Further Reading

Tony Hammond, Timo Hannay, and Ben Lund. (2004) The Role of RSS in Science Publishing. Syndication and Annotation on the Web. *D-Lib Magazine*. 10(12) [online] Available at http://www.dlib.org/dlib/december04/hammond/12hammond.html [Accessed 20 August 2005]

Intranets

Getting to Eureka!

Martin White

The last few years have seen many organizations investing in content management software to support web sites and in particular intranets. One of the motivations for this has been the vision of everyone in the organization being able to contribute to an intranet, though in reality this is rarely achieved, and in fact is probably not a desirable state of affairs. CMS applications are good at getting content into a repository, and at first glance look to be the way in which content can also be retrieved. Looking at the feature list of most CMS products there is usually a reference to 'powerful retrieval functionality', and in many cases this will be based on the Verity or Convera search suites.

Notice carefully the word 'based'! Many organizations make the assumption that in buying a particular CMS product they are also buying an enterprise search engine. However this is not the case. The search functionality in a CMS is there solely to enable authors to find content in the CMS repository so that it can be re-purposed. This content is in HTML/XML formats and so usually the CMS search functionality does not even allow for the searching of content in related Word and pdf files, though that does vary from vendor to vendor. Certainly the license will only cover the searching of content on the web server, and even then only for internal purposes.

From the perspective of a user they just want to find information, no matter on what server or application. Their requirement is for 'enterprise search' and that is where there is going to be a lot of interesting developments over the next 12–18 months. If you are not an information professional used to the search functionality of a service like Factiva then your expectations of search are based around Google. I've lost count of the number of times I've heard senior managers say that all they want is Google on their intranet. That is indeed possible, and I'll come back to Google later.

With the arrival of desk-top search many users are now getting a different perspective on search, such as the value of being able to highlight the search keywords even in a pdf. This has lifted the expectations of how effective a good search engine can be. Searching through unstructured text is not new – indeed the first text search application was launched in 1963! Another factor in raising the profile and value of search has been the dramatic decrease in the cost of enterprise search from not only smaller companies such as Isys Search http://www.isys-search.com but also Verity, which is offering Ultraseek free for collections of less than 25,000 documents http://www.verity.com

Another low cost approach is to use a search appliance. This is a 'black box' that is plugged into a standard equipment rack and can be implemented in less than an hour. This is the basis for the Google Search Appliance http://www.google.co.uk/enterprise/ , but there are other companies in this market, such as Thunderstone http://www.thunderstone.com/texis/site/pages/Home.html. The pricing of the Google Mini is currently £1995 for searching 100,000 documents. The relevance algorithm is similar to, but not the same as, the web version of Google.

The problem with implementing a search engine is that it is not until it is actually installed and indexing the document collection that its performance can be evaluated. There will need to be continuous tuning of the search engine as new document sets are added, and there is usually very considerable flexibility in presenting the results of search. Search can also be computing intensive, especially at the indexing stage. Querying is in fact much less intensive. More problems come when the user wants to download a number of large documents. This can result in some processor and network bandwidth problems.

Another misnomer is that search engines offering multiple language search enable the user to enter search terms in English and then the search engine will translate the terms into other languages. This is not the case unless a comprehensive look-up dictionary has been developed. The multiple language capability just means that searches can be carried out on a wide range of different languages using search terms in the language concerned.

As in the case of any piece of software a successful implementation starts with a clear understanding of user requirements, and this is where using personas to develop some search scenarios is very valuable, as these scenarios can then be used in performance and usability tests at a later stage.

Looking around for good sources of information on how to select a search engine I found that there was not much out there. Avi Rappoport's web site at http://www.searchtools.com is a good place to start, and has a pretty comprehensive list of vendors. CMSWatch has a report at $1325 that profiles the major search engine vendors and their products http://www.cmswatch.com and there is a good blog at http://www.unstruct.org/. So I spent much of the summer writing the Enterprise Search Guidebook, which was published in September by Freepint. Details at http://www.freepint.com/shop/report/

End of sales pitch. Happy searching.

Reference Management

Best Practice on Using Reference Software

Tracy Kent (t.k.kent@bham.ac.uk)

There are a number of ways in which reference management (no matter which package you choose) can be utilised for research and so here are a number of useful tips which you could consider:

To add value to individual references include some notes about the reference for later retrieval. You may wish to make a distinction between "quotes" that you can relocate or your [own notes] that could be placed in square brackets. This would help you keep track of materials and to re-purpose materials at a later date.

Alternatively you could use your own word-processing package to retain extensive notes and include your reference software ID number for cross referencing. This would mean that each reference would contain a file name and location in a field such as a label field or research notes option. You should ensure that the file format is consistent though to aid retrieval such as c:\documents\research\filename.doc Most packages allow you to open up additional file formats from within the references and to retain more than one link. We have yet to see how this might be affected by desk top search tools!

Several packages such as Endnote or Procite (for example) make reference to an image field which put a copy of the file into an internal software folder. Endnote for example has a data folder. This way you can only (often) locate one file per record while you can add multiple file links using other fields.