



Search Engine Optimisation Basics for Government Agencies

Prepared for State Services Commission by Catalyst IT

Neil Bertram

May 11, 2007

Abstract

This document is intended as a guide for New Zealand government agencies on optimising their websites for modern search engines, describing briefly the current best practices for search engine optimisation (SEO) for major search engines.

1 Introduction

Over the past decade, the techniques required to rank highly in search engine result pages (SERPs) have changed dramatically. While in the earlier days of the web, simply including key search terms in meta tags and page titles ensured good placement, due to widespread search engine abuse and end-user dissatisfaction with search results, the major search engines have now implemented ranking algorithms that are focused on returning pages that they believe are what the user actually wants to read, rather than what the website owner wants to be listed for.

Some general guidelines on ensuring your websites are search engine friendly are listed below. Please note these are subject to constant change, but represent current best practices as of 2007.

2 Advice

2.1 Make sure your content is high quality and relevant

The single most important way to rank well in search engines is to provide content that is useful to a visitor, and won't disappoint them should they choose to click on that page's title in search results. This means keeping the content up-to-date and relevant, while providing some value that the visitor cannot get from another site on the same topic.

This is important because search engines are now designed to look for signs that a page is relevant and useful to people visiting it, which they do in part by tracking search users visiting the page from their search results, and either staying to read the content, or returning to the search results to try another link.

Of course not every page on a site will offer visitors high value, but most sites will feature some original articles and documents that provide content that isn't available elsewhere. If such pages or documents become popular, they will enhance the entire site's search engine profile.

Another benefit of having high quality and relevant content is that other sites will start linking to it. Most major search engines will count a link coming to a page on your site as a "vote" for that page, which is one of the most effective ways to improve your search engine exposure. Links from related but external major sites carry the most "vote", whereas links inside a site and between sites seen to be run by the same organisation carry less weight. Asking completely unrelated sites to link to your site is generally a bad idea however, as the search engine sees these as "forced" rather than "organic" links, and they will generally not help your rankings.

Finally, it is important that articles are well proof-read for common grammatical and spelling mistakes, as search engines will take note of any such errors and may impose a penalty on their perceived quality of the content. On top of this, if a key term is misspelled on the page, users searching for that term may be unable to locate the page at all.

2.2 Make sure your content is accessible and correctly describes itself

While it is common sense to make your site accessible to a wide audience of users, many overlook the fact that search engines do not see their site as most other users would. The search engine crawlers see your site's content in a similar manner to a blind user; without any multimedia content such as images, audio, Flash or video. It is therefore important that your

site usefully labels such content with text-based alternative captions. There are mechanisms in HTML such as the `title` and `alt` attributes to accomplish this.

The use of semantic HTML (or XHTML) with separated visual style and content can also be very helpful to search engines and non-visual human users alike. HTML includes facilities for conveying semantics, such as the various levels of header tags for headings (`H1`, `H2`, ...) You should ensure that the key topic for each page is listed in an `H1` heading, and that there is ideally only one such level-1 heading on any given page. It is also useful if that heading at least partially matches the page's `title` tag. Please refer to the e-Government standards published at <http://www.e.govt.nz/standards/> for more information on conveying semantics in your HTML, as well as other best practices.

Avoid the use of graphics-based navigation where there is no text-based alternative. While search engines may follow a link with no discernable title, the title of a link places great weight on what the search engine believes a target page is actually about, so is very important. A link without a text-based title will not transfer any "vote" for the topic of the target page.

You may also want to optimise your site's metadata to assist search engines in discovering the key topics of the page. While this used to be a key practice to rank well, it now only works well if the content of the page backs up the keyword claims. Nevertheless, all sites should be providing the various `<meta>` tags that describe a page, and you may additionally want to research using the Dublin Core Elements standard (<http://www.dublincore.org>) to more completely describe each page, including author and copyright information. For sites providing structured content such as blog entries or formal articles, the use of invisible microformats (see <http://www.microformats.org>) may also provide an advantage, as search engines will likely embrace such self-describing content in the near future.

2.3 Don't abuse the search engines

Many common practices to rank well in the past now actually hurt your ranking. In general, any change you make to your site with the intention of making it appeal more to search engines, while having the effect of making it appeal less to a human user, is going to impose a penalty.

Such practices include placing many general search terms in the title of a page which may not have anything at all to do with what that particular page is about. This not only confuses users, but also leads to demerit points from search engines that note that you're lying about the actual content of that page. As mentioned earlier, your top-level heading should match your page title, so if you're adding keywords to the title that you wouldn't put in a large heading on the page, they probably shouldn't be there.

Another current common practice is to include the site's name in the page's title, such as `About e-government - New Zealand E-government Programme`, which is usually acceptable as it helps users see what site they're on. The search engines will tolerate this, and it may even help your rankings if users search for a combination of your site's name as well as keywords from page content.

Always ensure that your title and meta tags (if present) correctly describe the content of the page they are on. Adding superfluous keywords to either will result in lower rankings and the possibility of being blacklisted for those terms.

2.4 Guide the search engines to the useful content

While search engines are relatively good at finding their own way around a site, occasionally they need further guidance. If there are any pages or areas of your site that you would prefer weren't added to a search index, you should place a `robots.txt` file in the root of your site specifying them. See the Robots Exclusion Standard at <http://www.robotstxt.org> for more details. This mechanism should be honoured by all search engines.

Another recent method available is the Sitemap Protocol (see <http://www.sitemaps.org>, an XML format that is now supported and encouraged by most major search engines. Using this protocol, you can give search engines guidance on which pages of your site you would like to have indexed, as well as how important those pages are in relation to each other, and how often you expect them to change. If your site is not currently capable of producing these XML sitemaps, it may be worth including them in a future redevelopment where practical.

Of course, classic page-based sitemaps and index pages on your site are also of benefit to search engines, but be aware that most search engines only read the first 100 links they encounter in an HTML page, so a large sitemap or index should be broken up into multiple pages to get maximum benefit.

It may seem counterintuitive to restrict search engine access to certain pages on the site if you are trying to increase your rankings, but remember that quality is more important than quantity to search engines; by only allowing the search engines to index pages that would be useful to a searcher, you are helping keep the search engine indexes clean, and therefore people will find what they want faster.

Common pages you should prevent search engines from indexing include:

- Web-based applications such as a webmail service that would provide no value to a searcher
- Content that is temporary in nature and would likely be gone by the time anyone tried to visit
- Any search result pages from an internal search engine – Search engines prefer not to index the results of other search engines
- Any content that isn't for public consumption – search engines find pages that normal users probably wouldn't

2.5 Use a meaningful and consistent URL scheme and keep it long-term

While the URL of a page on your site may not seem important, search engines may rank your site very favourably for terms discovered within the URL itself. In addition to this, meaningful URLs are more friendly to users and are generally more likely not to change.

If not doing so already, it is worth investigating and implementing a "speaking URL" scheme, where the URL does not point to actual files on your web server, but rather describes the path to the content. For example:

- Normal URL: `http://example.govt.nz/somedirectory/page.html?page_id=12345&lang=mi`
- "Speaking" URL: `http://example.govt.nz/maori/information/contact-us/`

Such speaking URLs are less likely to change over site revisions, so the search engine should never have to de-list and rediscover a page at a new location. In many cases, implementing such a URL scheme can be done without changing much of the site's backend architecture. In addition to this, search engines prefer URLs with fewer HTTP GET parameters (URLs containing a ? and & characters especially).

2.6 Properly use the HTTP protocol

Many sites do not use facilities available in the HTTP protocol correctly, which often has no impact on users, but can have a dramatic impact on the value a machine can derive from the site.

The first common problem is the use of old-fashioned "meta refresh" tags to direct users from one page to another. While these work, search engines *will not* follow them. A better solution would be to use standard HTTP response codes:

- HTTP 301 "Content Moved" – This should be used to direct someone to a replacement URL for the content they were attempting to access. Search engines will remember this and will visit the new URL from that point on
- HTTP 302 "Temporarily relocated" – This should be used in general situations where a page has moved temporarily, or you are redirecting based on some decision that you would not want the client to cache

Using the correct HTTP redirection code is critical for search engines. If a page has moved and visitors are directed to the new location with a 302 response, search engines will continue to direct visitors to the old address. Also, if your site needs to move to a new domain name, it is best to use a permanent 301 redirection of pages on the old domain to the new one, so the new domain comes up in search results instead of the old one.

Another common mistake is to serve up "Page not found" messages without using the 404 HTTP response code. If this happens, search engines will not remove the non-existent page from the index, instead replacing it with the content from the "Page not found" page. All such pages should be served with non-200 response code to prevent this from occurring. Another available response code is 410 Gone, which indicates that you are aware that content used to exist at that location, but is no longer available. Currently search engines will handle a 404 and 410 response identically; by removing the page from their index immediately.

Finally, for sites that have content in multiple languages, such as English and Māori, it is important to specify the correct ISO language code at the top of the HTML document or in HTTP response headers. Serving a page as English (usually the default) when it is not in fact in English will lead to it not being included in language-specific indexes that users may restrict their search to. See <http://www.w3.org/International/> for information on properly specifying the language of an HTML document.

2.7 Don't move content around unnecessarily and keep it on one domain name

Many sites will undergo major changes periodically, but in most cases the same content will still be available from each iteration. Search engines typically rank pages on a per-page basis, where the URL is the unique identifier, so changing URL means losing any rankings that page may have previously had.

Sites should not move to new domain names often, but if they do, a 301 redirect (as described in the previous section) is the best approach. Note that it may take up a year for a relocated site to rank well again in search results.

Disruption to search engines from site changes can be greatly minimised by following the advice given earlier on URL schemes and correct use of redirection codes.

Another issue is that of sites that are available on multiple domain names. Best practice is to redirect all alternate names to a common domain using 301 redirection. Failing to do this can cause each of the domains to be indexed separately and to compete against each other for search engine rankings. In a worst-case scenario, the pages of the site may be placed into “supplemental indexes”, an area where search engines such as Google place pages that appear to contain no original or useful content.

This includes normalising whether your site’s URL has a `www` prefix or not. Search engines may see `http://www.example.govt.nz` and `http://example.govt.nz` as different sites, so it is best to choose which is the preferred form and issue a redirect to visitors using the other.

2.8 Sign up for search engine webmaster tools and monitor search traffic

Many search engine operators now appreciate how difficult it has become to understand why a site isn’t ranking well, and have offered help in the form of webmaster tool areas on their sites. Such sites give you access to information such as current page or site rankings, common search terms that match your site, as well as any problems they’ve encountered while indexing your site.

Monitoring information in these tools (especially those from Yahoo and Google) may give you valuable insight into why your site is not performing well against searches.

Many web log analyser packages also allow an administrator to see what search engines and terms visitors are arriving from. It may be worthwhile following these statistics as a guide representing the type of content users would like to find on your site, and enhancing articles or pages relating to those terms.

3 Additional Notes

3.1 Optimising for a New Zealand audience

Many search engines have regional indexes, `http://www.google.co.nz` being one example of a New Zealand regional search page. In Google’s case, a website can only be in one regional index at once, so it’s important that your site is listed, you must satisfy at least one of the following requirements:

- Ensure the site has a `.nz` domain name, or
- Ensure the site is hosted on a server located within New Zealand

Sites that are included in Google’s New Zealand index will show up higher for people searching from within New Zealand, or other searchers that have restricted their search to New Zealand websites only.