Ethics and Law on the Web

Information Retrieval

© Crista Lopes
Main Point

• Just because you can doesn’t mean you should
Guidelines for Bot Writers

• An old but still relevant text
  • [http://www.robotstxt.org/guidelines.html](http://www.robotstxt.org/guidelines.html)

• Summary
  • **Be Accountable** - if your actions cause problems, be available to take *prompt* action in response;
  • **Test Locally** - expand the scope gradually before you unleash your crawler on others;
  • **Don't hog resources** – web servers are for people primarily
  • **Stay with it** - "it's vital to know what your robot is doing, and that it remains under control".
Guidelines for Bot writers

• Find out the sites’ crawling policies
  • E.g. sourceforge:
    • http://sourceforge.net/apps/trac/sitelegal/wiki/Crawler%20policy
  • Contact the Web admins
Ethics and Law

Ethical Issues

Laws
The Web and the Law

• Web: almost 20 years old (your age!)
• Wild Wild West, for the most part
  • Very few laws
  • Existing laws either untested or outdated
• Our own judgments/actions matter
Who can offend who

Users

Providers

Personal Information (Privacy)

Content (Security)
Privacy Statements

• Web sites may have privacy statements
  • What info they collect, how it’s used
• E.g. http://uci.edu/privacy.php
Privacy – Apache Log

66.249.66.18 - - [25/Jan/2012:00:18:27 -0800] "GET /xwiki/bin/export/Stats/CurrentYearActivity?format=rtf HTTP/1.1" 404 335 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

180.76.6.28 - - [25/Jan/2012:00:18:31 -0800] "GET / HTTP/1.1" 200 45 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)"

66.249.66.18 - - [25/Jan/2012:00:19:02 -0800] "GET /calswim/coverage-s1.png HTTP/1.1" 200 961543 "-" "Googlebot-Image/1.0"

173.192.98.90 - - [25/Jan/2012:00:33:00 -0800] "GET / HTTP/1.1" 404 290 "-" "Python-urllib/2.4"


180.76.5.65 - - [25/Jan/2012:02:18:20 -0800] "GET / HTTP/1.1" 200 45 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0;"
Terms of Use

• Web sites may have Terms of Use, or Terms of Service
  • Sometimes there is an explicit step for users to accept them
  • Most times there isn’t
• Express wishes from providers
• e.g. http://www.cnn.com/interactive_legal.html
• Bot writers: look for this
Copyrights

• Products of intellect are ruled by “Copyright Law”
• Web content, source code, etc. are products of intellect
Copyrights

- [http://www.copyright.gov/](http://www.copyright.gov/)

- (next slides taken from here)
What is Copyright?

“Fair Use”

• Various purposes for which the reproduction of a particular work may be considered fair
  • Criticism, comment, news, teaching, scholarship, and research
  • It’s a grey area

• [http://www.copyright.gov/fls/fl102.html](http://www.copyright.gov/fls/fl102.html)
“Fair Use”

• four factors to be considered in determining whether or not a particular use is fair:
  • The purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes
  • The nature of the copyrighted work
  • The amount and substantiality of the portion used in relation to the copyrighted work as a whole
  • The effect of the use upon the potential market for, or value of, the copyrighted work
Legal Case Studies

• eBay vs. Bidder's Edge
  • eBay successfully stopped crawlers from Bidder’s Edge
  • “trespass to chattels”

• American Airlines vs. FareChase
  • Online fare comparisons