

The new Penn State Search Engine

Presentation and discussion for
Penn State Webmasters

Web Developer's Lunch – September 2003
Presented by Jeff D'Angelo

Prerequisites

The presentation covers topics that assume a basic understanding of computers, the World Wide Web, HTML and search engines.

Headlines

- Spring 2003 – Search Engine Committee chooses Google Search Appliance
- Aug 11 – search.aset.psu.edu goes online
- Aug 25 – GSA replaces search.psu.edu
- Sept 25 - search.cac.psu.edu goes offline

Cliff Notes

- Aug 25 – all search engine invocations must change
- No longer need to ask search engine administrators to index a site
- New contact: search@psu.edu;
webmaster@psu.edu no longer responsible

How to invoke the search engine

- New HTML `<form>` code – See <http://aset.its.psu.edu/googledocs/instructions.html#invoke>
- Common mistake: `<input name="q">`
- Common mistake: “PennState” no space
- Copy/Paste best option

Search invocation options

- `as_sitesearch` – limit a search to a single domain or site
- `sitesearch` – same as above, but hide the “`site:domain`” query parameter
- `restrict` – search within a subcollection
- Details:
<http://aset.its.psu.edu/googledocs/instructions.html#restrict>

How to be indexed, and rank well

- Chances are, you are already indexed.
- Clear organization and good site design first step.
- Set a title, meta keyword and description.
- Avoid frames, or at least include a `<noframes>` version.
- Use text, not text in images.

How NOT to be indexed

- Domain Webmaster control – robots.txt
 - User-agent: *
 - Disallow: /
- Content Editor control – robots meta tag
 - `<meta name="robots" content="noindex, nofollow, noarchive">`
- Ultraseek specific `<!--stopindex-->` not effective
- Acrobat Standard Security = noarchive for .pdf
- password protection / other restriction (IP?)

New Policies

- HTML docs no larger than 2.5 MB, non-HTML no larger than 30 MB (only first 2.5 MB indexed)
- cgi-bin, ?, not indexed unless requested
- Excluded by policy (don't ask to include):
 - Personal Web server www.personal.psu.edu
 - Residence Halls xyz123.rh.psu.edu
 - Dial-up modems, mobility ports, vpn and mobile wireless connections

New Features

- SSL secured pages (https://) now indexed
- More document types indexed, including:
 - Adobe Portable Document Format (.pdf)
 - Microsoft Office Suite: Word, Excel, Powerpoint
 - Postscript
 - Full list: <http://aset.its.psu.edu/googledocs/filetypes.html>
- Site restricted searches and subcollections

Questions?

- Questions?
- Pizza?
- Tour?
- Easter Egg?