MIT Google - Add and Restrict Web Pages

Add Pages Into the Search Index

Google begins indexing at the top of the MIT website and follows links to find all indexable pages within the MIT search collection. For your pages to be indexed by Google, you simply need to make sure your page can be reached by clicking links from the MIT home page. If you would like your certificate-protected documents to be searchable, see MIT-Google - Secure Search.

There is no need to submit pages to the index; the Google crawler will pick up changed, new, and removed pages automatically during its continual crawling of the MIT website.

If you would like your page(s) to be listed in the MIT secondary pages, send a request to the MIT Webmasters. Also, if your site is affiliated with a department, lab or office, consider having your departmental or office site link to your site.

If you need a link to a personal home page, you can request to be listed on the MIT Community Home Pages.

Restrict Pages from the Search Index

- **Page-by-page**
  If you don't want a page to be indexed, insert `<meta name="robots" content="noindex, nofollow">` within the `<head>` tag of your document. This will prevent crawlers (robots) from indexing the page and from following any links from the page. If the page has already been indexed, it will be removed from the index the next time Google crawls the page.

- **If you run your own server**
  You can use the robots.txt file to exclude search engines from indexing the site. The Googlebot user-agent refers to the world-wide `www.google.com` crawler; the MIT-Google crawler is called `gsa-crawler`.

- **Emergency removals**
  If you need to get a page out of the index urgently, send a request to the MIT-Google Team, and provide the URL of the page you want removed.

Troubleshooting

**Why isn't my site in the search index?**

Please check the following:

- Is your site linked to from another MIT site, and is that site searchable by MIT-Google? If your site is new, it may take a few days before MIT-Google finds and indexes your page.

- If your site is hosted on your own server or by scripts.mit.edu, verify that your-server.mit.edu/robots.txt does not exist. If it does exist, ensure that MIT-Google (user-agent gsa-crawler) is allowed to crawl your content.

- Verify that your pages do not contain `<meta name="robots" ... >` tags.
Why isn’t my site near the top of the search results?

We don’t have much control over or visibility into the ranking algorithm used to produce search results. However, here are a few tips that may help improve your site’s ranking:

- Be sure to specify a clear site title in the `<title>` tags. Generally, anticipate what phrase(s) visitors might use when searching for your site, and strive to incorporate those phrases in your title.

- Use text and HTML markup for your content, rather than relying solely on graphics; especially in titles and headings. For instance, make effective use of `<h1>`, `<h2>`, etc, tags rather than graphical banners.

  Search engines are primarily text-based and do not read content within graphics. It is possible, however, to use both textual markup (for search engines) while displaying graphical banners (for users) with effective use of Cascading Style Sheets (CSS).

- Avoid embedding information in Flash; MIT-Google cannot read Flash. Encourage as many other sites as possible to link to your site. Ensure that the link text they use to refer to your site is descriptive.

- Meta tags don’t hurt, but their effectiveness with improving your site’s ranking on MIT-Google is limited.

Additionally, there is a large amount of information on Search Engine Optimization techniques on the web.

Google Search Configuration

MIT has customized the Google Search Appliance for our environment, with the following changes:

No caching

The commercial Google search engine caches a copy of each page that it indexes. If page content has been changed since the index was last updated, the user can view the cached version of the page (that is, the page as it existed when it was indexed). For security and privacy reasons, the MIT index does not use the caching feature.

Search collection

MIT’s search collection includes all the web pages in the mit.edu domain, specifically:

- `http://web.mit.edu` and
- `http://host.mit.edu`

...that are not specifically excluded by:

- the search administrator
- a `noindex` tag in the page’s HTML
- certificate protection although content available to all MIT users is indexed
- dynamically-generated content

Web pages excluded by the search administrator

Web pages in the following directories (and their sub-directories) are excluded from the MIT search collection:

- URLs containing this string:
  - `athena.mit.edu`
  - `sipb.mit.edu`
  - `dev.mit.edu`
  - `net.mit.edu`
  - `lees.mit.edu`
  - `ops.mit.edu`


- Hypermail and pipermail (archives)
- Java, Perl, Python documentation
- Debian, GNU/Linux mirror
- Specific pages kept out of the index at the request of their owners
- Dynamically generated pages, such as URLs containing `cgi-bin` and question marks.

These pages have been excluded for a variety of system performance, copyright, license, and Institute policy reasons. Additional directories or pages not listed here may have been excluded by the search administrator. If you think your page may have been excluded and don’t want it to be, contact the MIT-Google Team.
Crawling Schedule

The search appliance continuously crawls documents on the MIT domain. If your new page must be included in search results immediately, or if you have questions about the indexing of your content, contact the MIT-Google Team

Related Links

- Google Search Engine at MIT
- Add a Search Box on Your Site
- Handling Redirection When Your Site Moves
- Changing Your Directory Information
- Secure Search