Meetings Reports

Metasearching - Better Searching?

British Computer Society Electronic Publishing Specialist Group one-day seminar
Said Business School, Oxford, Thursday 22 July 2004

This was a lively meeting that provided good insights on a key topic for journal and for reference publishers. 'Metasearching' is the use of tools to provide better search and retrieval, across multiple databases, services, platforms, protocols, and vendors. From the rapidly increasing number of vendor solutions on the market, it is clearly a growth sector.

Andy Powell of UKOLN provided a clear overview of metasearching, based on the JISC-funded work UKOLN carried, out, for example towards the JISC Information Environment, a set of standards for seamless access across multiple resources. He pointed out the problems of searching across incompatible Web resources: they have different user interfaces; everything is in HTML, so it is difficult to merge, copy and paste the information you have retrieved, and you don't necessarily get access to the resource itself. Metasearching has two approaches: cross-searching, which is real-time searching across several databases, and harvesting, batch processing of pages into a local database. The latter provides a quicker result for the user. The JISC Research Discovery Network is a nightly harvest of several smaller gateways. Searching across databases typically uses the Z39.50 protocol; harvesting is typically done using the OAI (Open Archives Initiative) protocol. Metasearching websites involves the wonderfully named 'web scraping' of information from the site, which, as its name suggests, is an imperfect solution. As soon as the web page is updated, the scraping may no longer work.

Increasingly, solutions are appearing to the problem that provide better standards for metadata, for example the NISO Metasearch Initiative, and repositories of metadata such as the eGov metadata standard and the LOM metadata profile.

Other presentations showed how these approaches are being implemented in specific sectors. John Davidson of Sentient Learning described how Sentient tools can be used to make a university reading list shared across all the stakeholders, including the library, students, and local bookshops, thereby enabling better use of existing resources. James Culling of ExLibris described widely used tools such as SFX, an Open-URL compliant link server, providing context-sensitive linking for journals, and MetaLib, providing access to an institution's e-resources, whether local or remote.

Hilary Ollerenshaw described a recent initiative from the North Bristol NHS Trust. Entitled Knowledge4Health, it is a portal of internal and external resources, providing a single point of access for evidence- and Trust-based patient information, including local, regional, and national services.

Chris Knowles of Magus Research described a rather different problem: searching across an organised, structural and managed body of information. Merchant banks, for example, frequently want to search across several structured but incompatible resources. He described tools to facilitate such a search from a single search screen; a similar approach is now being implemented for law firms.

Finally, Martin Kelly of the Institute of Physics Publishing talked about initiatives to improve searching across large-scale scientific databases. These included using a proprietary search tool, Vivisimo, which groups search results into clusters, and compared it to Verity. Vivisimo worked well with unstructured data, while Verity was better where a taxonomy already exists.

Michael Upshall
Consultant advising publishers
on electronic product development and dissemination
michael@consultmu.co.uk

Book Reviews

Metadata for information management and retrieval
David Haynes. London: Facet Publishing, 2004
ISBN 1-85604-489-0 xiv, 186pp. £39.95

It's over eight years since I last worked full-time as a librarian. Though I do still return to the fray occasionally, temping or doing contract work, my primary occupation now is as a freelance indexer – mostly of books or other print materials, but I've also ventured into web site indexing, dabbled in indexing of other electronic texts, and generally tried to stay on the track of emerging ideas and new developments in ways of getting information from electronic resources.

Anyone who has followed a similar trail will know that it's fairly easy to pick up a basic understanding of metadata as a tool for resource description and

subject indexing. Going beyond that, though, in breadth or in depth, can be time-consuming. In my case, there were certainly plenty of gaps in the understanding, and some near-bottomless chasms of ignorance about metadata's other uses, in and outwith the LIS community.

David Haynes' book offers plenty of material to help fill those gaps. Its aims are summarised as 'describing recent progress in metadata standards and applications and focusing on the concepts behind metadata'. The book's target audience is 'information professionals who want to develop their knowledge and skills in order to manage metadata effectively, and managers who are faced with strategic decisions about adoption of IT applications that use metadata.'

Haynes first looks at the historical background of the term, and concepts associated with it. He then puts forward his own approach: in place of the usual 'data about data', he defines metadata as 'data that describes the content, format or attributes of a data record or information resource', applicable to structured or unstructured information, in print or electronic form, and stored either in the resource or in a separate database. Subsequently, he outlines a five-point model of metadata's purposes: resource description; information retrieval; management of information resources; documenting ownership and authenticity of digital resources; and interoperability.

Two chapters then deal with metadata in general, the first looking at ways of defining, expressing and storing it. Mark-up languages such as XML, with its tags, schemas and Document Type Definitions, embed metadata in the document it refers to; databases of metadata store it separately. Haynes then reviews various contexts in which metadata is used: word-processing, cataloguing, records management, e-commerce, and content management.

Next, he considers some of the data modelling systems underlying metadata standards. Almost all of those covered were new to me, apart from the Resource Description Framework (to which Haynes supplies the most comprehensible introduction I've yet come across). The others discussed are the ABC Ontology; Functional Requirements for Bibliographic Records; the Indecs metadata framework; and the Open Archival Information Standard. This chapter ends with a review of metadata standards themselves: Dublin Core and its extensions/derivatives; MARC; ISAD (International Standard Archival Description); ONIX; and standards for multimedia and educational resources.

Haynes then revisits his model of purposes. Each of the five has a chapter devoted to it, providing both an outline of current developments and an introduction to relevant concepts in each of the specific areas. These are followed by a chapter on managing metadata. The five-point model returns again in 'Looking forward – the future', this time as a framework for consideration of trends and possible developments in each of the five areas. Finally, he looks at trends in metadata management, speculates on the durability of metadata, and makes some predictions about what the future may hold for it and those who work with it.

The book is chock-a-block with information, on virtually every aspect of metadata. The one omission I can identify is faceted metadata classification; there is no mention of it or its associated language, XFML. Slightly surprising was the absence of any reference to topic maps in the discussion of metadata's durability and future development. Topic maps share at least two of metadata's purposes – resource description and information retrieval – but it isn't yet clear whether they will compete with metadata in those arenas, or complement it in some way. A discussion of the possibilities could have made interesting reading.

The text is well-organised and well-presented, each chapter beginning with an overview of its content and ending with a summary, plus a list of references and further reading. A glossary would have been a useful addition, particularly as the index doesn't immediately identify where definitions of terms can be found.

The index has other weaknesses, too. Its coverage of the text is patchy: 'artificial intelligence' refers only to page 98, for example, ignoring the interesting comment on page 176 about the role of AI in future systems; 'controlled vocabularies' refers to a major discussion on pages 152-4, but not to the point made on p. 138 about their importance to interoperability.

Nor can it be depended on to collocate those references which have been indexed. The entry for 'automatic indexing' has two references more than the inverted form 'indexing, automatic'; the page referenced at 'multimedia, intellectual property rights' does not appear in the entry for 'intellectual property rights'; and the same is true of the entries for 'semantic interoperability' and 'interoperability'. Cross-referencing is not always adequately done: there is nothing, for example, to show that information about 'controlled vocabularies' may also be found under headings such as 'thesauri', 'taxonomies', and 'synonyms'. Finally, the two

entries and one cross-reference covering the subject of records retention are badly in need of sorting out.

If all that sounds like nit-picking, it isn't; it's groundwork for a point I think important to make. LIS professionals know that a poor index limits the usability of a book for reference purposes. We also claim, as one of our core skills, expertise in organising information for retrieval. Yet here we have a professional publication, about a new technique in information management and retrieval, in which the traditional, built-in tool of retrieval hasn't been made to function as it should.

That the defect blights an otherwise excellent book is disappointing. What bothers me even more, though, is this: if we do not, in our own professional literature, demonstrate the ability to make old methods work, what kind of message does that send about our abilities to cope with the new?

Linda Sutherland
http://www.users.zetnet.co.uk/tarristi/
linda.sutherland@zetnet.co.uk

Freedom of information: a practical guide to implementing the Act
Kelvin Smith. Facet Publishing. London. 2004
ISBN 1-85604-517-X 200pp. £39.95

If you don't already know the significance of the date 1 January 2005, you will do when you have read this book. This is when the Freedom of Information (FOI) Act 2000 and the FOI (Scotland) 2002 Act become fully effective, and when public organisations will have to make their information available. This book is for all those in public authorities, including central government, non-departmental public bodies, all tiers of local authorities down to parish councils, higher and further education, schools, police authorities and the National Health Service. At last the UK has FOI legislation bringing it into line with other countries.

Kelvin Smith is Head of the Cataloguing and Accessioning Unit and Records Management Consultant in the Records Management Department of the National Archives, and has been immersed in the subject for 6 years. He has succeeded in his aim to write a practical guide to enable those involved in enacting FOI for their organisations without fear or worry. The book focuses on implementation from a user's point of view, and has chapters on:

Background to the Freedom of Information Act
The legislation
Exemptions
Publication schemes
Enforcement and appeal
Records management
Data protection, human rights and other legislation
Staffing and training
Getting ready for Freedom of Information

The Appendices have the full text of the Codes of Practice under sections 45 and 46 of the Act, Definitions, and Further help and guidance.

The Act is retrospective, and organisations need to:

Know what information they hold
Manage their information holdings effectively
Have in place the infrastructure for dealing with FOI requests
Meet challenging deadlines in responding to individual requests for information
Proactively disseminate information through a publication scheme
Set up arrangements to handle complaints and appeals
Ensure consistency in discharging their duties under the Act.

For information professionals used to the reference interview, it is of note that you cannot make enquiries as to why the information is being sought or what it will be used for. However, if a request is ambiguous, you can seek reasonable clarification. Information can also be refused if it is exempt under the Act as defined by 'public interest' (although this is likely to be tested in law), if the request is a repeat, or 'vexatious'. Exemptions are covered in detail; Smith thinks that it can be considered that information is not reasonably accessible if it is available only in digital form, but unlikely that the reverse will be upheld.

The use of publication schemes will save time and money by providing information which authorities publish as a matter of course. Most organisations publish these schemes on their websites, and screenshots of schemes from different types of organisation are included. Smith recommends that authorities look more closely at version control of documents, as earlier drafts of documents may be required to be released. He considers it likely that some information about public servants in connection with their work will be accessible, for example phone number, work address, role, responsibilities and grade. He suggests making available summaries of released information on the Internet (a database of requests submitted under Canada's Access to Information Act is available at http://faculty.maxwell.syr.edu/asroberts/foi).

Smith emphasises throughout that the legislation will only be as good as the quality of the records that are subject to its provisions. Responsibility for capturing, maintaining and ensuring access to records rests