Open Source Data Sources
Academy of Management PDW
13 August 2006, Atlanta

James Howison
PhD Candidate
Syracuse University
School of Information Studies

Supported by The Syracuse FLOSS project with Prof. Kevin Crowston.
(NSF Grants 03-41475 and 04-14468. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.)
Overview

Types of data on open source teams
Ethical issues
Where and how can I get this data?
Difficulties in using data
Integrating types of data

Slides and References at:
http://floss.syr.edu/presentations/FlossDataTutAoM2006/
What’s available?

Project level data

- ‘Demographics’ (Start date, license etc)
- Team (Founder, roles etc)
- Communications (Email lists, IRC etc)
- Code repositories and release history

Cross project data

- Project lists and counts
- Relative statistics (Downloads, activity etc)
Ethical Issues with Data Use

Action in public, intended to be shared and observed
✓ But not for research … consider risks
Anonymized data can easily be traced
Should your research be available to the community it is based on?
Sources of open source data

Manual collection & ‘spidering’

Academic data and analysis sets
- Notre Dame’s Sourceforge Dumps
- FLOSSmole
- CVSanalY

Non-academic data and analysis sets
- OpenBRR
- Ohloh
Notre Dame Sourceforge dumps

Greg Madey working with Sourceforge
✓ Single interface to academic community
Monthly dumps of (almost) entire Sourceforge database
✓ ‘Demographics’
✓ Communications (except Mailing Lists!)
✓ Bug Tracker details
Contract with Madey’s group needed
Web form for SQL query, text file download
Wiki recently setup for community interaction
FLOSSmole

- Collaborative group of academic researchers
- Collective spidering of Sourceforge, Rubyforge, Freshmeat and ObjectWeb
  - Scripts to collect mailing lists from Sourceforge
  - Some data from Savannah and Apache
- Web SQL interface, script access available on request
- Analysis scripts largely available
- Mailing list and blog for communication
CVSanaly

Gregorio Robles and Libre Software Engineering project from Spain
Scripts convert code repository (eg CVS) logs into relational database
✓ “Who’s contributed the most code?”
MySQL dump of all Sourceforge projects available for download
Scripts can run against any CVS server
Non-academic sources

Ohloh

✓ “Objective metrics”
✓ Contributor graphs, COCOMO cost estimates

Open Business Readiness Rating

✓ Attempt at systematic ratings of projects to be used in software specification
✓ Aim to share ratings done by different organizations
Data difficulties

Dirty data
- Not all use all features of repositories
- Many projects outside your scope (e.g. single person or ‘dumped’ school projects)
- Highly skewed data (sampling difficulties)

Non-research data have response bias and low variance
- Includes Freshmeat ratings or Sourceforge’s ‘trove’ categories

Manual creation of comparable sets, manual confirmation of data comparability
Integrating Data and Next steps

Most studies use one only type of data
I’m currently developing a ‘Browser’
which combines sources using a simple
‘Actor’ does ‘Action’ structure
Data sharing is good, analysis script
sharing is excellent :-)

Open Source Data Tutorial, James Howison, Academy PDW
References

Slides, References and links at:

http://floss.syr.edu/presentations/FlossDataTutAoM2006/