everything. Faced with a focussed request to retrieve richly structured information such as that to be found in the databases of our memory institutions, hospitals, schools, colleges or universities, Google and others among the current generation of Internet search engines struggle. What little information they manage to retrieve from these repositories is buried among thousands or millions of hits from sources with widely varying degrees of accuracy, authority, relevance and appropriateness (Miller, 2004).

There are other fundamental limitations too. Google cannot meet what Svenonius (2000) calls the:

> ... essential and defining objective of a system for organizing information [...] to bring essentially like information together and to differentiate what is not exactly alike.

So Google is unable to distinguish between "J.K. Rowling" as subject and "J.K. Rowling" as author. The Google Scholar service does allow searches to be restricted to authors or dates, but the facility is flawed and its effectiveness limited (Jacso, 2004, 2005a, b). Searching for words which have many synonyms is also problematic, whereas in a more sophisticated information system (perhaps one designed by librarians), structure would be given through the use of controlled subject headings. There are also internet resources that search engines cannot index; perhaps because the pages are dynamically generated, or require registration before access, or contain no indexable text, or because they have been excluded by webmasters, either deliberately or inadvertently.

Alternatives to Google

In the UK the proposed solution to the problems described by Miller is the creation of a "common information environment" for all UK citizens. This is an ambitious extension of the type of solution that has been developed in more restricted subject areas. Examples include the GEM gateway to educational materials (www.geminfo.org), which provides "quick and easy access to thousands of educational resources" and the open language archives community (www.language-archives.org) which aims to create "a worldwide virtual library of language resources".

The JISC common information environment *Vision for the Future* states that there are many technical, commercial and administrative barriers in its path.

> Overcoming these barriers will require concerted action on the part of all organisations in the field. It will take time and it will not be easy (JISC, 2004).

But if, after great effort and expense, such a "common information environment" were to be created, the question arises, would people use it, or would they just continue to use Google anyway? Becker (2003), in a paper on enhancing student search skills, raises the difficulty of "interrupting the automatic Google search response". The current reality is that for most internet users, and that includes academics and students

(Urquhart *et al.*, 2003), a powerful and pervasive common information environment already exists, and it is called Google.

Metadata vs Google

The proposed JISC solution is an example of a trend that sees commercial producers courting the search engines while most of the academic world is taking a different direction. This trend can be seen in the development of the metadata community, which at times seems introverted and embroiled in debates about the fine detail of competing standards. Despite a mantra of interoperability, attention is rarely given to the question of how to ensure that meticulously crafted metadata is used beyond the confines of its immediate surroundings. The existence of search engines is ignored or denigrated. It seems that the creators of many useful web sites are happy to belong to the invisible or deep web of items which standard search engines cannot or will not find (Sherman and Price, 2001).

The value of metadata in principle is undeniable. In different guises, such as cataloguing according to AACR2, it has long been a fundamental part of traditional information management. In theory metadata should make search engines more efficient, so the question arises, why have search engines not been making use of metadata? The main reason stems from a difference between traditional information retrieval and the digital world, as Lynch (2001) points out in an important article, "When documents deceive". Traditionally:

> Metadata (surrogate records) for documents can be taken at face value as honest attempts to accurately describe documents, and should be treated this way in retrieval systems.

But in the digital world, this is not necessarily the case:

> ... the metadata may be carefully constructed by any number of parties to manipulate the behavior of retrieval systems that use it, rather than simply describing the documents or other digital objects it may be associated with (Lynch, 2001).

It is because of this type of abuse that Google and other search engines do not use information in HTML meta tags (de Groat, 2002; Sullivan, 2002). Of course this situation may change. Methods of assuring the trustworthiness of metadata may be introduced, and even now there are indications that Google privileges information from reputable sites, such as those with ".gov" or ".edu" or ".ac.uk" domain names. There are also indications of focused approaches, for example, the joint project between Google and the DSpace institutional "superarchive" system to improve access to institutional archives of 17 universities worldwide (Open Access Now, 2004). The Google Scholar and Google Print initiatives (Banks, 2005) represent even more ambitious attempts to enhance the quality of Google search results by concentrating on publications from reputable sources.

The other major problem with metadata, once standards have been agreed, is the question raised by Thomas and Griffin (1998) in their paper "Who will create the metadata for the internet?" This clearly sets out the issues, and little has changed since its publication other than an increase in papers debating the question. Creation of detailed metadata is expensive, it is still not clear who should accept responsibility for the work, and it is difficult to organise efficient metadata creation procedures for distributed contributors even within a single institution, so the incentives to invest in

metadata creation are not compelling, particularly if Google can index web sites effectively without it.

A new approach: metadata and Google

It is ironic that those making the greatest investment in valuable electronic resources are not benefiting from the freely available power of search engines. This does not have to be the case. Instead of accepting that a mountain of high-quality and richly structured content is beyond the reach of Google, and therefore, looking for new (and expensive) solutions to make it more readily accessible to users, the alternative strategy is to get Google to come to this mountain and do justice to it by indexing it fully and effectively. The software is already proven, the networks are already in place, and the users are already convinced. Yet authors such as Miller seem to ignore the possibility that the problem may not lie with the search engines but with the data providers. All that is required is some adjustment to existing information repositories to make them more visible to Google users, and a strategy to ensure those adjustments take place.

An indication of the desirability of this process is provided by Calishain (2003), co-author of the acclaimed book *Google Hacks* (Calishain and Dornfest, 2003), who speculates on ways that "Google could expand what they have got". One suggestion is that Google should "reach out to information-collection publishers" by providing more details about how it indexes and ranks its content, along with guidance for responsible use. Calishain comments that:

> I'm referring to librarians and other information professionals who are often in charge of putting large collections of information online. Usually, those kinds of content publishers have far better things to do than spend extensive amounts of time trying to make sure their content gets indexed. This is a pity because it is exactly these kinds of information collections (extensive, unique, often not available anywhere else online) that are so valuable to search engines.

Clearly, someone who understands very well how Google works can appreciate the importance of optimising large unique collections for Google access. Calishain is not alone. There is already good evidence that some important organisations understand the value of taking this approach, as will be seen below.

Google-friendly publishing

Just as in the area of self-archiving of publications, it is the physics community that is leading the way in opening up their content for Google. Both the American Institute of Physics (AIP) and the Institute of Physics (IoP) have undertaken major initiatives to ensure that their content is indexed by Google, although by different means (Aldrich, 2004; Scholarly Information Strategies Ltd, 2003; Tenopir, 2004). AIP have created PhysicsFinder; a new, hierarchical version of their web site, and the results they report are remarkable:

> Since, the Google indexing of articles, PhysicsFinder has helped create a nearly 200 per cent increase in average monthly article sales for AIP (Aldrich, 2004).

IoP have maintained their database version but set up procedures to remove barriers to spidering by Google and other web crawlers, with Google IP addresses treated as legitimate subscribers able to access all content. The arXiv physics preprint archive

has also ensured that it is indexed by Google, and Inger (2004) reports that in the first month of indexing, usage of the service increased by 50 per cent.

Other commercial publishers are taking similar steps. CrossRef, the cross-publisher citation linking system, began a pilot collaboration with Google in April 2004 to allow indexing of full text content from 29 academic publishers. Searching of these resources is available both on the web sites of participating publishers and also through the usual Google site (CrossRef, 2004). Similarly, in March 2004, Google was given access to the full text of most of the journal articles hosted by Ingenta, a leading provider of online publishing services to academic and professional publishers. Previously Google had crawled only material available without charge. Under the new system, Google users reach the abstract page but are then offered either pay-per-view access or an authentication route to the full-text (if their host institution is a subscriber). Ingenta reports a dramatic jump in usage, and sums up the advantages:

> Greater visibility of publisher's full text articles; all words in an article are indexed as "searchable" on Google, not just the metadata; no security issues, as users are always passed to the abstract page and recognized as a subscriber or given pay-per-view options (Ingenta, 2004).

The Institute of Electrical and Electronics Engineers (IEEE, 2004) has not gone quite this far, but its publications have also seen greatly increased access since their abstracts began to be indexed by Google.

Similar examples of innovative developments are also appearing in the traditional library world. For example, an initiative between OCLC and Google resulted in records from the OCLC WorldCat union catalogue being made available to Google searchers (Jordan, 2004). OCLC (2004) reported that:

> During its pilot phase in 2004, the number of links from search engines to Open WorldCat's web interface grew substantially, from tens of thousands per month at the turn of the year to over three million per month in September 2004.

A smaller-scale example is the deliberate effort by the National Library of Scotland to ensure that its *Word on the Street* collection (www.nls.uk/broadsides) is indexable by search engines (Bains, 2004).

The developers of the open language archives community have taken a similarly inclusive view, stating that "encouraging people to document resources and make them accessible to search engines is part of our vision" (Bird and Simons, 2003). They have ensured that, as well as producing finely focused metadata for their own system, they have also made the data available for web crawlers.

A further development is the DP9 (2001) Gateway, which is described as:

> . . . an open source gateway service that allows general search engines (e.g. Google, Inktomi, etc.), to index OAI-compliant archives. DP9 does this by providing a persistent URL for repository records, and converting this to an OAI query against the appropriate repository when the URL is requested. This allows search engines that do not support the OAI protocol to index the "deep web" contained within OAI compliant repositories.

Such initiatives show a willingness to work with search engines rather than criticise them, but this is far from being the dominant approach in the academic and scientific world. Hunter and Guy (2004), for example, claim to illustrate "how existing information about available resources can be repurposed fairly easily and cheaply

using standard tools". However, their repurposing is only for OAI harvesters. They claim that their technique "offers the prospect of resource discovery far beyond what is currently available to users of the web via standard search engines" but they do not consider that they could include search engine users too.

Data shoogling

The above examples show that some commercial publishers already attach great importance to getting their content indexed by Google, and demonstrate the dramatic impact such indexing can have on usage figures. These publishers have been prepared to change their publication procedures in order to make their content more accessible to Google users. This process of reengineering a collection of information to make it more readily accessible via Google is sufficiently important to require a specific term. The Scots verb "shoogle" means to shake or jog, and therefore, the term "data shoogling" will be used in the rest of this paper to refer to the process of rejigging, or republishing, existing digital collections, and their associated metadata, for the specific purpose of making them more easily retrievable via Google.

There is an existing activity which attempts to achieve precisely this result, called "search engine optimisation" (SEO), and even an associated profession. The crucial elements that differentiate data shoogling from mere SEO are:

(1) It applies only to structured collections held in databases or content management systems.

(2) It is particularly relevant to organisations that wish to capitalise on the investment they have already made in creating metadata.

(3) It assumes that content managers are concerned with adhering to sound principles of collection management and digital preservation, which means enthusiasm for matters such as consistent and accurate resource description, adherence to established standards, and interoperability with other collections.

These three criteria apply to most large collections held by public-sector institutions such as universities, libraries, museums and government departments, so the data they hold ought to be of relatively good quality and consistency, but it needs shoogling to help people find it.

The rest of this paper describes some strategies and practical measures that content providers can adopt to ensure that their material is more easily and precisely retrieved via Google, thereby achieving the same goals that writers such as Miller and Calishain are seeking. Just as importantly, it emphasises how these measures can be put in place without compromising any of the standards for resource description or structural integrity that are common in digital libraries and other large online collections. The proposals, therefore, go well beyond standard techniques for SEO, by taking into account the need for adherence to well-established library standards, and the demands of interoperability and digital preservation. The proposals are also intended to be future proof, so that as Google inexorably develops, or becomes superseded by a better service, content providers can readily adapt to a different global information environment.

There are four main components of the data shoogling process, each of which is explained below:

(1) search engine optimisation;
(2) metadata cleaning;
(3) metadata optimisation; and
(4) metadata exporting.

However, before beginning any data shoogling, it is worth assessing each collection to decide whether it makes strategic sense to encourage and simplify worldwide access to it. While this may seem obviously desirable for digital libraries, it is not necessarily desirable for all metadata collections, e.g. catalogues of common items of local interest only.

Search engine optimisation

There is no need to go into detail about SEO here, as the principles and processes are well-established and documented elsewhere. For example, Kent (2004) has produced a useful and readable guide. However, it is worth summarising the most important points of SEO, as it is a key stage in the data shoogling process.

Use a well-established domain name. Many large online collections are hosted by public-sector institutions, such as universities, libraries and museums. These have the great advantage that their web sites have existed for years and are already well indexed by Google. Specific projects are often given a new domain name ending in .org or .com, but this can make them less likely to be found and indexed by search engines. Also, they will not be discovered if a search filter such as "site:edu" or "site:ac.uk" is used.

Use robot-friendly design. In brief, this means minimising or avoiding those elements that cause problems for search engine robots, such as frames, forms, scripts, animations, logins and session identifiers.

Use stylesheets and avoid markup clutter. Cascading stylesheets are immensely useful in their own right, because they help separate structure from display format and content from design. They are also good for SEO, as web pages are smaller and cleaner, with a high content-to-markup ratio.

Include good text content. It is unclear how search engines rank the value of content, but it is certain that names and nouns are more highly rated than common words or instructions such as "click here" particularly in link anchors (the text between '' and ''). Search engines look for text, and so pages whose main purpose is to display images are more likely to be indexed and retrieved if they also contain some paragraphs of descriptive text. If images are used as links, they should be supplemented by ALT tags and meaningful text links.

Use persistent URLs. It is possible for dynamically generated pages to be found and indexed by Google, but they are often invisible. Static pages are best for SEO (and impose a lighter load on servers), but dynamic pages can be indexed if they have persistent URLs with no session identifiers. This point is discussed in more detail below.

Use meaningful and variable title tags. This is an important part of SEO, and is also the single most important element in data shoogling, so is described in detail below.

Include good keywords. SEO is important, but not important enough for collection managers to jettison long-established cataloguing principles and practices. Adding keywords is regarded as essential by those concerned solely with SEO.

However, data shoogling has different priorities, which require that keywords and subject terms only be added if directly relevant to the item concerned and in accordance with institutional policies for resource description and interoperability.

As well as assisting resource discovery, another advantage of SEO is to help make web sites compliant with accessibility standards and legislation.

Metadata cleaning

To be effective, metadata must be accurate and consistent. Data shoogling assumes a context where there is a concern for sound principles of collection management and efficiency of content maintenance. However, it is easy for inaccuracies and inconsistencies to creep into any large database or digital collection. It is, therefore, worth carrying out data checking and cleaning routines, to verify internal consistency, remove duplicates, rectify omissions, check adherence to standards and local conventions, and generally ensure the metadata is in good shape. The feasibility of this depends on the size of the collection, the flexibility of the database software, and the staff available to carry out the necessary checks and corrections. There are software tools available to help, but they can only assist informed content editing, not replace it. Even if full metadata cleaning is not possible, ensuring the accuracy of title, author, type and date fields is well worthwhile.

Metadata optimisation

The MARC 21 concise format for bibliographic data (MARC 21, 2003) defines over 180 metadata fields, most of which have several further subfield codes. The IEEE LOM draft standard for learning object metadata (IEEE Learning Standards Technology Committee, 2002) defines 77 metadata fields, along with detailed specifications for valuespaces and datatypes. The Dublin Core (2004) metadata standard, which was specifically designed to simplify description of electronic resources, defines a mere 15 fields (plus optional qualifiers). Yet in 2004, by far the most widely used program for finding internet resources was Google, which uses just one metadata field: the HTML title tag. The challenge of metadata optimisation is to retain the richness of resource description made possible by standards such as MARC, IEEE LOM and Dublin Core, but to get them to work with Google too. The obvious place to start is with the one field that Google does use: the title tag.

Use of the title tag. The importance of the HTML title tag may be widely known, but it is not necessarily acknowledged or acted upon, and its value is rarely mentioned in academic literature. MacDougall (2000) may be stating the obvious in pointing out:

> The naming of the title is important . . . and its creation should be treated to some extent as an indexing function by the author of the web document

but it is an obvious point worth emphasising.

Title tags are vital for three reasons: first, because the Google search algorithms give them significant weight; second, because users see title tag contents highlighted in their search results and have to browse numerous titles to identify items of relevance; and third, because title tags become the default names for bookmarks in web browsers. It, therefore, follows that anyone who wishes to encourage access to their collections has a vested interest in ensuring the use of accurate title tags, as in effect they serve as main entry points to items in distributed web-based collections.

This situation is reminiscent of the library card catalogue before computerisation. Main entries were created in which the author and date of publication were appended to the title, so that users of the card catalogue could quickly see the most important metadata in one place. For non-book items, the physical form of an item (called the general material designation) was also included, e.g. [sound recording].

This strategy can be adopted for the Google world by regarding the title tag as the single main entry point. There are no rules about the content of title tags in web pages, so a title can look like this:

<title>Tom Bell</title>

or this:

<title>Photograph of Tom Bell, 1914</title>

or this:

<title>Tom Bell, pioneer of the socialist movement on Clydeside</title>

Although the name Tom Bell alone is concise and accurate as an item title, the additional information is clearly useful to Google users faced with 26,900 titles matching a search for "Tom Bell". However, long titles might not be displayed in full (in 2004, Google displayed only the first 60 characters) so lengthy text does not help users scanning search results, although it does help with indexing and retrieval.

This simple example illustrates that the question of what to include in the title tag is more complex than one might think. All three titles above seem valid, and there are many other options, so how can content providers decide what text to include in title tags? A good way to answer this question is to turn to well-established cataloguing principles and standards. The *Anglo-American Cataloguing Rules* (AACR2, 2002) specify a standard for the syntax of the main entry point which was designed for card catalogues but can equally well be applied to the title tag. This could turn the above example into:

<title>Tom Bell, pioneer of Clydeside socialist movement [photograph], 1914</title>

Unlike books and articles, photographs often have no clearly defined title, so content managers and cataloguers can choose to include useful descriptive text in item titles.

Where an author's name is known and relevant, that can also be slotted into the title tag using AACR2 syntax, e.g.

<title>Fifteen thousand Glasgow tenants on strike [newspaper article] / P.J. Dollan, 1915</title>

The syntax here is consistent with AACR2, but the general material designation, or item type, deviates from the standard AACR2 wording "[electronic resource]" which is too vague to be useful to Google users. Including a meaningful item type in the title tag helps users with resource selection, although there is currently no standard taxonomy of item types.

Although it might appear to break basic metadata rules to put the author, date and item type alongside the title, this is simply a display format. The four fields are concatenated for this purpose but can easily be derived from an underlying database in which they are stored separately, in accordance with basic data management principles. This is important, as Google may not be dominant forever, so alternative output formats might be required in future. Even AACR2 is not set in stone, so the syntax might need to be modified to comply with the emerging AACR3 specification.

Another option for the title tag is to include the collection name as a prefix, e.g.

<title>Red Clydeside: What socialism really means [leaflet], 1906</title>

This provides useful context for Google users, but has disadvantages too. The collection name may be quite long and take the whole string beyond the 60-character title display length. Also, this creates redundancy when searching within the collection itself, where it is unhelpful for every search result to begin with the same prefix. This problem can be avoided if local searching is based on direct database querying (as opposed to Google-like harvesting and indexing), as the output can then be customised to produce whatever combination of fields is judged most useful to users.

Although AACR2 comprises a large and complex set of rules, a thorough grasp is not necessary for creating consistent, accurate and useful title tags that comply with its syntax for main entry points. Bowman (2003) provides a concise and readable introduction to AACR2 that provides sufficient guidance. By studying his examples, and following the strategy proposed above for the content of title tags, collection managers should be able to put their valuable metadata to more widespread use, thereby increasing the visibility and use of their collections. In other words, they should be able to optimise their metadata for Google users with no adverse affect on any other applications.

Descriptions and subtitles. Constructing a title tag from the content of four metadata fields (title, type, author, date) is the most important step in metadata optimisation, but it is not the only option. Three other commonly-used metadata fields that have high value for information retrieval are the description (or abstract or summary), the subtitle (where applicable) and the subject terms. These are more difficult to optimise for Google, but there are possibilities.

Although HTML does have a <meta> tag, its value is currently negligible owing to misuse. Even if web pages do include a <meta name = "description"...> tag, this will not be used by Google, which instead extracts its own snippets (the official term) from the full text of documents to serve as page summaries. The value of Google snippets as descriptions is highly variable. One way to improve them is to include the content of the metadata description field as visible text near the start of the page; perhaps beneath the title and author name. This should ensure that when a user search term matches a word in the title or author, the Google snippet will include the item description (or at least the start of it). In other words, the arrangement of text on the page is optimised to generate meaningful snippets for matches on significant words. This might result in the top of each page looking like the start of a catalogue record, but that could have a dual purpose, in making the most important metadata explicit for users as well as optimising it for Google.

Where items have subtitles, a similar strategy can be adopted, so that subtitles appear in Google snippets. Alternatively, a subtitle can be appended to the main title in the title tag, to enhance retrieval, though this may impede display of item type and author in the search results. Either choice can be justified, depending on the nature of the collection and the average length of titles.

Subject terms. One of the drawbacks of using Google is the lack of vocabulary control, which is an inevitable result of its post-coordinate indexing. Where collections have been catalogued using a controlled subject vocabulary, at least two shoogling techniques can be used to help extract more value from this precision of terminology. One option is to ensure that the subject terms appear in the web page. This may sound

obvious but is not always the case. A good position would be above the description, so that when a search matches a subject term, the description is used for the Google snippet, for example:

Title: Fifteen thousand Glasgow tenants on strike
Author: P.J. Dollan
Subject: rent strikes
Description: Article about the Glasgow rent strikes published in *Forward* newspaper in 1915.

In this particular case the subject term appears in the description too, but that is not always true.

A more complex option is to use the subject terms as explicit indexes to a collection (or to a set of collections), so that each subject term has its own web page, comprising a list of links to all relevant items. The beauty of this option is that it becomes perfectly justifiable to use the subject term in the title tag. The disadvantage is that it can introduce redundancy if the item titles appear in both the subject index page and in the pages of the digital resource itself.

The technique of promoting subject access by creating web pages for each subject term is an example of the practice of metadata exporting, which is the fourth main component of data shoogling.

Metadata exporting

The procedure for metadata exporting will vary from collection to collection, but will typically involve running a program that carries out a repetitive series of carefully formulated database queries. There are several possible scenarios; four examples are given below:

(1) A collection of digital content that already uses static pages, e.g. a small digital library. In this case data shoogling should simply require recreating the collection from its source database to ensure that its metadata is optimised for Google, as described above, with particular attention given to the content and syntax of the title tag.

(2) A collection of digital content that uses dynamically generated pages, e.g. a journal archive. In this case data shoogling may entail creating a large number of static pages, each with optimised metadata. This could be a substantial task for a large collection, if the pages are designed as part of a well-structured web site with suitable navigation. Alternatively, it may be possible to adjust the publication process to generate dynamic pages that can be indexed by Google, with replicable and relatively short URLs, and no session identifiers. These are sometimes called "Search Engine Safe URLs". An illustration of how to adjust the URLs of dynamically generated pages using one particular technology (IIS and ColdFusion) is given at cfhub.com/contributions/SES (CFHub, 2004).

(3) A metadata collection describing digital resources, e.g. an annotated directory of web sites. In this case the rationale and method of data shoogling requires careful thought, as in theory the digital content could render the directory redundant, if it were accompanied by fully optimised metadata. In practice this is very unlikely, and the aggregation of metadata from distributed but related resources in a single place can be useful for searching as well as browsing.

The aim of data shoogling in such cases (as ever) should be to maximise the value of the carefully assembled metadata for Google users as well as for those who reach the directory web site.

(4) A metadata collection describing physical resources, e.g. a library catalogue. In this case the collection manager should take a strategic view concerning which parts of the catalogue, if any, require shoogling. The process is most likely to be justified for national institutions or special collections in which the items described are not commonly available. Organising exported pages by subject term may be the best approach, as there is no associated digital resource to create redundancy.

If a series of static pages is generated, it is important that a procedure is set-up to recreate them as often as necessary. For some historical collections the static pages may never need updating, whereas the static pages of volatile collections may require automatic regeneration every week or even every night. If items are sometimes deleted from a collection then the process for generating the static pages needs to incorporate a deletion routine as part of each updating cycle. This can be automated by referring to the file creation date.

The mechanics of exporting optimised metadata are not complex, but the process does require careful thought and periodic review. It is the final stage in the data shoogling process, and can be highly effective, as the examples below illustrate.

The effectiveness of optimisation and precision. Development of the Glasgow Digital Library (Dawson, 2004) has shown that the extra work involved in putting shoogling theory into practice is manageable for a relatively small library (around 10,000 items) with few resources. In fact, the ideal position is one where no extra work is required at all, as the procedures for content updating automatically generate accurate, optimised and exported metadata. This can be remarkably effective. For example, on 26 August 2004 a new electronic book called *The Old Country Houses of the Old Glasgow Gentry* was published via the Glasgow Digital Library (gdl.cdlr.strath.ac.uk). By 7 September 2004, searching Google for "old country houses" (without quotation marks and without specifying the key word "Glasgow"), produced about 5,770,000 hits, with the home page of *The Old Country Houses of the Old Glasgow Gentry* ranked number 1, and the preface of the same book at number 2. This high ranking was further evidence of a trend that had previously been noticed, whereby GDL content would appear high in Google search results shortly after publication, and before there were any external links to it. This contradicts the received wisdom that Google rankings are primarily based on the number of linking web pages. In fact, searching Google to check how many sites linked to the newly-published ebook, with the syntax:

link:gdl.cdlr.strath.ac.uk/smihou/

produced no matches, even when this collection was top of the search results for "old country houses". This result proves that external links are not essential to achieving a high ranking in Google (although they may help). The high ranking was, therefore, attributed to the optimisation described above, notably the relatively simple search-engine-friendly design, the high content-to-markup ratio, the use of meaningful text in link anchors for internal navigation, and the use of precise title tags in every page.

In the case of the Glasgow Digital Library, this optimisation was not the result of deliberate attempts to achieve high placements in Google search results: it arose from the use of good cataloguing practice and simple page design. The understanding of shoogling theory described above emerged from an investigation into why GDL content was appearing so high in Google rankings so soon after publication. This investigation was also prompted by the realisation that almost all users who contacted the GDL with enquiries and feedback did so after discovering GDL content via a Google search for a specific topic, having no previous knowledge of the library's existence. It is certainly part of GDL philosophy to make its content widely accessible, easily discoverable, and compliant with relevant international standards, but even so the extent of its prominence in Google was rather startling for a relatively small and unpublicised library, and helped to initiate the current line of research.

A further example illustrates how a controlled subject vocabulary can be effective, even though this is usually seen as irrelevant to Google. One of the research strands of the GDL is the application of controlled subject terms at a high degree of granularity. Library of Congress Subject Headings (LCSH) have been applied to individual images and to chapters and sections of some ebooks. This has enabled development of a browsable subject interface that links to relevant items from different collections, as an alternative to the collection-centred view. The subject interface currently uses a static page for each subject, with the subject term in the title tag, which means that the subject metadata is highly optimised for Google. The value of this can be demonstrated by searching Google for a multi-word LCSH term that applies to a GDL item. For example, in September 2004 a phrase search for "service industries workers" produced 872 matches, with the relevant GDL page as the top item, while searching for the same three words with no quotes produced 4.8 million matches, with the GDL page as number 6. Numerous similar GDL pages were just as highly placed, e.g. searching for "industrial equipment in art" produced 3.7 million matches, with the relevant GDL page as the fifth item.

Ideally these GDL pages should not be that highly ranked, as the relevant content is small and specific to the history of Glasgow. The high ranking is due to routine metadata optimisation, not external links or quantity of content. If other sites that have more extensive content and some external links also used precise vocabularies and optimised metadata then the GDL pages should slip down the rankings (unless subject searches were qualified by location, e.g. "service industries workers Glasgow"). This would not be any cause for concern. The achievement of high rankings in search results for libraries such as the GDL is not an end in itself but simply a means to help searchers find useful and relevant content.

While few casual users will use LCSH terms deliberately, they can assist users who happen to choose matching words in their search term, even if not the precise term. They are also of value to subject specialists. Google is often criticised because searching for common words produces such a huge number of matches, but this is inevitable for a service that indexes several billion web pages. Fortunately, the richness and variety of language means that multi-word searches can be very effective in finding a small number of relevant pages. The use of a precise subject vocabulary by content providers can undoubtedly enhance that effectiveness, even if users are not aware of its use.

Shoogling in detail

Large publishers have the financial incentive and staff resources to make sure that their data shoogling is effective, but it can be difficult to find sufficient detail about their procedures to permit emulation. Evidence from the Glasgow Digital Library shows that the process does not require complex technical infrastructure or expensive software. The GDL does use a variety of databases and search tools, in order to fulfil its role as a testbed for methodologies as well as a user service, but the retrieval results reported above were achieved using common Microsoft desktop software: Word, Access and Visual Basic (VB). The next few paragraphs illustrate some of the methods used within the GDL that have helped make it so accessible to Google, in order to demonstrate their relative simplicity and to assist other potential shooglers. With larger collections the processes would be more complex but the basic principles would remain the same.

Image and caption collections

All captions and metadata are held in Access databases (as required by some funding bodies), with a separate database for each collection. Images are held in separate files (TIFF for preservation, JPG for web access), and are related to the metadata by the image file name matching the ItemId field in the database. An additional GDL database holds all the XHTML markup required to generate the digital library web pages, as well as the syntax for metadata formats that are required for other purposes, e.g. MARC 21 and Dublin Core. Storing the markup in a database rather than embedding it in templates or program code makes it simpler to modify and ensures consistency. A VB program reads the markup from this additional database and stores it in a VB collection (like an array but with meaningful subscript names), then reads the content from the collection database, one record at a time, via a simple SQL statement such as:

```
SELECT * FROM Records ORDER BY EarliestDate, Title
```

The program then simply outputs the metadata and content embedded in the relevant markup. For example, the VB statement:

```
Print #Filenum, X("TitleStart") & PageTitle & X("TitleEnd")
```

writes out a title tag using a record from the collection database, where,

#Filenum refers to the XHTML output file,
X("TitleStart") refers to the open title tag i.e. <title>
PageTitle refers to the text to appear in the title tag
X("TitleEnd") refers to the close title tag, i.e. </title>

A subsequent statement:

```
Print #Filenum, X("Heading1Start") & ItemTitle & X("Heading1End")
```

ensures that the title is visible on the web page as well as appearing in the title tag. If the PageTitle variable is equal to the ItemTitle variable, then the result of these two statements will be almost identical, for example:

```
<title>Socialism : what it is and what it means</title>
```

and:

```
<h1>Socialism : what it is and what it means</h1>
```

However, if the title tag is to be optimised as described earlier, then PageTitle needs to be constructed from variables representing different database fields. For example:

```
PageTitle = ItemTitle & "[" & ItemType & "], " & EarliestDate
Print #Filenum, X("TitleStart") & PageTitle & X("TitleEnd")
```

would produce:

```
<title>Socialism : what it is and what it means [booklet], 1906</title>
```

while:

```
If ItemAuthor <> Then _
PageTitle = ItemTitle & " [" & ItemType & "] / " & ItemAuthor & "," &
EarliestDate
```

would add the author name after the / character, to give:

```
<title>Socialism : what it is and what it means [booklet] / R. Wells,
1906</title>
```

This example from the GDL Red Clydeside collection is shown in detail to illustrate just how straightforward the process of optimising titles can be. It is of course more complex to generate entire web sites, with labelling, navigation, links, wrapper pages, contents and indexes pages, and so on, but the principles are similar. Although this example is from a VB6 application, the same code could be used in an Access 2000 module. Use of VB allows for a more sophisticated user interface, and has advantages for accessing other kinds of databases (e.g. SQL Server) and for scheduling automated processes, but the same results can be achieved just using Access.

Electronic books

When creating ebooks, pages are digitised using a desktop scanner or digital camera, text is converted to machine readable form and then stored in a Word document for proof reading, with a separate Word document for each ebook. The document is then formatted to reflect the structure of the original book, using inbuilt Word styles (Title, Normal, Heading 1, etc.) and user-defined styles. The document is converted to a single XHTML file using a Word macro (not the Word "save as web page" option, which produces pages cluttered with formatting markup that is bad for SEO) so that only text, image links and style names are saved: "Heading 1" style becomes "<h1>" "Normal" style becomes "<p>", etc. This XHTML file is then read into an Access database, using an Access module that parses the XHTML and stores it in a Records table, with each section (defined by <h1> or <h2> headings) stored as a separate record. The section heading is held in one field and the entire body of the section in another field, along with its embedded markup. Thereafter, the output process is as described above for image and caption collections, except that an Access module rather than a VB program is used to generate the web pages. The main difference is in the

content of the title tags. As each web page is a section of a book, the section title needs to be contextualised to include the book title ("Introduction" is not a useful web page title), so that the relevant statement looks something like this:

```
PageTitle = TitlePrefix & SectionTitle & " [" & XA("ItemType") & "] / " &
XA("ItemAuthor") & "," & XA("PubDate")
```

where TitlePrefix is the title of the whole book. Unlike the earlier example, the item type, author and date are the same for every section of the book, so they are stored as constants (in a VB collection called XA) rather than as variables. From a local perspective, it seems redundant to include the same book title, author name and date in the title of every section of an ebook (some GDL ebooks have over 100 sections). However, this apparent redundancy is useful to Google users, as each ebook section is effectively an independent web page that may well be retrieved in isolation from the rest of the book, and could be used as an individual learning object, along with objects from quite different sources.

Although the process described here may sound complicated, once it has been set-up it only takes two minutes to convert a 500-page book from a single Word document to an electronic book containing a series of interlinked web pages, each with precise and optimised title tags. Once the book has been finalised there is no need to keep converting it from Word and loading it into a database, so it is possible to adjust optimisation parameters, if necessary, and recreate the entire ebook within a few seconds, ready for the next visit from Google robots.

Subject indexes

Browsable subject indexes to digital libraries or metadata collections can be created by following similar methods to those outlined above. Typically, each distinct term is retrieved in turn from a Subjects table, then a series of repetitive SQL queries retrieves the records matching each term from one or more collections, and outputs the matching titles (along with authors, descriptions and other fields if appropriate). The value and feasibility of doing this depends on the nature of the collection and the precision of the subject cataloguing. Evidence from the Glasgow Digital Library indicates it is highly effective in assisting retrieval of relevant resources via multi-word subject terms.

Further optimisation issues

The methods outlined above are simple and extremely effective, but are illustrative not prescriptive. More sophisticated solutions will be required for larger collections, but the principles will be the same. There is also scope for further research and development of the optimisation process, particularly in the area of semantics rather than syntax. For example, there is no standard vocabulary for digital item types, to replace the inadequate AACR2 standard "[electronic resource]". Although item types such as [photograph], [article] and [ebook chapter] are used within the GDL, the lack of any widely accepted standard limits their value as search filters, although they are still useful in results listings.

It is also not clear how to optimise long titles for both retrieval and results display. While it may be tempting to abbreviate titles by omitting common words that have little value for searching, this would offend those committed to following established cataloguing procedures. Subtitles present a further dilemma. Current practice within

the GDL is to include a subtitle in the title tag of the home page of an ebook, but not in the title tag of individual sections. This choice is largely determined by Google's 60-character title display limit, and may impede retrieval for a search term that matches one word in the subtitle and another in the section title or full text.

Another optimisation issue is to prevent duplication by stopping Google from indexing pages that only contain links to other local pages, e.g. contents pages and index pages. Such indexing is redundant as long as the linked pages, containing the actual text and titles, are indexed. The well-established solution is to include the following meta tag in the head element of the page:

```
<meta name="robots" content="noindex, follow">
```

This instructs search engines to follow the links from that page but not index its content. A statement to generate this tag can easily be inserted into the program or module that generates the web page, but it is important that it is inserted selectively and not in all pages. Some of the large GDL ebooks have contents pages for each chapter, as well as an overall ebook contents page, so the statement needs to be inserted in each of these contents pages but not in the text pages. The program, therefore, requires a means of automatically identifying when it is writing a contents page.

Beyond Google

The current popularity and effectiveness of Google may not persist. Already there are signs of a Google backlash (Bradley, 2004). The strategies and techniques outlined in this paper are intended to help information providers reach out to Google users, but it is important to remain flexible, and to establish procedures that will allow output and optimisation for different applications in future. It is also worth investigating enhancements and alternatives. The "common information environment" referred to earlier is a sound idea in principle, but its focus on communication protocols and metadata syntax make it largely irrelevant to many potential users. Making content more accessible to search engines is a good start, but would be much more effective if supported by a widely-used set of semantic standards, covering issues such as persistent digital object identifiers, subject classifications, item type taxonomies, conventions for titles of digital resources, guidelines for abstracts and descriptions, and co-ordinated use of authority files. All these are compatible with use of Google at some level, and would be invaluable components of a common information environment. They could even lead to overdue improvements to existing information standards such as LCSH and AACR2, or the development of controlled enhancements such as UK subject headings with simple, easily-applied syntax, and mappings to LCSH.

The need for semantic interoperability has long been recognised (Berners-Lee *et al.*, 2001), but it presents more problems than syntactic interoperability. While there are mappings between metadata schemes such as MARC 21 and Dublin Core, mapping between subject taxonomies is much more difficult (McCulloch, 2004). Yet a successful information retrieval system also requires a third concept, that of "user interoperability" whereby information structures, such as metadata schemes, have to usefully interact with those who are supposed to benefit from them, that is, information seekers. Perhaps the difficulty of dealing with semantics explains why recent metadata development seems to have been largely concerned with syntax. Yet ultimately it is the actual words used in the metadata that matter – the titles, subject terms and

descriptions, not the containers. And even the words only matter if people can find them. This paper has sought to document a route by which the metadata community can re-establish contact with the wider community, and in particular Google users – a very large community indeed.

Conclusion

The evidence described above shows that commercial publishers are attaching great importance to getting their content indexed by Google. They are willing to make significant changes to their procedures in order to increase access to their content and thereby generate more income. Complementary evidence from the Glasgow Digital Library demonstrates how some simple adjustments to web pages, and title tags in particular, can be extremely effective in ensuring that content can easily be discovered by Google users. There is substantial scope for universities, libraries, museums, government departments and other institutions to take similar steps. Compared to the vast effort and expenditure that has gone into creating and cataloguing the resources held by these organisations, the steps required to make their metadata more functional and their digital content more accessible are trivial. Millions of users may well be interested in these institutional resources but would never think to visit the institutional web sites; they simply pursue their interests via search engines. Rather than deploring this reality, institutions can reach out to users by ensuring that their content is easily located via the primary method of resource discovery.

References

AACR2 (2002), *Anglo-American Cataloguing Rules*, 2nd ed., American Library Association/Canadian Library Association/CILIP, Chicago, IL.

Aldrich, S. (2004), "Using search engines to find new customers at the American Institute of Physics: scientific publisher combines findability with 'buy by the piece' to grow its customer base", Strategic Research Service (Patricia Seybold Group), 4 March, available at: www.aipservices.org/newsroom/white_papers/pdf/PhysicsFinder.pdf (accessed 26 October 2004).

Arms, C.R. and Arms, W.Y. (2004), "Mixed content and mixed metadata: information discovery in a messy world", in Hillman, D.I. and Westbrooks, E.L. (Eds), *Metadata in Practice*, American Library Association, Chicago, IL.

Bains, S. (2004), "Breaking through the walls: developing the virtual National Library of Scotland", paper presented at Electric Connections Conference, Stirling.

Banks, M.A. (2005), "The excitement of Google Scholar, the worry of Google Print", *Biomedical Digital Libraries*, Vol. 2 No. 2, available at: www.bio-diglib.com/content/2/1/2 (accessed 5 July 2005).

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

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific American*, Vol. 284 No. 5.

Bird, D. and Simons, G. (2003), "Extending Dublin Core metadata to support the description and discovery of language resources", *Computers and the Humanities*, Vol. 37 No. 4, pp. 375-88.

Bowman, J.H. (2003), *Essential Cataloguing*, Facet Publishing, London.

Bradley, P. (2004), "Search engines: the Google backlash", *Ariadne*, No. 39, available at: www.ariadne.ac.uk/issue39/search-engines (accessed 26 October 2004).

Brandchannel (2004), *Brand of the Year Survey Results 2003*, available at: brandchannel.com/features_effect.asp?pf_id = 195 (accessed 26 October 2004).

Calishain, T. (2003), "Why try to out-Google Google?", O'Reilly Web DevCenter, available at: www.oreillynet.com/pub/a/javascript/2003/05/16/googlehacks.html (accessed 26 October 2004).

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

CFHub (2004), *Search Engine Safe URLs*, available at: cfhub.com/contributions/SES/ (accessed 26 October 2004).

CrossRef (2004), "Press release: CrossRef Search Pilot now includes 29 publishers, 3.4 million research articles", available at: www.crossref.org/01company/pr/press092104.html (accessed 26 October 2004).

Dawson, A. (2004), "Building a digital library in 80 days: the Glasgow experience", in Andrews, J. and Law, D. (Eds), *Digital Libraries: Policy, Planning and Practice*, Ashgate, Aldershot.

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.

Dowling, T. (2004), "On Monday, Google went down", *The Guardian*, 28 July.

DP9 (2001), "DP9: an OAI service provider for web crawlers", available at: dlib.cs.odu.edu/dp9 (accessed 26 October 2004).

Dublin Core (2004), available at: dublincore.org (accessed 26 October 2004).

Hunter, P. and Guy, M. (2004), "Metadata for harvesting: the open archives initiative and how to find things on the web", *Electronic Library*, Vol. 22 No. 2, pp. 168-74.

IEEE (2004), "Google users flock to IEEE XPLORE", *What's New @ IEEE for Students*, Vol. 6 No. 3.

IEEE Learning Standards Technology Committee (2002), *1484.12.1-2002 IEEE Standard for Learning Object Metadata*, IEEE, New York, NY.

Ingenta (2004), *Ingenta Partners with Google to Enable Full Text Indexing*, available at: www.ingenta.com/isis/general/Jsp/?target = /about_ingenta/press_releases/google.jsp (accessed 26 October 2004).

Inger, S. (2004), "Google vs traditional information services: a comparison of search results", National Federation of Abstracting and Indexing Services (NFAIS), 22 February, available at: www.scholinfo.com/presentations/GoogleversusTraitionalInformationServices.pdf (accessed 26 October 2004).

Jacso, P. (2004), "Péter's digital reference shelf – Google Scholar", December, available at: googlescholar.notlong.com (accessed 5 July 2005).

Jacso, P. (2005a), "Péter's digital reference shelf – Google Scholar (Redux)", June, available at: www.galegroup.com/servlet/HTMLFileServlet?imprint = 9999&region = 7&fileName = reference/archive/200506/google.html (accessed 5 July 2005).

Jacso, P. (2005b), "Google Scholar: the pros and cons", *Online Information Review*, Vol. 29 No. 2, pp. 208-14.

JISC (2004), *A Vision for the Future: Towards a Common Information Environment*, Joint Information Systems Committee, London, available at: www.jisc.ac.uk/uploaded_documents/vision.pdf (accessed 26 October 2004).

Jordan, J. (2004), "From the President: extending WorldCat, raising the visibility of libraries", *OCLC Newsletter*, No. 263, available at: www.oclc.org/news/publications/newsletters/oclc/2004/263/letter.html (accessed 26 October 2004).

Kent, P. (2004), *Search Engine Optimization for Dummies*, Wiley, Indianapolis, IN.

Lynch, C. (2001), "When documents deceive: trust and provenance as new factors for information retrieval in a tangled web", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 1, pp. 12-17.

McCulloch, E. (2004), "Multiple terminologies: an obstacle to information retrieval", *Library Review*, Vol. 53 No. 6, pp. 297-300.

MacDougall, S. (2000), "Signposts on the information superhighway: indexes and access", *Journal of Internet Cataloging*, Vol. 2 Nos 3/4, pp. 61-79.

Mangold, T. (2004), "We can't live without Google", *Evening Standard*, 28 July.

MARC 21 (2003), *MARC 21 Concise Format for Bibliographic Data*, Library of Congress, Network Development and MARC Standards Office, Washington, DC, available at: www.loc.gov/marc/bibliographic/ecbdhome.html (accessed 26 October 2004).

Marchiori, M. (1998), "The limits of web metadata, and beyond", *Computer Networks and ISDN Systems*, Vol. 30 Nos 1/7, pp. 1-9.

Miller, P. (2004), "Towards the digital aquifer: introducing the common information environment", *Ariadne*, No. 39, available at: www.ariadne.ac.uk/issue39/miller/ (accessed 26 October 2004).

OCLC (2003), *2003 Environmental Scan*, OCLC, Dublin, OH, available at: www.oclc.org/membership/escan/social/satisfaction.htm (accessed 26 October 2004).

OCLC (2004), *Open WorldCat Program*, OCLC, Dublin, OH, available at: www.oclc.org/worldcat/open/default.htm (accessed 26 October 2004).

Open Access Now (2004), "Google and DSpace launch joint project", 10 May, available at: www.biomedcentral.com/openaccess/news/?issue = 16 (accessed 26 October 2004).

Scholarly Information Strategies Ltd (2003), *How to Get Premium Content Indexed by Google and Other Web Crawlers*, Scholarly Information Strategies Ltd, Didcot, Oxon.

Sherman, C. and Price, G. (2001), *The Invisible Web: Uncovering Information Sources Search Engines Can't See*, Cyber Age Books, Medford, NJ.

Sullivan, D. (2002), "Search engine features for webmasters", available at: searchenginewatch.com/webmasters/article.php/2167891 (accessed 26 October 2004).

Svenonius, E. (2000), *The Intellectual Foundation of Information Organization*, MIT Press, Cambridge, MA.

Tenopir, C. (2004), "Is Google the competition?", *Library Journal*, Vol. 129 No. 6, p. 30, available at: www.libraryjournal.com/article/CA405423 (accessed 26 October 2004).

Thomas, C.F. and Griffin, L.S. (1998), "Who will create the metadata for the internet?", *First Monday*, Vol. 3 No. 12, available at: www.firstmonday.dk/issues/issue3_12/thomas/index.html (accessed 26 October 2004).

Urquhart, E. *et al.*, (2003), "Uptake and use of electronic information services: trends in UK higher education from the JUSTEIS project", *Program: Electronic Library and Information Systems*, Vol. 37 No. 3, pp. 168-80.

Wallis, J. (2003), "Information-saturated yet ignorant: information mediation as social empowerment in the knowledge economy", *Library Review*, Vol. 52 No. 8, pp. 369-72.

Corresponding author
Alan Dawson can be contacted at: alan.dawson@strath.ac.uk