Aggregation Hacks

Richard Clamp

Birmingham Perl Mongers, June 2005

So I was trying to be lazy...

Richard Clamp

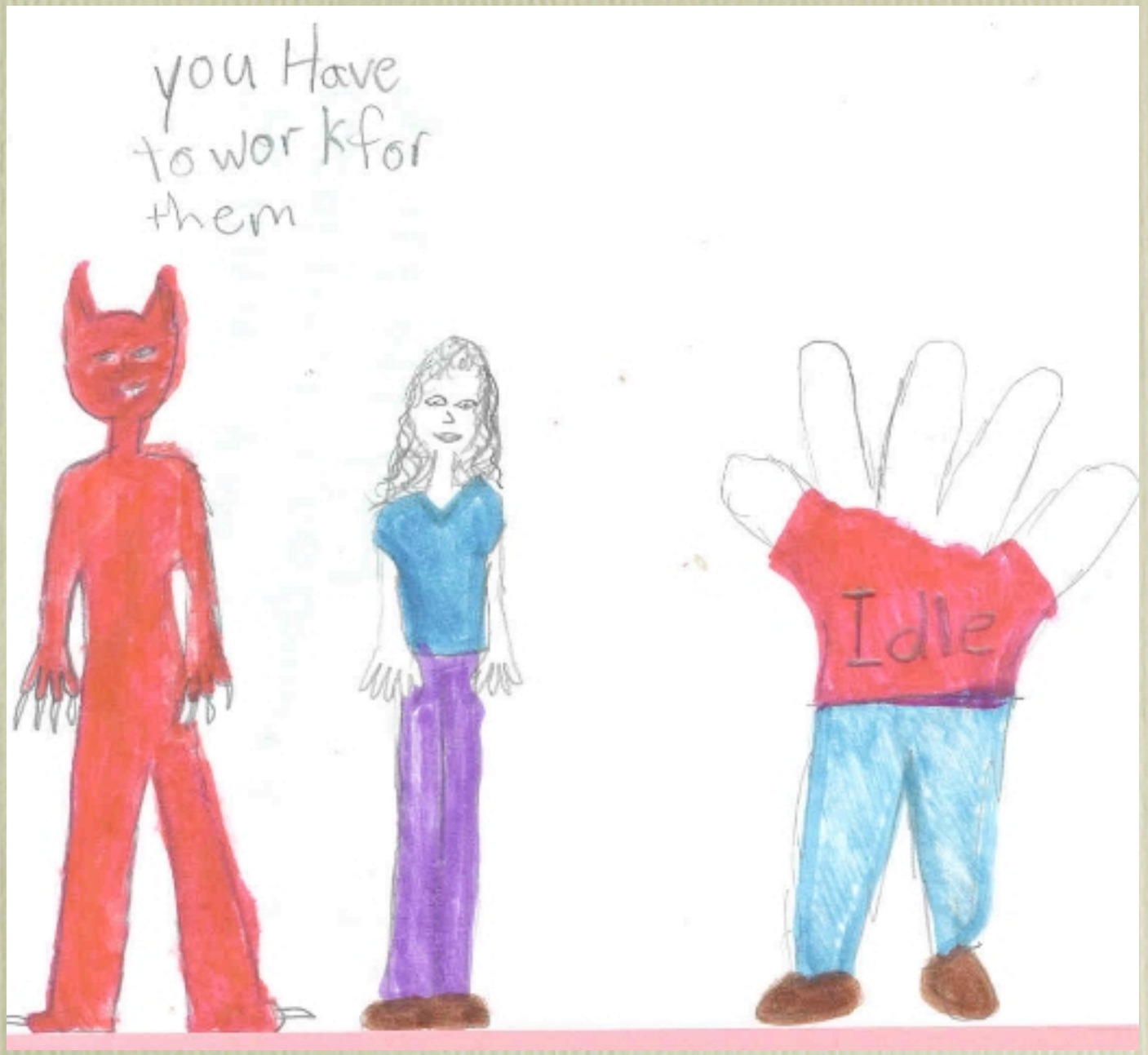
Birmingham Perl Mongers, June 2005

Obvirtues

- You can't mention laziness, without the other two
  - Impatience
  - Hubris
- But sometimes, they combine oddly

Laziness

- So this talk is about laziness
- Or rather the quest for idle hands
- I found something out about idle hands

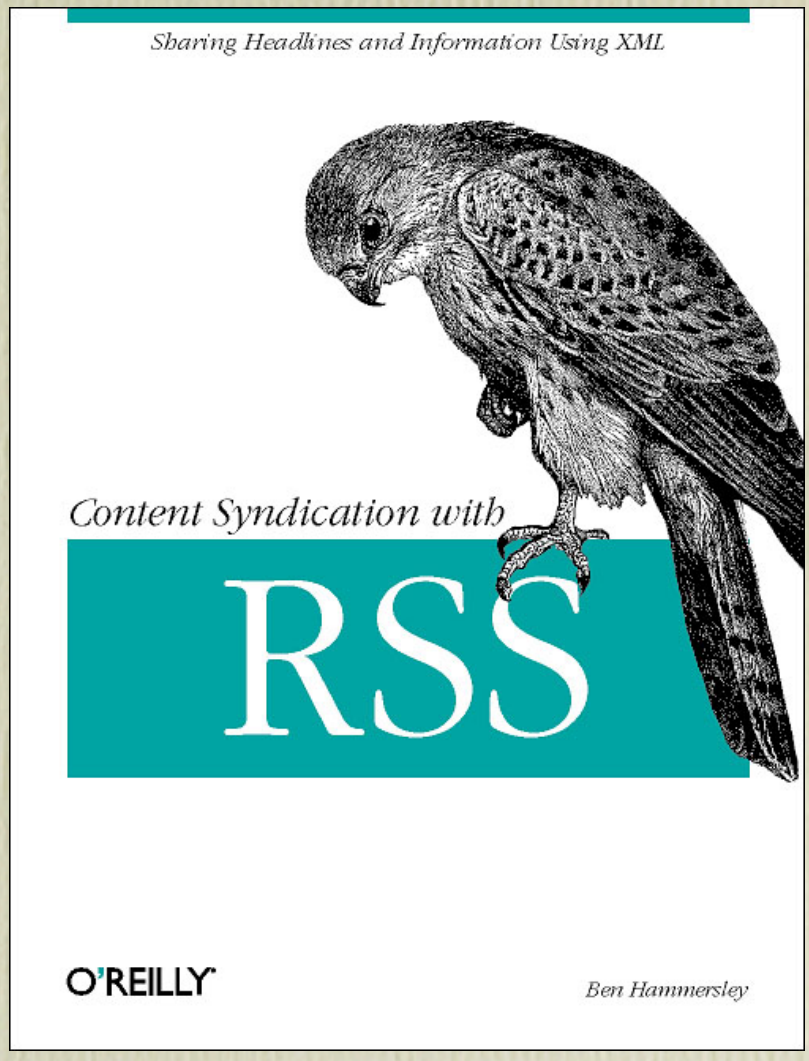

you Have
to wor k for
them
Idle

Buzzword Bingo

RSS	atom	opml
Class::DBI	Template::Toolkit	XPath
WWW::Mechanize	XML::Feed	YAML

RSS

- Rich Site Summary
- A Feed is a list of items
- An item is typically a news story
- More typically it's a blog post...



Timesink

- Had a Mac, which was poorly
- Also I'd bought a Zaurus
- Only mac app I'd really miss was NetNewsWire
- So I replicated the functionality as a web application

Timesink architecture

- Very simple data model - Feed, Items, Subscriber, Subscription, Seen
- Turned into classes with Class::DBI
- A sprinkling of the Template::Toolkit

More Timesink

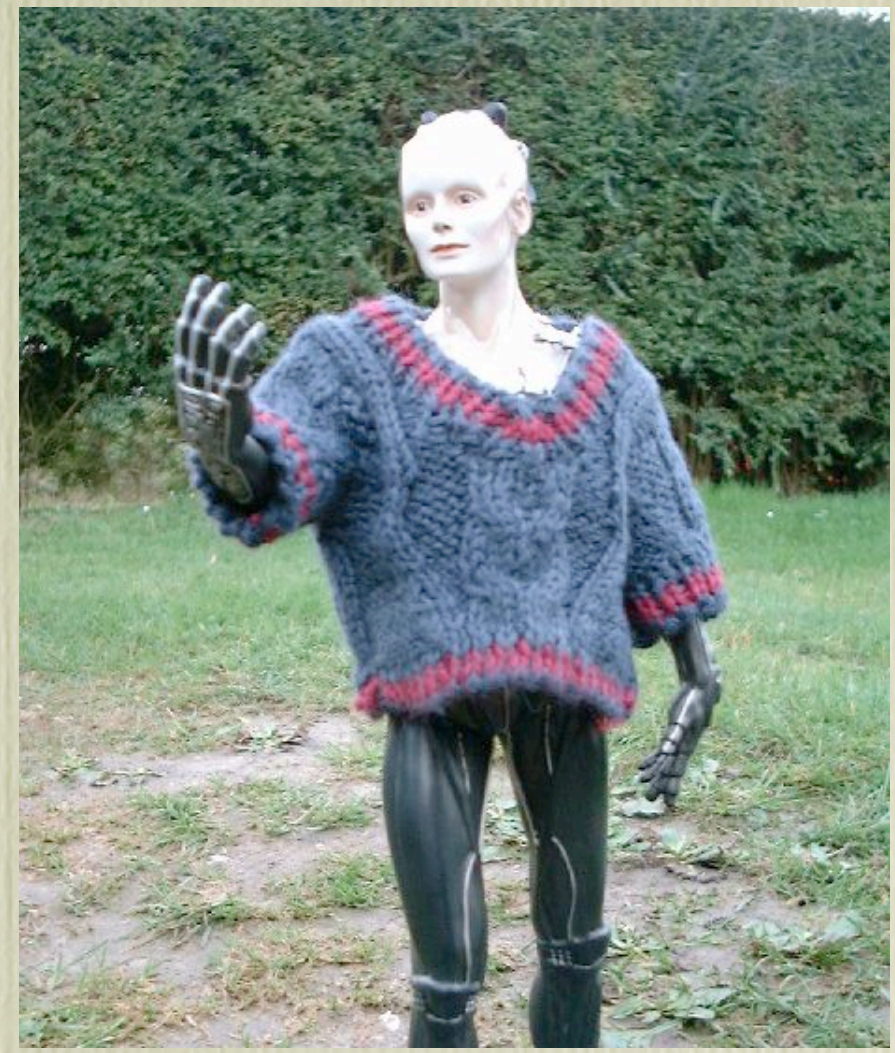
- Also command line app for managing feeds
- Uses XML::Feed so it magically deals with Atom and RSS
- New for 2005 - OPML export!

OPML

- Outline Processor Markup Language
- Just an XML schema for outline documents
- Also used by most RSS readers to export/ import subscription lists

rssborg

- Original idea by Simon Cozens
- Grab a page
- Use Template::Extract to pull out the news items
- Turn them into a rss feed using XML::RSS



Template::Extract

- looks like TT
- ```
  [% FOR records %]
  <!--START OF ABSTRACT OF NEWSITEM-->
  [% ... %]
  <a href="[% url %]"><acronym title="Click here to read this article">
  [% title %]</acronym></a>     ([% date %]) <BR>
  [% ... %]<font size="2">[% content %]</font></font>
  [% ... %]
  <!--END OF ABSTRACT OF NEWSITEM-->
  [% END %]
  ```
- but it gets compiled into this really really twisted regex, can be a pain to debug
  ```

comic2rss

- Lots of my browsing was keeping current with web comics
- I already had a mechanism for tracking things that change
- So I just needed to get the comics into RSS feeds

comic2rss 1.0

- WWW::Mechanize make this job simple
  - go to front page
  - while there are pages left to see
    - extract the image and add to feed
    - follow the link back

pesky interweb

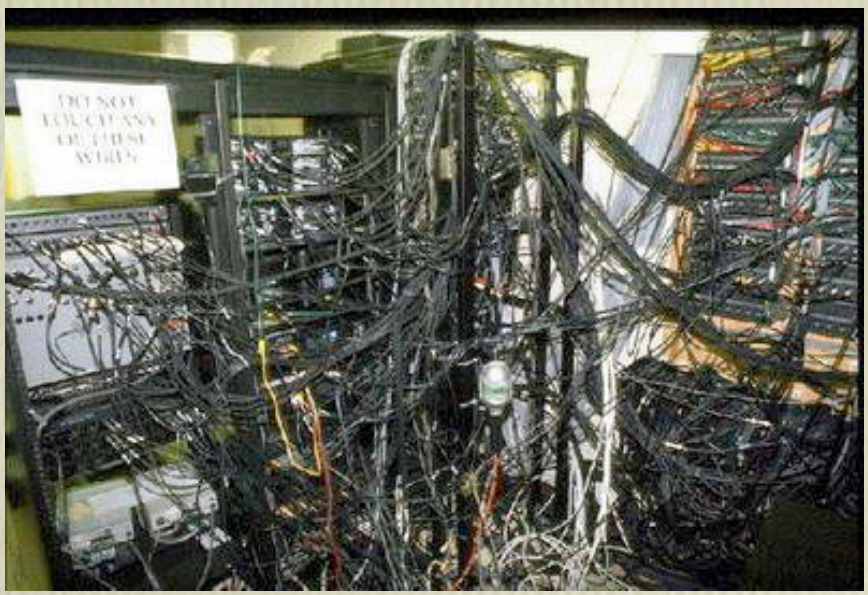
- Some sites are awkward though
- WWW::Mechanize can only follow links by URL or link text
- You can't say:
  - "Follow the link indicated by the back gif"

XPath

- “regexes for XML”
- `//img[@src = "image/back.gif" ]/../@href`
- That’s “the link surrounding the back image”

Bad HTML

- It's 2005
- XHTML 1.0 was released in 2000
- Most of the web is still cruddy HTML



Tidy to the rescue!

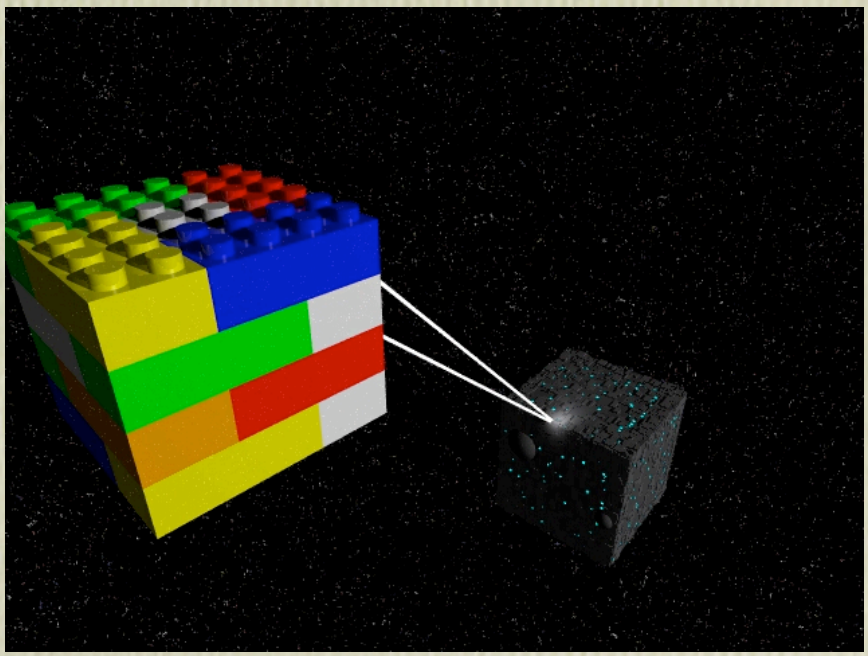
- Tidy takes in HTML
- and tries very hard to give you XHTML back

comic2rss 2.0

- Can handle the tricky navigation with XPath
- While it's there it'll grab you the news post
- And generates better RSS feeds

rssborg 2.0

- XML::XPath based rewrite
- Not really much else to see here



Resources

- Timesink
  - `http://unixbeard.net/svn/richardc/timesink/`
- comic2rss
  - `http://unixbeard.net/svn/richardc/misc/comic2rss/`
- rssborg
  - `http://blog.simon-cozens.org/bryar.cgi/id_6522`
  - `http://unixbeard.net/svn/richardc/cgi/`