Email Extracting

Abram Hindle

Kitcher/Waterloo Perl Mongers
Canada
http://softwareprocess.es/
abram.hindle@softwareprocess.es

Email Extraction

- You have a set of emails
- You want to abstract the emails and put them into a database
  - Easier searching
  - Easier querying
  - Extract attachment
  - Cross Link

MBox

- Common format
- All emails raw concatenated together seperated by newslines a From header
- If you can cat raw emails you can make an mbox file
- Very easy to create

Popular Modules

- Mail::Box
- Mail::MboxParser

Mail::Box

- Mail::Box
  - Large suite, all encompassing
  - Parses many formats
  - Modern
  - Has addon to handle threads

Mail::MboxParser

- Mail::MboxParser
  - Handles quotes in bodies
  - Smaller
  - Less surprises
  - Less Modern
  - Easier to use interface for most things

Problem with both suites

- Dates!
- Mail::Box has timestamp, which is a best guess
- MBox::Parser is inconsistent, multiple date methods
  - We chose MBox::Parser anyways because we want to strip quotes from messages so we can do text analysis.

Note though...

- You can always use both!

Mail::MboxParser Patterns

- Create a new parser on a file descriptor
- Iterate through messages use

  `$mb->next_message`
- Get body and values
- Clean the values
- Store the data

Issues with email data

- Dates!
  - Inconsistent
  - Different Languages
- Message bodies are not clear
  - `my $body = $msg->body($msg->find_body);`
- Avoid signatures !

Dates

- Date from header
- Recieved header
- x-list-recieved-date
- `$msg->from_line`
- Then parse it! Date::Parse, Date::Manip

We want to see who's talking!

- Email::Find
  - `find_emails($text, $callback($email,$name))`
  - Annoying interface but works
- To / CC /From etc.

Things to watch out for with emails

- Multiple addresses same person
- Email addresses embedded in emails
- Attachments
- Data in attachments
- Parsing or avoiding quotes
- Threading

Resources

- http://search.cpan.org
  - Mail::MboxParser
  - Mail::Box
  - Email::Find
  - Date::Parse
  - Date::Manip