RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

# The Srofulous port of Screen Scraping: A Tutorial for the Tyro

John P. Joergensen
Associate Dean for Information Services
Rutgers University School of Law - Newark

# Introduction: What is it?

- Harvesting content from the Internet.
  - Static Harvesting
  - Data Mining
  - Dom Manipulation
- Methods
  - Manual
  - Scripting
  - Wholesale copy
  - Something in Between

# WGET As a very good friend

- Easy-ish to use:
  - *$ wget -i inputfile* Reads lines of URLs in a file
  - *$ wget -r -l 4* http://example.com Recursive: get the URL and everything it links to, here to 4 levels.
  - *$ wget* http://example.com *-nc* No Clobbering: Don't re-download anything you already have.
  - *$ wget* http://example.com *--timestamping* Put a timestamp on the filenames as you download.
  - Also:
    - --load-cookies cookies.txt
    - --user-agent=”Mozzilla”

# Examples

Circuit Courts: search engines

Fdsys: URL analysis

Library of Congress: Session Data

Common threads: CGI and URLs and Post v. GET

# FDSYS: They like to make it easy.

Simple (ish) URL analysis.

Nested Loop to get documents.

# Circuit Courts: Data Mining

Database manipulations.

2nd Circuit: POST data.

6th Circuit: Simple URL manipulation in GET.

3rd Circuit: Just get the recent stuff.

# Circuit Courts: File Processing.

- Search result page contains metadata and links to doc.
- Grab the document
- Grab the Metadata.

# Library of Congress Catalog

Session data stored in variables that need to be noticed and accounted for.

- They may be in URL's and they may be stored as POST data.

See URL: PID and SEQ.  Where do they get assigned?

See MARC Designation.

# Let's be polite . . .

Use bandwidth limits and wait times.

- *$ wget http://example.com --limit-rate=128* Limit the download rate to something reasonable.
- *$ wget http://example.com --wait=4* Wait x number of seconds between retrievals.
- *$ wget http://example.com --wait=4 --random-wait* Wait a random number of seconds around the wait interval.