A cloud for Project Census: Data Center for Understanding Software Development

Zhou Minghui
zhmh@pku.edu.cn
Peking University
2011-03-23
Software development:
What questions do we have?
A large network switch: several thousand developers

- Offshoring and Retirement
  - What’s the best practice?
  - How to measure the project’s performance?
  - How could we speed up newcomers?
Gnome and Mozilla

- **Big community**
  - Gnome: 156,332 people completed 517,801 issues, 1999-2011.1
  - Mozilla: 153,975 people completed 599,400 issues, 1998-2011.1

- **Questions**
  - How could so many people distributed over the world cooperate with each other to produce a product (without compensation)? Innovation in OSS?
  - How should I participate in an OSS project?
Hi guys

The vote is to decide if we change the project lifecycle maturation criteria.
The proposal is here:
http://www.ow2.org/view/TechnologyCouncil/Project+Maturation (rev 18.1)

[ ] +1 Use this set of criteria
[ ] 0 Don't care
[ ] -1 Veto, do not change anything (must be argued)

Happy voting :)  
--Guillaume
What could we do?

Open Source Data-Center
“Digital Archeology”
### Audris Mockus: Public repository: an Universal Version History

<table>
<thead>
<tr>
<th>Forge</th>
<th>Type</th>
<th>Files</th>
<th>File/Ver.</th>
<th>Unique File/Ver.</th>
<th>Branching</th>
<th>From</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large cmpny.</td>
<td>Var.</td>
<td>3,272K</td>
<td>12,585K</td>
<td>4,293K</td>
<td>2.9</td>
<td>1988</td>
</tr>
<tr>
<td>SourceForge</td>
<td>CVS</td>
<td><strong>26,095K</strong></td>
<td>81,239K</td>
<td><strong>39,550K</strong></td>
<td>2.1</td>
<td>1998</td>
</tr>
<tr>
<td>code.google</td>
<td>SVN</td>
<td>5,675K</td>
<td>14,368K</td>
<td>8,584K</td>
<td>1.7</td>
<td>1996</td>
</tr>
<tr>
<td>repo.or.cz</td>
<td>Git</td>
<td>2,519K</td>
<td>11,068K</td>
<td>5,115K</td>
<td>2.2</td>
<td>1986</td>
</tr>
<tr>
<td>Savannah</td>
<td>CVS</td>
<td>852K</td>
<td>3,623K</td>
<td>2,345K</td>
<td>1.5</td>
<td><strong>1985</strong></td>
</tr>
<tr>
<td>git.kernel.org</td>
<td>Git</td>
<td>12,974K</td>
<td><strong>97,585K</strong></td>
<td>856K</td>
<td><strong>114</strong></td>
<td>1988</td>
</tr>
<tr>
<td>OpenSolaris</td>
<td>Hg</td>
<td>77K</td>
<td>1,108K</td>
<td>91K</td>
<td>12.2</td>
<td>2003</td>
</tr>
<tr>
<td>FreeBSD</td>
<td>CVS</td>
<td>196K</td>
<td>360K</td>
<td>75K</td>
<td>4.8</td>
<td>1993</td>
</tr>
<tr>
<td>Kde</td>
<td>SVN</td>
<td>2,645K</td>
<td>10,162K</td>
<td>527K</td>
<td>19.3</td>
<td>1997</td>
</tr>
<tr>
<td>gnome.org</td>
<td>SVN</td>
<td>1,284K</td>
<td>3,981K</td>
<td>1,412K</td>
<td>2.8</td>
<td>1997</td>
</tr>
<tr>
<td>Gcc</td>
<td>SVN</td>
<td>3,758K</td>
<td>4,803K</td>
<td>395K</td>
<td>12.2</td>
<td>1989</td>
</tr>
<tr>
<td>Eclipse</td>
<td>CVS</td>
<td>729K</td>
<td>2,127K</td>
<td>575K</td>
<td>3.7</td>
<td>2001</td>
</tr>
<tr>
<td>OpenJDK</td>
<td>Hg</td>
<td>32K</td>
<td>747K</td>
<td>60K</td>
<td>12.4</td>
<td>2008</td>
</tr>
</tbody>
</table>
Why important?

- Understand the entire open source ecosystem
- Radically improve software engineering by discovering effective code reuse patterns
- Conduct a census of software development

Discovery strategy

- Sites with many projects: e.g., SourceForge, GoogleCode, Savannah, repo.or.cz, github.com
- Ecosystems: e.g., Gnome, KDE, NetBeans, Mozilla, ...
- Famous: e.g., Mysql, Perl, Wine, Postgres, and gcc
- In wide use: e.g., git.debian.org
- Directories: e.g., RawMeat and FSF
- Published surveys of projects
What do we have now?
A cloud for project census

- Our Data Center keeps track of variation of both commercial and open source projects. This “universal” repository records data from:
  - Task tools
  - Code tools
  - Other tools
  - Email lists

- Infrastructure: Mem/Proc.: R910(4U), 64GbRAM, 16-cores X7550
### Projections in problem tracking systems

<table>
<thead>
<tr>
<th>Who</th>
<th>When</th>
<th>What</th>
<th>Removed</th>
<th>Added</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:mclasen@redhat.com">mclasen@redhat.com</a></td>
<td>2010-10-30 02:40:42 UTC</td>
<td>CC</td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="mailto:cosimoc@gnome.org">cosimoc@gnome.org</a></td>
<td>2010-11-01 12:02:51 UTC</td>
<td><strong>Attachment #173510</strong> Attachment status</td>
<td>none</td>
<td>reviewed</td>
</tr>
<tr>
<td><a href="mailto:cosimoc@gnome.org">cosimoc@gnome.org</a></td>
<td>2010-11-01 13:13:37 UTC</td>
<td><strong>Attachment #173629</strong> Attachment status</td>
<td>none</td>
<td>accepted-commit_now</td>
</tr>
<tr>
<td><a href="mailto:hadess@hadess.net">hadess@hadess.net</a></td>
<td>2010-11-01 13:18:39 UTC</td>
<td>Status</td>
<td><strong>UNCONFIRMED</strong></td>
<td><strong>RESOLVED</strong></td>
</tr>
<tr>
<td><a href="mailto:hadess@hadess.net">hadess@hadess.net</a></td>
<td>2010-11-01 13:18:42 UTC</td>
<td>Resolution</td>
<td></td>
<td><strong>FIXED</strong></td>
</tr>
<tr>
<td><a href="mailto:hadess@hadess.net">hadess@hadess.net</a></td>
<td>2010-11-01 13:18:44 UTC</td>
<td><strong>Attachment #173629</strong> Attachment status</td>
<td>accepted-commit_now</td>
<td>committed</td>
</tr>
<tr>
<td><a href="mailto:hadess@hadess.net">hadess@hadess.net</a></td>
<td>2010-11-01 13:18:44 UTC</td>
<td><strong>Attachment #173510</strong> Attachment is obsolete</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Projections in version control systems

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3.1120</strong> enndeakin%gmail.com</td>
<td>2008-10-24 04:17</td>
</tr>
<tr>
<td><strong>3.1119</strong> <a href="mailto:gavin@gavinsharp.com">gavin@gavinsharp.com</a></td>
<td>2008-10-19 00:58</td>
</tr>
<tr>
<td><strong>3.1118</strong> bzbarsky%mit.edu</td>
<td>2008-10-13 12:20</td>
</tr>
<tr>
<td><strong>3.1117</strong> dbaron%dbaron.org</td>
<td>2008-07-24 12:41</td>
</tr>
<tr>
<td><strong>3.1116</strong> <a href="mailto:gavin@gavinsharp.com">gavin@gavinsharp.com</a></td>
<td>2008-04-22 13:48</td>
</tr>
<tr>
<td><strong>3.1115</strong> jonas%sicking.cc</td>
<td>2008-04-18 10:20</td>
</tr>
<tr>
<td><strong>3.1114</strong> roc+%cs.cmu.edu</td>
<td>2008-04-18 02:21</td>
</tr>
<tr>
<td><strong>3.1113</strong> roc+%cs.cmu.edu</td>
<td>2008-04-17 04:52</td>
</tr>
<tr>
<td><strong>3.1112</strong> roc+%cs.cmu.edu</td>
<td>2008-04-17 01:15</td>
</tr>
<tr>
<td><strong>3.1111</strong> Olli.Pettay%helsinki.fi</td>
<td>2008-04-15 11:02</td>
</tr>
<tr>
<td><strong>3.1110</strong> dbaron%dbaron.org</td>
<td>2008-04-14 18:04</td>
</tr>
<tr>
<td><strong>3.1109</strong> jonas%sicking.cc</td>
<td>2008-04-08 20:20</td>
</tr>
<tr>
<td><strong>3.1108</strong> jonas%sicking.cc</td>
<td>2008-04-08 19:06</td>
</tr>
<tr>
<td><strong>3.1107</strong> reed%reedlodden.com</td>
<td>2008-04-08 11:52</td>
</tr>
</tbody>
</table>
- Developers plateau after 20 months in big project
- Centrality of tasks continues to increase over 3 years

Developer fluency: Achieving True Mastery in Software
-FSE 2010
Relative sociality impacts who will become LTCs

- Proportion of joiners who become LTCs (long term contributors)

Does initial environment impact the future of developers?
-ICSE 2011
About OW2’s Mature Transition Criteria

1) The project maturity classification should include a criteria related to the use of the Issue Tracking System. It's unlikely that mature projects would have no activity there. I propose to use, for example, OTC's, ReCO, LTCs, open defects, resolution time...

2) "Besides "Last commits", other aspects could be considered:
i, new committers recently joining in
ii, Long Term Contributors and their activities,
iii, core developers and their activities,
iv, experience profile currently,
v, the source code scale, the number of code lines added or deleted,
vi, the view of branches or sub-projects,
vii, the geographical distribution of developers."

3) It would also make sense to consider mechanisms that could attract or help novices to contribute, such as wiki, mail indexes...

4) Project history pages or project news?

-----Ma Xiujuan, Peking University
“Data and the Compute-Driven Transformation of Modern Science”

- exponential growth in the last couple of decades in data from scientific experiments and simulation, and in the computational power to deal with it, and the greater need for collaboration.

- No longer can science be done with just a single professor working with just a couple of students. ”Software is the Modern Language of Science.”