Web scraping with Ruby

Remigijus Jodelis
Ruby workshop, 2011-03-19, Vilnius

step 1 - fetching data

- *net/http*
- *mechanize*
- *wget*
- *curl, curb*
- *httrack*

...

step 2 - processing

- regexp
- hpricot
- nokogiri
- scrapi

...

step 3 - storing / publishing

- file (csv, yaml, plain ruby)
- database
- your API

obstacles

1. changing websites - break scrappers
2. invalid / inaccessible data
3. lack of unique identifiers

obstacles + solutions

1. sessions	1. mechanize & co.
2. obfuscation	2. decoding
3. flash	3. decompiler
4. IP traffic limits	4. distribute / proxies
5. proxy detection	5. elite proxies
6. captchas	6. workers in India :)

suggestions

1. cache all data
2. time delays to prevent DOS
3. standard inteface (rake fetch, process)
4. don't trust anything (checks everywhere)
5. use loggers
6. collect statistics
7. iterate often
8. *http://scraperwiki.com/*