Search Engine Optimization

Please be aware that individual search engines also have their own guidelines and recommendations for inclusion. While the guidelines below apply to **most** DSpace sites, you may also wish to review these guidelines for specific search engines:

- "Indexing Repositories: Pitfalls and Best Practices" talk from Anurag Acharya (co-creator of Google Scholar) presented at the Open Repositories 2015 conference
- Google Scholar Inclusion Guidelines
- Bing Webmaster Guidelines

Ensuring your DSpace is indexed

Anyone who has analyzed traffic to their DSpace site (e.g. using Google Analytics or similar) will notice that a significant (and in many cases a majority) of visitors arrive via a search engine such as Google or Yahoo. Hence, to help maximize the impact of content and thus encourage further deposits, it is important to ensure that your DSpace instance is indexed effectively.

DSpace comes with tools that ensure major search engines (Google, Bing, Yahoo, Google Scholar) are able to easily and effectively index all your content. However, many of these tools provide some basic setup. Here's how to ensure your site is indexed.

For the optimum indexing, you should:

1. Keep your DSpace up to date. We are constantly adding new indexing improvements in new releases
2. Ensure your DSpace is visible to search engines.
3. Enable the sitemaps feature – this does not require e.g. registering with Google Webmaster tools.
4. Ensure your robots.txt allows access to item "splash" pages and full text.
5. Ensure item metadata appears in HTML headers correctly.
6. Avoid redirecting file downloads to Item landing pages
7. Turn OFF any generation of PDF cover pages
8. As an aside, it's worth noting that OAI-PMH is generally not useful to search engines. OAI-PMH has its own uses, but do not expect search engines to use it.

Keep your DSpace up to date

We are constantly adding new indexing improvements to DSpace. In order to ensure your site gets all of these improvements, you should strive to keep it up-to-date. For example:

- As of DSpace 5.0, the DSpace robots.txt file now includes references to Sitemaps by default (see DS-1936), and also blocks known bad bots (see DS-2335).
- As of DSpace 4.0, DSpace has provided several enhancements, which were requested by the Google Scholar team. These included providing users (and web indexers) a way to browse content by the date it was added to DSpace (see DS-1482), ensuring the "dc.date.issued" field is set more accurately (see DS-1481), and enhancing the logic behind the "citation_pdf_url" HTML <meta> tag (see DS-1483)
- As of DSpace 1.7, DSpace has improved how its Item-level metadata is made available to Google Scholar. For the 1.7.0 release, the DSpace Developers worked directly with the Google Scholar developers, to ensure DSpace is generating the "citation_*" HTML "<meta>" tags (i.e. Highwire Press tags) that Google Scholar recommends in their Indexing Guidelines.
- As of DSpace 1.5, DSpace has support for sitemaps (both simple HTML pages of links, as well as the sitemaps.org protocol). It also includes item metadata in the HTML HEAD element of item display pages, ensuring that the metadata can be effectively indexed no matter what changes you might have made to your DSpace's layout or style.
- As of DSpace 1.4, DSpace has support for the "if-modified-since" HTTP header. This basically means that if an item (or bitstream therein) has not changed since the last time a search engine's crawler indexed it, that item/bitstream does not have to be re-retrieved, sparing your server.

Additional minor improvements / bug fixes have been made to more recent releases of DSpace.

Ensure your DSpace is visible to search engines

First ensure your DSpace instance is visible, e.g. with: https://www.google.com/webmasters/tools/sitestatus

If your site is not indexed at all, all search engines have a way to add your URL, e.g.:

- Google: http://www.google.com/addurl
- Yahoo: http://siteexplorer.search.yahoo.com/submit
- Bing: http://www.bing.com/docs/submit.aspx

Enable the sitemaps feature

DSpace provides a sitemap feature that we **highly recommend** you enable to ensure proper indexing. Sitemaps allow DSpace to expose its content in a way that makes it easily accessible to search engine crawlers. Sitemaps also help ensure that crawlers do NOT have to visit every page in your DSpace (which means the crawlers can get in and get out quickly, without taxing your site). Without sitemaps, search engine indexing activity may impose significant loads on your repository.

HTML sitemaps provide a list of all items, collections and communities in HTML format, whilst Google sitemaps provide the same information in gzipped XML format.

To enable sitemaps, all you need to do is run `[dspace]/bin/dspace generate-sitemaps` once a day.

Just set up a cron job (or scheduled task in Windows), e.g. (cron):

```
# Regenerate sitemaps at 6:00 AM local time each morning
0 6 * * * [dspace]/bin/dspace generate-sitemaps
```

Once you've enabled your sitemaps, they will be accessible at the following URLs:

- XML Sitemaps / Sitemaps.org syntax: `[dspace.url]/sitemap`
- HTML Sitemaps: `[dspace.url]/htmlmap`

So, for example, if your "dspace.url = http://mysite.org/xmlui" in your "dspace.cfg" configuration file, then the HTML Sitemaps would be at: "http://mysite.org/xmlui/htmlmap"

The generate-sitemaps command

This command accepts several options:

Option	meaning
-h --help	Explain the arguments and options.
-s --no_sitemaps	Do not generate a sitemap in sitemaps.org format.
-b -no_htmlmap	Do not generate a sitemap in htmlmap format.
-a --ping_all	Notify all configured search engines that new sitemaps are available.
-p *URL* --ping *URL*	Notify the given URL that new sitemaps are available. The URL of the new sitemap will be appended to the value of *URL*.

You can configure the list of "all search engines" by setting the value of `sitemap.engineurls` in `dspace.cfg`.

Make your sitemap discoverable to search engines

Even if you've enabled your sitemaps, search engines may not be able to find them unless you provide them with a link. There are two main ways to notify a search engine of your sitemaps:

1. **Provide a hidden link to the sitemaps in your DSpace's homepage.** If you've customized your site's look and feel (as most have), ensure that there is a link to `/htmlmap` in your DSpace's front or home page. *By default, both the JSPUI and XMLUI provide this link in the footer*:

   ```
   <a href="/htmlmap"></a>
   ```

2. **Announce your sitemap in your robots.txt**. Most major search engines will also automatically discover your sitemap if you announce it in your robots.txt file. *By default, both the JSPUI and XMLUI provide these references in their robots.txt file.* For example:

   ```
   # The FULL URL to the DSpace sitemaps
   # XML sitemap is listed first as it is preferred by most search engines
   # Make sure to replace "[dspace.url]" with the value of your 'dspace.url' setting in your dspace.cfg file.
   Sitemap: [dspace.url]/sitemap
   Sitemap: [dspace.url]/htmlmap
   ```

   a. These "Sitemap:" lines can be placed anywhere in your robots.txt file. You can also specify multiple "Sitemap:" lines, so that search engines can locate both formats. For more information, see: http://www.sitemaps.org/protocol.html#informing
   b. Be sure to include the FULL URL in the "Sitemap:" line. Relative paths are not supported.

Search engines will now look at your XML and HTML sitemaps, which serve pre-generated (and thus served with minimal impact on your hardware) XML or HTML files linking directly to items, collections and communities in your DSpace instance. Crawlers will not have to work their way through any browse screens, which are intended more for human consumption, and more expensive for the server.

Create a good robots.txt

The trick here is to minimize load on your server, but without actually blocking anything vital for indexing. Search engines need to be able to index item, collection and community pages, and all bitstreams within items – full-text access is critically important for effective indexing, e.g. for citation analysis as well as the usual keyword searching.

If you have restricted content on your site, search engines will not be able to access it; they access all pages as an anonymous user.

Ensure that your robots.txt file is at the top level of your site: i.e. at http://repo.foo.edu/robots.txt, and NOT e.g. http://repo.foo.edu/dspace/robots.txt. If your DSpace instance is served from e.g. http://repo.foo.edu/dspace/, you'll need to add /dspace to all the paths in the examples below (e.g. /dspace/browse-subject).

NEVER BLOCK THESE PATHS

Some URLs can be disallowed without negative impact, but be ABSOLUTELY SURE the following URLs can be reached by crawlers, i.e. DO NOT put these on Disallow: lines, or your DSpace instance might not be indexed properly.

- `/bitstream`
- `/browse` (UNLESS USING SITEMAPS)
- `/*/browse` (UNLESS USING SITEMAPS)
- `/browse-date` (UNLESS USING SITEMAPS)
- `/*/browse-date` (UNLESS USING SITEMAPS)
- `/community-list` (UNLESS USING SITEMAPS)
- `/handle`
- `/html`
- `/htmlmap`

Example good robots.txt

Below is an example good robots.txt. The highly recommended settings are uncommented. Additional, optional settings are displayed in comments – based on your local configuration you may wish to enable them by uncommenting the corresponding "Disallow:" line.

```
# The FULL URL to the DSpace sitemaps
# XML sitemap is listed first as it is preferred by most search engines
# Make sure to replace "[dspace.url]" with the value of your 'dspace.url' setting in your dspace.cfg file.
Sitemap: [dspace.url]/sitemap
Sitemap: [dspace.url]/htmlmap

##########################
# Default Access Group
# (NOTE: blank lines are not allowable in a group record)
##########################
User-agent: *
# Disable access to Discovery search and filters
Disallow: /discover
Disallow: /search-filter
# For JSPUI, replace "/search-filter" above with "/simple-search"
#
# Optionally uncomment the following line ONLY if sitemaps are working
# and you have verified that your site is being indexed correctly.
# Disallow: /browse
#
# If you have configured DSpace (Solr-based) Statistics to be publicly
# accessible, then you may not want this content to be indexed
# Disallow: /statistics
#
# You also may wish to disallow access to the following paths, in order
# to stop web spiders from accessing user-based content
# Disallow: /contact
# Disallow: /feedback
# Disallow: /forgot
# Disallow: /login
# Disallow: /register
```

WARNING: for your additional disallow statements to be recognized under the `User-agent: *` group, they *cannot be separated by white lines* from the declared `user-agent: *` block. A white line indicates the start of a new user agent block. Without a leading user-agent declaration on the first line, blocks are ignored. Comment lines are allowed and will not break the user-agent block.

This is OK:

```
User-agent: *
# Disable access to Discovery search and filters
Disallow: /discover
Disallow: /search-filter
Disallow: /statistics
Disallow: /contact
```

This is **not OK**, as the two lines at the bottom will be completely ignored.

```
User-agent: *
# Disable access to Discovery search and filters
Disallow: /discover
Disallow: /search-filter

Disallow: /statistics
Disallow: /contact
```

To identify if a specific user agent has access to a particular URL, you can use this handy robots.txt tester.

For more information on the robots.txt format, please see the Google Robots.txt documentation.

Ensure Item Metadata appears in the HTML HEAD

It's possible to greatly customize the look and feel of your DSpace, which makes it harder for search engines, and other tools and services such as Zotero, Connotea and SIMILE Piggy Bank, to correctly pick out item metadata fields. To address this, DSpace (both XMLUI and JSPUI) includes item metadata in the <head> element of each item's HTML display page.

```
<meta name="DC.type" content="Article" />
<meta name="DCTERMS.contributor" content="Tansley, Robert" />
```

If you have heavily customized your metadata fields away from Dublin Core, you can modify the crosswalk that generates these elements by modifying `[dspace]/config/crosswalks/xhtml-head-item.properties`.

Google Scholar Metadata in HTML HEAD

In addition to Dublin Core <meta> tags in the HTML HEAD, DSpace also includes Google Scholar specific metadata fields in each item's HTML display page.

```
<meta content="Tansley, Robert; Donohue, Timothy" name="citation_authors" />
<meta content="Ensuring your DSpace is indexed" name="citation_title" />
```

These meta tags are the "Highwire Press tags" which Google Scholar recommends. If you have heavily customized your metadata fields, or wish to change the default "mappings" to these Highwire Press tags, they are configurable in `[dspace]/config/crosswalks/google-metadata.properties`

Much more information is available in the Configuration section on Google Scholar Metadata Mappings.

Avoid redirecting file downloads to Item landing pages

Make sure that you never redirect "direct file downloads" (i.e. users who directly jump to downloading a file, often from a search engine) to the associated Item's splash/landing page. In the past, some DSpace sites have added these custom URL redirects in order to facilitate capturing statistics via Google Analytics or similar.

While these URL redirects may seem harmless, they may be flagged as cloaking or spam by Google, Google Scholar and other major search engines. This may hurt your site's search engine ranking or even cause your entire site to be flagged for removal from the search engine.

If you have these URL redirects in place, it is highly recommended to remove them immediately. If you created these redirects to facilitate capturing download statistics in Google Analytics, you should consider upgrading to DSpace 5.0 or above, which is able to automatically record bitstream downloads in Google Analytics (see DS-2088) without the need for any URL redirects.

Turn OFF any generation of PDF cover pages

While DSpace offers a PDF Citation Cover Page option, this option may affect your content's visibility in search engines like Google Scholar. Google Scholar (and possibly other search engines) specifically extracts metadata by analyzing the contents of the first page of a PDF. Dynamically inserting a custom cover page can break the metadata extraction techniques of Google Scholar and may result in all or much of your site being dropped from the Google Scholar search engine.

For more information, please see the "Indexing Repositories: Pitfalls and Best Practices" talk from Anurag Acharya (co-creator of Google Scholar) presented at the Open Repositories 2015 conference.

In general, OAI-PMH is not useful to Search Engines

Feel free to support OAI-PMH, but be aware that in general it is not useful for search engines:

- No reliable way to determine OAI-PMH base URL for a DSpace site.
- No standard or predictable way to get to item display page or full text from an OAI-PMH record, making effective indexing and presenting meaningful results difficult.
- In most cases provides only access to simple Dublin Core, a subset of available metadata.
- **NOTE:** Back in 2008, Google officially announced they were retiring support for OAI-PMH based Sitemaps. So, OAI-PMH will no longer help you get better indexing through Google. Instead, you should be using the DSpace 'generate-sitemaps' feature described above.

T