CASE STUDY

Deploying an effective search engine

A search engine is often the first method used to find a page, and yet, most users suffer frustration and failure. More still are put off by the complexity of the search engine, and the confusing manner in which the results are displayed.

A good search engine must:

- Be easy to use.
- Assist users to find the correct information.
- Display results in a meaningful way.
- Help authors to improve the site.

This case study looks at a recent project to deliver online documentation to frontline staff in a major organisation. When the system went live, there were over 3,000 pages of information; this is expected to grow to more than 15,000 over the coming year.

As frontline staff are always under time pressure, they need to find answers quickly and easily. An effective search engine is a key way of meeting this goal (good navigation and a 'back-of-the-book' index are also vital).

In this case study, we relate our experiences in what has been a very successful project. (Happy users are always a good sign.)

Summary This paper covers key areas in the design and deployment of an effective search engine:

- Selecting a search engine

  Before taking any action, determine your business and technical requirements. Once this is complete, research the currently available engines, and select to meet your needs.

- Designing the interface

  Take extra time and effort when designing your search pages. They should be clear, easy, and above all, *simple*. Don't bother with an 'advanced search' facility: your users won't understand it.

- Behind the scenes

  Make your search engine quietly work for the user, to correct their mistakes, and to help them find the right page.

Step Two Designs Pty Ltd
www.steptwo.com.au • contact@steptwo.com.au
Knowledge Management Consultancy • SGML, XML & HTML

Selecting a search engine

There are hundreds of search engines, ranging in cost from free to tens of thousands of dollars. Most of these are packed with features, from fuzzy matching to indexing Word documents, and much besides. Choosing from such a range is not easy.

Requirements

Take the time to identify and document your business and technical requirements, in consultation with your stakeholders. For our project, we listed over two pages of requirements, including:

- Easy to use.
- Must run on Unix.
- Ability to index PDF files and HTML pages.
- Straightforward to implement and maintain.

These requirements were generated by looking at our:

- End users — frontline staff, with very little time and limited computer experience
- Project goals — deliver a large and structured online resource
- Technical restrictions — must work on existing servers

Every project and business is different, however, and there is no 'one size fits all' search engine. It is only possible to be confident that you have selected the right search engine once you have a list of requirements to judge it against.

Our selection process

We started by creating a list of all the search engines on the market that fitted the criteria outline above. This was a long list. From this, we narrowed down the list to a single search engine in each of three categories: free, moderate-cost, and expensive.

With a short-list in hand, we downloaded evaluation copies of each engine. This proved to be a very useful exercise, which allowed us to assess whether actual capabilities matched marketing promises. It also gave us a sense of how 'polished' the product was, and how meaningful the documentation was.

Obviously the expensive search engine ($10,000+) did everything we could ever ask for. In the end, however, we went for the free search engine, as it met 95% of our requirements, and all of our key needs.

As it has turned out, having access to the source code for the search engine has allowed us to tweak some critical components to meet our needs, which would not have been possible with a commercial system.

Despite being free, the search engine also impressed us with its level of support, including an active developer mailing list. While this has worked for us, selecting a commercial search engine may provide better support in other situations.

And no, we're not going to name our selected tool (this is a whitepaper, not free advertising). Every situation is different: so research the products yourself, and draw your own conclusions.

Designing the interface

While much of the work of deploying a search engine goes on behind the scenes, the design of the user interface greatly influences how successful the system will be. While the interface design must be consistent with the rest of your online material, we recommend the following guidelines:

Search page

- Keep it simple

  There are two key elements on a search page: a field to enter the search terms, and a 'search' button. There is no reason to make the page any more complex than this.

- Provide hints

  A list of tips and examples on the main search page help users when they first use the search engine. This list should be written in plain English, and should cover the common issues and questions.

- No advanced searching

  Normal users have enough difficulty with search engines without confronting them with a complex set of 'advanced search' methods. Users want to quickly find a single page, and you must design your interface to meet this need.

  If the search engine doesn't do what the users would, *by default*, change it. Searching should always be as simple as 'enter the search terms and hit enter'.

  However, if your users are researchers, librarians or other *very* advanced users, you may want to consider providing an advanced search. Wait for your users to ask for it, though.

- Always 'and'

  Few users understand the concept of 'boolean operators'. Instead, they expect that when they type in three words, they will be given only those documents that contain *all three*. Furthermore, typing in more words should provide less hits, not more.

  The search engine must therefore default to 'and-ing' the words together. In fact, eliminate support for boolean operators all together, unless there is a clear case that they will be of value to your users.

- Place the cursor

  When the search page is opened, the cursor should already be in the search field (this is known as 'setting the focus'). This allows the user to simply type in their words, and hit enter. It's a small point, but it took only days for our users to specifically ask us for it.

Results pages

- Make it attractive

  A results page should encourage users, not frighten them off with tiny text, difficult layouts, and hard-to-read fonts. Remember that you are expecting users to spend time browsing through the list of results, so it is worth spending some extra time making the pages easy on the eye.

- Keep it simple

  There are only three things that you need to present for each hit: title (a hyperlink to the actual page), page summary and ranking. Nothing more.

  Why, for example, would the user want to know the size of the page in kilobytes? The less you say for each hit, the easier it is for the user to scan through the list and find the page they want.

- Use a star ranking

  Each hit should have a ranking showing how well it matches the search terms. Display this as a star ranking, from 1 to 5 stars. This is simple, familiar, understandable and attractive.

  (What does a ranking of '83%' really mean? 83% of what?)

- Make the description meaningful

  Ideally, each hit should provide a useful description of the page, obtained from the 'meta' tags within the page. If this information is not available, provide a brief extract, highlighting where the search terms are used.

  Ensure that the extract always shows some useful text, and not the standard headings on every page (how many listings have you seen that start with '[Home] [Contents] [Index] …'?). Most search engines allow you to indicate in each page where the actual text begins and ends. Take the time to implement this throughout your site.

Behind the scenes

Effort should be spent 'behind the scenes' to improve the effectiveness of your search engine. Most engines have capabilities that, when implemented carefully, will help users to find the pages they are looking for.

These features must operate transparently, so that the user is not even aware of their impact. They should simply find the search engine both easy to use and effective.

Fuzzy searching, stemming, and more

Our selected search engine provided a number of powerful searching capabilities:

- Fuzzy searching, or 'sounds-like'

  There were three closely-related options which were essentially designed to find terms which 'sounded like' those entered by the user. In this way, it becomes possible to handle spelling mistakes and other inconstencies.

  While our expectations were high for these features, we ended up turning them off. The simple reason was that the fuzzy matching was too 'loose', and matched too many other terms. This tended to produce an unmanageable large list of results.

  Instead, we chose to make more extensive use of a synonyms list (this is discussed later).

- Stemming

  This feature takes the terms entered by the user, and tries other combinations of endings. For example, searching for 'walks' would also find 'walk', 'walking', 'walked'.

  We found this to be very effective, and it eliminated differences in singular vs plural uses of terms in our pages.

There are a wide variety of other tools available in modern search engines, beyond those mentioned above. Evaluate each carefully, as it is our experience that just because a feature exists, it doesn't mean it will help *your* users. Explore and test (with real users).

One tip: we configured the search engine to display the resultant list of search terms (after being processed through the stemming and synonyms) in tiny text at the bottom of each results page. This helped us to check what the search engine was actually doing.

Weightings and rankings

The order in which results are displayed by a search engine is the product of a number of complex weighting and ranking factors behind the scenes. These vary from engine to engine, and are not well understood. They also have a big impact on how effective the search engine is. (After all, users expect the page they are looking for to be in the top ten, not hit number 198.)

To confess, although we were well aware of these issues, we failed to adequately research and document how *our* search engine worked in this area. Very quickly, though, our technical writers started asking us some pointed questions. Like: 'why does this page appear as number 2, while this other page (which seems the same) is number 50?'.

So we went back, and documented how the search engine sets its weightings and rankings. This was not an easy document to write, and it took us several edits to make it simple enough to understand.

We then sat down with our technical writers, and worked through the details. We ended up simplifying the weightings: while the engine incorporated several 'advanced' weightings, the value of these was outstripped by the confusion they caused.

The message here: understand how your search engine works, and modify it (if required) to meet your specific requirements. The key is to have the search engine work in a 'transparent' and understandable way. (After all, searching is not black magic.)

Usage statistics

In this project, we also generated a number of web-based usage statistics. These online reports allowed the technical writers and administrators to track the operation of the search engine, and the behaviour of the users.

These reports included:

- Full search usage

  A raw list of every search term entered by the users, the time and date, and the number of matches returned. We started with this, as it was quick and easy, and allowed our developers to check that the search engine was working.

- Search usage summary

  The top 100 terms entered by the users in each month, with the number of hits for each.

  This provided invaluable information on what the users were interested in, and highlighted several key areas where our documentation was inadequate. It also indicated sections which were being heavily used, and therefore were worth spending some extra time refining.

- Failed searches

  A list of all the searches that returned zero hits. These resulted either from user errors (such as spelling mistakes), or because there was simply nothing written about those topics. This is obviously a 'hit list' of areas for the technical writers to create some new material.

  It also highlighted the need for an enhanced synonym list (see below).

We cannot over-emphasise the importance of putting in place usage statistics. These are vital in monitoring the ongoing health of the search engine, as well as gathering key information on user behaviour.

With effective usage information, it becomes possible to further refine the behaviour of your search engine to be more effective in your specific situation. It also allows the content creators to know what information is being used, and how often.

Synonyms

One of the first things we discovered as a result of our usage statistics was that our users couldn't spell. (Actually, their biggest problem is lack of time: only 15 seconds between calls. Hurried typing leads to common mistakes.)

For example, 'quarterly' is a often-used search term. Unfortunately, it is just as often entered as 'quartly', 'quartely', quaterly' and 'quately'. All of these spelling errors were turning up in the 'failed searches' report.

Another issue became apparent: 'rego' and 'registration' were used interchangeably, both by the users and the technical writers.

These two problems were solved by a feature of the our chosen search engine: a synonyms list. We developed a web-based interface that allowed the technical writers to enter lists of words that are equivalent. The system then automatically updates the configuration files behind the scenes to make this work.

By developing an easy method to identify problem terms, and create new synonyms, the technical writers are able to effectively manage the ongoing use of the search engine. In this way, the list of 'failed searches' is kept manageably small.

(As a side note, the other way of dealing with mis-spellings is to have the search engine run a spell-checker across the search terms. While this can be very effective, our free search engine didn't have this capability.)

Summary

In brief, we learned the following lessons as a result of this project:

- Spend a lot of time identifying your needs, and researching the right search engine. Choosing the wrong search engine is a costly mistake that is not easy to rectify half way through a project.
- Keep the interface simple. The search page should have a field to type in and a 'search' button. Complex interfaces and advanced searches will confuse users: by default, your search engine should simply do what the users expect.
- Take the time to configure the intelligence 'under the hood'. The search engine should quietly assist the user to find the desired page (via synonyms, fuzzy searching, and so forth).
- Track the usage of your search engine, and use this to assess how well it is working. You should be gathering enough information to allow you to refine the engine's configuration to better meet user needs.

About the Author

James Robertson is the managing director of Step Two Designs, a knowledge management consultancy based in Sydney, Australia. James specialises in XML development, information management and systems design.

James can be reached via e-mail at: **jamesr@steptwo.com.au**

Content Management Requirements Toolkit

Fully-revised, and almost twice the size, the new version of the Content Management Requirements Toolkit captures the latest thinking in the content management industry. In addition to the expanded range of requirements, the Toolkit now provides a comprehensive guide to writing CMS scenarios, as well as a detailed overview of the whole selection process.

The single best resource for rapidly creating an effective tender, and selecting the right CMS.

Choose a content management system (CMS) is not easy. There are hundreds of products in the marketplace, all with highly-variable capabilities. In this rapidly-evolving environment, the challenge is to find the CMS that best matches your business needs.

The Content Management Requirements Toolkit provides a comprehensive starting point for identifying the business and technical requirements that will drive your selection process.

It contains **133 fully-developed requirements**, across five main categories.

- **content creation**
- **content management**
- **publishing**
- **presentation**
- **contract & business**

These ready to use requirements can be cut-and-pasted directly into your tender document. This will save you days of work, and is an invaluable checklist to ensure that no critical requirements are missed.

Clear and concise, these requirements have been distilled from real-world experience, and reflect actual business needs.

Version 2.0
Fully-downloadable package
123 pages, August 2004
Only US$ 550

For more information on the Requirements Toolkit, sample pages, and secure online ordering, visit:
www.steptwo.com.au

Step Two DESIGNS

CMSLIFECYCLE

Produced and published by **Step Two Designs**