Get Your Site Indexed

• Three basic questions from this chapter:
  – What if your site is not indexed?
  – How many pages on your site are indexed?
  – How do you get more pages indexed?
What If Your Site Is Not Indexed?

- Are you certain it is not indexed?
  - Look for a PageRank bar in the Google Toolbar
  - Perform a **site:** search in Google/Yahoo/Ask/MSN
What If Your Site Is Not Indexed?

• Verify your site is not banned or penalized
  – Has the number of your pages in the search engine decreased recently?
  – Can you only find your home page via a direct search with the URL? (not with relevant queries?)

• Search engines publish guidelines against spamming techniques
  – e.g., Google
  – What is search engine spam?
Search Engine Spam
Sometimes also known as web spam, spamdexing

- The use of techniques to manipulate search engines, typically to generate undeservedly high search engine result page rankings.
  - Cloaking (including hidden text, redirection)
  - Keyword stuffing
  - Link farms, link exchanges, web rings
  - Comment spam, referrer spam
  - Link bombing (a.k.a., googlebombing, "miserable failure", "out of touch executives")
  - Blog spam (splogs)
- Methods you want to avoid!
Cloaking Example

Examples credit: Kumar Chellapilla, Microsoft Live Labs
What If Your Site Is Not Indexed?

• Make sure search engine spiders are visiting
  - Examine your web server logs

- Missing spiders?
  • Perhaps site new, or down, or not linked
  • Could submit site to engines, or better, get links
What If Your Site Is Not Indexed?

- Get sites to link to you
  - Best method to attract search engine spiders
  - Get linked from a directory
  - Create a few links from your other pages
  - Start a campaign to attract links (chapter 13)
How Many Pages on Your Site Are Indexed?

• Determine how many pages you have
  − Ask webmaster
  − Check internal (intranet) search engine
  − Add up content sources (e.g., the number of items in your database)
  − Ask many search engines (site: query) to estimate

• Check how many pages are indexed
  − Again with site: query

• Calculate your inclusion ratio
  − Indexed/total documents
  − Want near 100%!
How To Get More Pages Indexed?

● Primary concern addressed by chapter

● Many possible approaches
  – Eliminate spider traps
  – Reduce ignored content
  – Create spider paths
  – Use paid inclusion
How To Get More Pages Indexed?

• Eliminate spider traps...
  – Carefully set robots directives

```bash
# robots.txt for www.davison.net

User-agent: ExtractorPro
Disallow: /

User-agent: DIIbot
Disallow: /

User-agent: *
Disallow: /admin
Disallow: /errors
Disallow: /lines
Disallow: /~kriser
Disallow: /~kai
Disallow: /cgi-bin
Disallow: /web-caching
```

Avoid infinite URLs!
How To Get More Pages Indexed?

- Eliminate spider traps...
  - Eliminate pop-up windows
  - Don't rely on pull-down-down navigation
How To Get More Pages Indexed?

- Eliminate spider traps...
  - Simplify dynamic URLs
  - Consider URL rewriting to look like a static URL
How To Get More Pages Indexed?

- Eliminate spider traps...
  - Eliminate dependencies to display pages
    - Cookies
    - JavaScript
    - Flash/Java
    - Login for personalized site
How To Get More Pages Indexed?

• Eliminate spider traps...
  - Ensure your web servers respond
    • Spiders will ignore sites that are down or too slow
  - Use redirects properly
    • Begs a few more questions:
      - What are redirects?
      - Why do we want to use them?
      - Are all redirects equally useful?
Redirects

- Redirects are how a web request for one page will automatically get redirected to another page.

- Four kinds of redirects:
  - JavaScript redirects
    - As in Dr. Davison's homepage
  - HTTP response code 301 (permanent redirect)
    - As when missing the trailing slash of a directory as in http://www.cse.lehigh.edu/~brian
  - HTTP response code 302 (temporary redirect)
    - As in Lehigh's home page
  - Meta refresh redirects (example next slide)
Meta Refresh Redirect

Transferring you to the Department of Civil & Environmental Engineering, Lehigh University Website

If you are not automatically transferred within 5 seconds,
click here: Civil & Environmental Engineering
Meta Refresh Redirect

Transferring you to
Civil & Environmental
University

If you are not automatically redirected in
seconds, click here: Civil & Environmen

This page has moved.
Please click here to be automatically redirected to the
new page location.

Tip: Remember to update any bookmarks that you
may already have to this page.

©2002-2005 Lehigh University, 27 Memorial Drive West, Bethlehem, Pa.
18015 Tel. (610) 758-3000
Useful Redirects

- Over time, pages get added and removed
- A request for a missing page will generate a **404 Not Found** error
- Redirection can send your browser to the new location automatically
- Server-side redirects will also affect crawlers
  - 301 redirects will transfer value of old links to new
    - Crawls new URL, removes old
  - 302 will index content of new at URL of old
How To Get More Pages Indexed?

- Reduce ignored content
  - Slim down your pages
    - Use an external JavaScript file (good for caching, too)
  - Validate your HTML
    - Spiders are less forgiving than browsers
    - Use tools like the W3C Validation Service.
- Reserve flash for content you do not want indexed
  - Crawlers generally don't parse it
- Avoid frames
  - Poor usability, often difficult for crawlers
How To Get More Pages Indexed?

- Create spider paths
  - That is, create pages with easy to follow links through your site
- Site maps
  - Useful for both human and robot visitors
- Country maps
  - Great when organization/products are spread out across many country-specific sites
- Google SiteMaps
  - Direct feed of list of (all) URLs, easily updated
  - Available through Google webmaster tools
Sample Site Map

The eBay site is organized into five sections: Buy, Sell, My eBay, Community and Help. Get an overview of each section below.

**Buy**
- Registration
  - Register to Buy
  - Confirm Registration
- Categories
  - All Categories
  - Antiques
  - Art
  - Automotive (eBay Motors)
  - Baby
  - Books
  - Business & Industrial
  - Cameras & Photo
  - Cell Phones & PDAs

**Sell**
- Selling Activities
  - Add to Your Item Description
  - Block Bidder/Buyer List
  - Buyer Block Exemption List
  - Cancel Bids on Your Listing
  - Change Your Item’s Gallery Image
  - End Your Listing
  - Manage Your Counter
  - Manage Your Store
  - Pre-approved Bidder/Buyer List
  - Promote Your Item
  - Promote your listings with link buttons
  - Rerlist Your Item

**Community**
- Feedback
  - Feedback Forum
  - Follow Up to Feedback You Left for Others
  - Leave Feedback for a Member
  - Leave Feedback for a Transaction
  - Make Feedback Public or Private
  - Reply to Feedback Received
  - View a Member’s Feedback
- Connect
  - About Me
  - Chat Rooms
  - Discussion Boards
How To Get More Pages Indexed?

• Use paid inclusion
  – Paid inclusion can make your life easier
    • It can index more of your site
    • It is cheaper than paid placement
    • It can adapt quickly to changes
    • It lets you test changes to your site quickly
    • It is easy to get stared with paid inclusion
  – Making the most of paid inclusion
    • Realize that all content is reviewed
    • Avoid unrelated keywords (can be considered spam)
Chapter Summary

• Chapter answered three basic questions:
  – What if your site is not indexed?
  – How many pages on your site are indexed?
  – How do you get more pages indexed?
    • Eliminate spider traps
    • Reduce ignored content
    • Spider paths
    • Paid inclusion