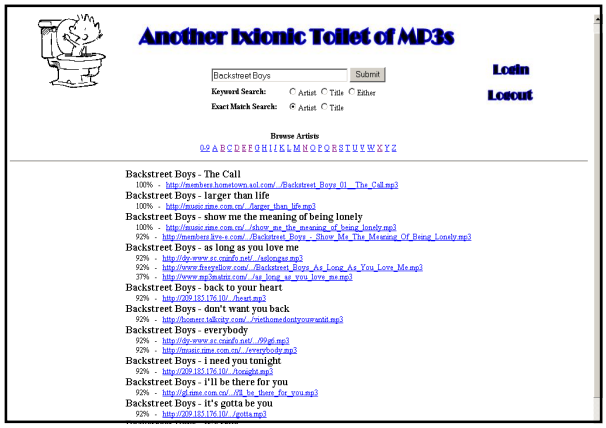
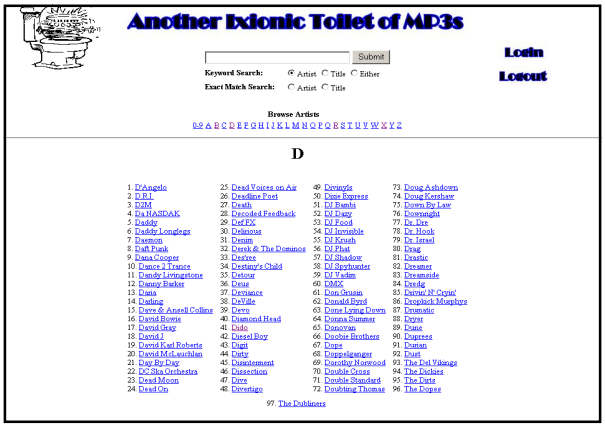
AIT Crawler

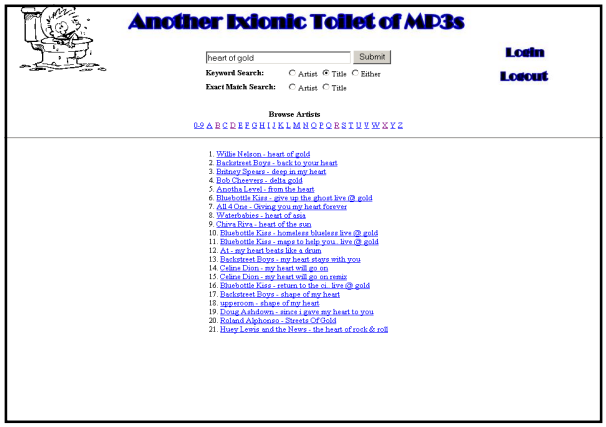
Anthony Johnson

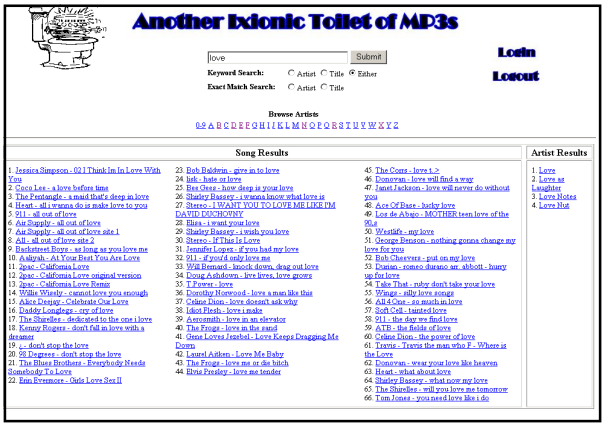
Igor Volynskiy

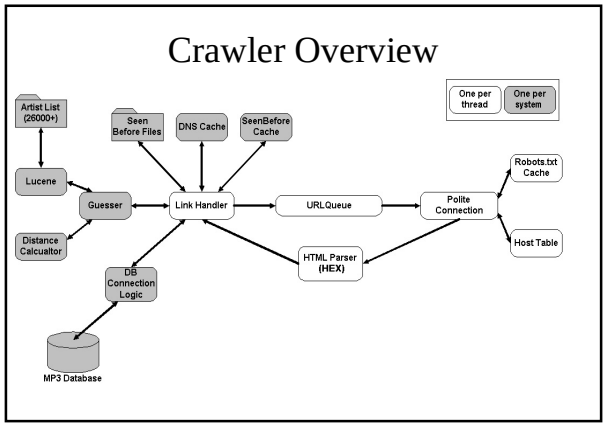
Travis Rothlisberger











Measurements

- First assignment crawl
  - CPU time 15m 40s
  - # pages visited 236
  - # Links followed 273
  - mp3s found 712
  - Pages/mp3s 0.33
  - CPU time per MP3 1.32 sec
- Guessing Accuracy
  - Overall: about 60%
  - Major Artists: about 80%

Lessons learned

- Teamwork
  - Project divisions, importance of communicating the interaction between individuals' parts.
- Planning
  - Concrete, well-thought out design. Foresight regarding potential issues and changes.
  - Design code to be flexible for any additions, reductions or changes that may need to be made in the future.
- End Product
  - Thoroughly test the code. Just because it works for a couple hours, doesn't mean it won't hit any snags.
  - Users prefer simple and easy to neat and complex.

Next Time…

- Better Crawling
  - Learn from the past, track statistics on servers.
  - Identify "hidden" MP3s, (.zip, .class, …)
  - Clean up afterwards. Remove dead/bad links.
- Better Planning
  - More in-depth design of website and UI before developing database.
  - More scalability. Allow the structure to expand more easily.
- Better Stuff
  - Artist and file info, discographies, bios, …
  - User interaction. Help name files, report dead links, request and post files, "punch the monkey" banners….