Empirical Project Monitor and Results from 100 OSS Development Projects

Masao Ohira
Empirical Software Engineering Research Laboratory, Nara Institute of Science and Technology
ohira@empirical.jp
EASE Project

- Empirical software development environment for tens of thousands of projects
  - Massive data collection
  - Intensive data analysis
  - Feedback for software process improvement in organizations/communities (not only a single developer/project)
Empirical Environment

- Code clone detection
- Component search
- Metrics measurement
- Project categorization
- Cooperative filtering

- Product data archive (CVS format)
- Process data archive (XML format)

- Format Translator
- Format Translator
- Format Translator

- Versioning (CVS)
- Mailing (Mailman)
- Issue tracking (GNATS)

- Other tool data

- EPM (developing)

- Widely used development support tools

- GUI

- Managers
- Developers
EPM: Empirical Project Monitor

- A partial implementation of Empirical Environment
- Collect, measure, and show various data for project control
- Data source from tools used in software development
  - Versioning system (e.g. CVS)
  - Mailing list manager (e.g. Mailman)
  - Issue tracking tool (e.g. GNATS)
Architecture of EPM

- PostgreSQL (Repository)
- Standardized empirical SE data (in XML)
- CVS, Mailman, GNATS (ShareSource™)
- versioning history
- mail history
- problem history
- prediction/schedule metrics value other tool data etc.
- measurement of intra and inter projects
- analysis tools

developer
Manager

developer
manager

12/21/2003
Characteristics of EPM

- Use open source development tools
  → Easy to introduce
- Small overhead of data collection
  - Most data from versioning history
  - Communication through e-mail, and recoding issues by tracking tool
- Easy to transform other data format to the standardized empirical SE data format
Application Area of EPM

- **Large project**
  - Share project status immediately
  - Reduce project management load
  - Reduce risk for tampering data

- **Small project**
  - Apply with small cost
  - Apply to various projects, including XP and distributed development
Data collection from OSS Development Projects

- SourceForge.net
  - hosted projects: 72,853 (Dec. 15)
  - registered Users: 753,428 (Dec. 15)

- A variety of collaboration tools
  - SourceForge Collaborative Development System (CDS) web tools
  - Project Web Server
  - Tracker: Tools for Managing Support
  - Mailing lists and discussion forums
  - MySQL Database Services
  - Project CVS Services
  - etc.

Available data source for EPM
Overview of Collected Data

- 100 Active projects @ SF.net
  - Data sources for EPM
    - CVS data (only 40 projects)
    - Mailing Lists data
    - Issue (Bug) reports data
  - Project info. in a summary page
    - number of developers
    - period of a project
    - development status
    - intended audience
    - programming language
    - number of bugs
    - number of CVS commits
    - etc.
SourceForge.net

- Links to available data source for EPM
- Information related to the project

SourceForge.net is a website that provides information related to the project and links to available data source for EPM.
Summary of 100 OSS projects@SF.net: Evolution?

Graph showing data points with an arrow indicating a trend or direction.
Result of CVS Product Data:
Lines of Code
(history of software growth)
Result of CVS Process
Data: Check in/out
(history of developer’s activities)
How can we use such a lot of data?
Gross Classification using EVIDII

- EVIDII: Interactive interfaces that visualize relationships among three sets of data

(original application domain: face-to-face communication support between clients and designers)
Demo: organizing dynamic community?

Project info.
numbers of developers, LOC, development terms, etc.
Scenario: organizing a dynamic community / providing feedback for improvement

1. Comparing other projects with a target project

2. Finding similarities and differences between them

3-a. Notifying to related project leaders of the existence of communities

3-b. Identifying factors of the similarities and differences

4-a. Asking them help/advises for improvement

4-b. Providing suggestions for improvement
Summary and Future Work

- EPM: Empirical Project Monitor
- Data Collection from 100 OSS projects (only 40 CVS data…)
- Two scenarios using EVIDII

- More data collection (mails and bug issues) and analysis using EPM/EVIDII