Institutional Web Management:
The Next Steps

Getting Robots to Index Your Web Site

Aims Of This Exercise

This exercise is intended for use in a hands-on exercise / discussion group. The aims of the exercise are to identify techniques for getting your web site indexed by search engines, and also ways to stop robots indexing parts of your web site.

1.1 Submitting URLs to Search Engines

Go to the address <`http://www.altavista.com/`>: Scroll to the bottom of the page and then click on the link to **Add A Page**. Submit a resource, such as your institutional home page.

Is this a usable service for your institutional web site?

1.2 Submitting Services

Go to <`http://www.submit-it.com/`>. Read the information provided. Then go to <`http://www.netcreations.com/postmaster/`> and read the information provided there. Do you think it would be cost-effective for your institution to subscribe to a service such as these?

1.3 eXcite

Go to <`http://www.excite.com/Info/listing.html`>. Read the information on indexing your web site. Give details of the page should you submit if you want your entire web site indexed.

1.4 The Home-Grown Approach

Describe techniques for creating a file which contains the URLs for all pages on your web site.

Describe techniques for using this information to have your web site indexed.

1.5 Restricting Access To Robots

Go to the address <`http://www.ukoln.ac.uk/web-focus/webwatch/services/`>. Choose the **/robot.txt Checker** service. Check the `robots.txt` file for your institutional web server. Then view the `robots.txt` file.

Select a number of other web sites and view their `robots.txt` file. Include IBM's and Sun's web site's.

Read the information about the Robots Exclusion Protocol at <`http://info.webcrawler.com/mak/projects/robots/norobots.html`>.

Use of the Robots Exclusion Protocol can improve your server's performance by removing unwanted access by robots. It can also improve the quality of hits relating to your web site by allowing only quality resources to be indexed.

Do you think you should make greater use of the Robots Exclusion Protocol on your web site? If so, how with you decide which resources to disallow robots from indexing?

1.6 Managing Restricted Access To Robots

Try to think of ways in which departmental information providers can ensure that resources which they are responsible for are not indexed unnecessarily.

1.7 Conclusions and Recommendations

What conclusions and recommendations have you drawn from this exercise?

Further Information

For further information see the following URLs:

<http://searchenginewatch.com/webmasters/features.html>

<http://www.eiffel.com/private/meyer/robots.html>