FLOSSMetrics: data about libre software development

Jesus M. Gonzalez-Barahona
(GSyC/LibreSoft, URJC)
http://identi.ca/jgbarah http://twitter.com/jgbarah
jgb@gsyc.es

Free/Libre and Open
Source Software Metrics

Master's Program on Development and Management of Free Software Projects
(Open Session) Vigo, Spain, March 17th 2011

FLOSSMetrics: base ideas

Libre software development:

- **Lots of opinions, few known facts**
- **Researcher-friendly: public data, reproducibility, validation of results, large samples**
- **Interest by volunteers and companies**

Main questions:

- **Can libre software development be improved?**
- **Can software engineering learn from libre software?**
- **Can projects better understand their processes and products?**

http://flossmetrics.org

FLOSSMETRICS goals

- Retrieval of data from (thousands of) libre software projects
- Analysis about actors, artefacts and processes involved in development
- Higher level studies: software evolution, human resources, effort estimation, productivity, quality, etc.
- Database available to other researchers, developers, etc.
- Providing tools for development follow-up
- Involvement with the libre software community

http://flossmetrics.org

Main results

- Huge database with factual details about libre software development (accessible to everyone)
- Higher level analysis and studies
- Sustainable platform for benchmarking and analysis
- Targeted reports (SMEs, industry, etc.)
- Focus on providing data and information that others can use for research, evaluation, follow-up

http://flossmetrics.org

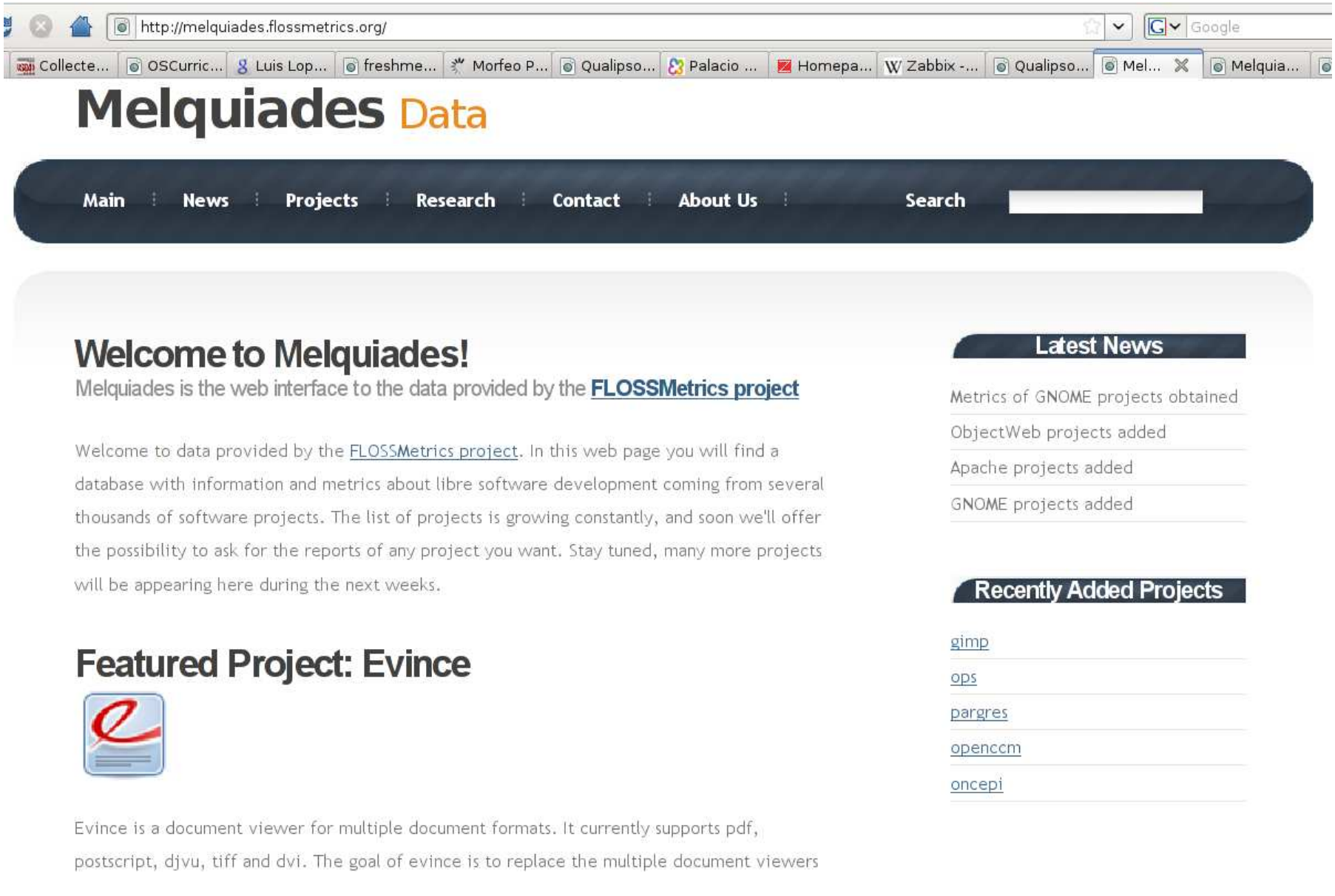
Partners

- Universidad Rey Juan Carlos (ES)
- University of Maastricht (NL)
- Wirtsshaftuniversitaet Wien (AT)
- Aristotle Univeristy of Thessaloniki (GR)
- Conecta s.r.l (IT)
- ZEA Partners (BE)
- Philips Medical Systems (NL)

Project funded by the European Commission
(FP6-IST programme)

Final results (2011)

- Full MySQL dumps for about 2,800 projects:
  - CVS, Subversion, git commit records (2,000)
  - Metrics (size, complexity) for source code (900)
  - Mailing lists main headers (400)
  - Issue tracking system (bug reports, etc.) (750)
- Focused report on SMEs, and focused studies
- Aggregated database (all projects together, separated dumps for SCM, mailing lists, issue tracking systems)
- Melquiades: `http://melquiades.flossmetrics.org`
- API for accessing the data


http://melquiades.flossmetrics.org/
Google
Collecte...
OSCurric...
Luis Lop...
freshme...
Morfeo P...
Qualipso...
Palacio ...
Homepa...
Zabbix -...
Qualipso...
Mel...
Melquia...
Melquiades Data
Main
News
Projects
Research
Contact
About Us
Search
Welcome to Melquiades!
Melquiades is the web interface to the data provided by the FLOSSMetrics project
Welcome to data provided by the FLOSSMetrics project. In this web page you will find a database with information and metrics about libre software development coming from several thousands of software projects. The list of projects is growing constantly, and soon we'll offer the possibility to ask for the reports of any project you want. Stay tuned, many more projects will be appearing here during the next weeks.
Featured Project: Evince
Evince is a document viewer for multiple document formats. It currently supports pdf, postscript, djvu, tiff and dvi. The goal of evince is to replace the multiple document viewers
Latest News
Metrics of GNOME projects obtained
ObjectWeb projects added
Apache projects added
GNOME projects added
Recently Added Projects
gimp
ops
pargres
openccm
oncepi

- Scheduled to analyze: 1091 projects.
- Of which some data is already available for: 954 projects.
  - Analyzed Source code management: 937
  - Analyzed Mailing lists repositories: 74
  - Analyzed Tracker systems: 0

Search: gn Show 10 entries

Name	SCM	MLS	TRK
flowdesigner	●		
gnat-gdb			
gnat-gps			
gnaural	●		
gnome-applets	●		
gnome-backgrounds	●		
gnome-control-center	●	◆	
gnome-desktop	●		
gnome-doc-utils	●		
gnome-games	●	◆	
gnome-icon-theme	●		
gnome-keyring	●		
gnome-keyring-manager	●		
gnome-mag	●		

Results

SCM repository analyses

Tool	Type	Date	Download
cvsanaly2	svn	2008-09-05	Donwload
cvsanaly2	svn	2008-08-30	Donwload
cvsanaly2	svn	2008-08-28	Donwload
cvsanaly2	svn	2008-06-13	Donwload
cvsanaly2	svn	2008-05-13	Donwload
cvsanaly2	svn	2008-03-31	Donwload
cvsanaly	svn	2008-02-14	Donwload

MLS repository analyses

Tool	Type	Date	Download
mlstats	mboxes	2008-08-28	Donwload
mlstats	mboxes	2008-06-13	Donwload
mlstats	mboxes	2008-05-13	Donwload
mlstats	mboxes	2008-03-31	Donwload
mlstats	mboxes	2008-02-14	Donwload

Tools

- LibreSoft Tools suite
  - CVSAnalY already work with CVS, SVN, git
  - CVSAnalY produces complexity metrics counts for each release of each file (C, C++, Java, Python)
  - Bicho: bug reports from SourceForge (Bugzilla coming soon)
  - MLStat: mailing lists, hiding real email addresses
- All of them integrated in Melquiades

http://melquiades.flossmetrics.org

http://forge.morfeo-project.org/projects/libresoft-tools/

Kinds of studies performed

- Evolution of projects (code, communities, responsiveness, etc.)
- Detection of deviations (disruptions in bug fixing practices)
- Human resources (what happens when the core team changes?)
- Effort and value estimation (how much time was put in a project?)
- Parameters to characterize the status of a project
- Sustainability conditions

But the imagination is the limit!

`http://flossmetrics.org`

CVSAnalY 2.0 features

- Browses an SCM repository producing a database with:
  - All metainformation (commit records, etc.)
  - Metrics for each release of each file
- Also produces some tables suitable for specific analysis
- Multiple SCMs: CVS, svn, git
- Whole history in the database, it's possible to rebuild the files tree for any revision
- Tags and branches support
- Option to save the log to a file while parsing
- Extensions system, incremental capabilities
- Multiple database system support (MySQL and SQLite)

CVSAnalY 2.0 Extensions

- Extension: a "plugin" for CVSAnalY
- Add information to the database, based in the information in the database and maybe the repository
- Usually: new tables for specific studies
- Simple example: commits per month per commiter
- Extensions add one or more tables to the database but they never modify the existing ones

Some CVSAnalY 2.0 Extensions

- **FileTypes: adds a table containing information about the type of every file in the database (code, documentation, i18n, etc.)**
- **Metrics: analyzes every revision of every file calculating metrics like sloc and complexity metrics (mccabe, halstead). It currently supports metrics for C/C++, Python, Java and ADA.**
- **CommitsLOC: adds a new table with information about the total lines added/removed for every commit**

MailingListStats

- Parses mbox information (RFC 822)
- Stores results (headers, body) in a MySQL database:
  - Sender, CCs, etc.
  - Time / Date
  - Subject
  - ...
- Incremental
- Can store multiple projects in a single database
- Provides also some stats at the end

Bicho

- Aimed at parsing issue tracking systems.
- Results stored in a MySQL database.
- Information about each issue (bug report), and its modifications.
- Currently supports:
  - SourceForge (HTML parsing, old interface, adapting to the new one)
  - BugZilla: GNOME, KDE, adapting to others

Demo: queries on the aggregated SCM database

- Aggregated SCM database of June 2nd 2009
- 518 projects
- 2,987,961 file revisions
- 2,440,127 commit records
- 20,583 committers

```
zcat fm3_aggregatedb_scm_20090602T12\:06\:00.sql.gz  |
   mysql --user=jgb --password=XXXXX fm3_aggregatedb_scm
SELECT COUNT(*) FROM projects
SELECT COUNT(*) FROM files
SELECT count(*) FROM scmlog
SELECT count(*) FROM people
```

Demo: some queries

- **Total number of SLOC for all revisions of all files**
- **Total number of SLOC and file revisions for C files**
- **Total number of file revisions and files for C files**

```
SELECT SUM(sloc) FROM metrics

SELECT SUM(sloc), COUNT(file_id) FROM metrics WHERE lang="ansic"

SELECT COUNT(file_id), COUNT(DISTINCT(file_id)) FROM metrics
WHERE lang="ansic"
```

Demo: some queries (2)

- **Number of commits by committer and quarter**
- **Same, for those committers with at least 10 commits per quarter, ordered by time**

```
select year(date), quarter(date), committer_id, count(id)
from scmlog group by year(date), quarter(date), committer_id
```

```
select year(date), quarter(date), committer_id, count(id)
from scmlog group by year(date), quarter(date), committer_id
having count(id) > 10
order by year(date), quarter(date), count(id) desc
```

Demo: some queries (3)

- **Commits by committer and quarter for Evolution**
- **Commits per quarter for Konqueror**

```
select year(date), quarter(date), committer_id, count(id)
from scmlog
where project_id = 119
group by year(date), quarter(date), committer_id
order by year(date), quarter(date), count(id) desc

select year(date), quarter(date), count(id)
from scmlog
where project_id = 237
group by year(date), quarter(date)
order by year(date), quarter(date) desc
```

Demo: some queries (4)

- **Create a "history" view for Evolution**
- **Commits per committer and quarter**

```
create view history (year, quarter, committer_id, commits)
as select year(date), quarter(date), committer_id, count(id)
from scmlog
where project_id = 119
group by year(date), quarter(date), committer_id
order by year(date), quarter(date), count(id) desc

select * from history
```

Demo: some queries (5)

- On the "history" view for Evolution
- Commits per committer and quarter for committers active in 2001
- Commits per quarter for committers active in 2001

```
select * from history
where committer_id in
  (select committer_id from history where year=2001)
order by year, quarter, commits

select year, quarter, sum(commits) from history
where committer_id in
  (select committer_id from history where year=2001)
group by year, quarter
```

Retrieving information: general problems

Diversity:

- **Kinds of forges: difficult to automate**
- **Kinds of projects: not all projects in SF are relevant**
- **Sources for same project: forge(s), distributions...**

Missing information:

- **Hidden information (eg: mail headers)**
- **Lost information (eg: transition from CVS to SVN)**
- **Bugs and errors (eg: old locks in SCM)**

Stress to projects infrastructure!!

Retrieving information: SCM problems

- Different systems (CVS, Subversion, git, Bazaar, Mercurial, etc.)
- Different models (file-based, commit-based, distributed)
- Bots performing commits
- Large transitions don't preserve information
- Performance issues (systems poorly designed for massive retrieval)

But at least we have facilities for incremental retrieval

Retrieving information: BTS problems

- Different systems (Bugzilla, SourceForge, GForge, trac, Launchpad, etc.)
- Different models (bug cycle, bug report parameters, etc)
- Different uses (issue tracker, only bugs, scheduler, etc.)
- Bots acting on bug reports
- Lack of facilities for incremental retrieval
- Performance issues (systems not really designed for massive retrieval)

Retrieving information: Mailing lists problems

- Different systems (usually accessible only through HTML)
- Partial information (missing headers)
- Bots sending email (eg: commit messages)
- Spam (mixed with real messages)

But email messages are pretty uniform in format

Retrieving information: All together

- How to track actors and products:
  - Different repositories of the same project
  - Different projects
- SourceForge helps a bit (?)
- Massive information (when dealing with 1,000s projects)
- Exchange formats (for third parties and reproduction)
- Tracking information (where did this commit record came from?):
  - Repositories change
  - Retrieval tools change
  - Errors do occur

Interested?

- Detailed description of work available from the website
- All the software used is libre software
- Tell us about your pet project, we can analyze it
- Interested in knowing how this is useful for you: provide feedback about your interests, needs
- Open to contributions: plug-ins for new analysis tools
- Willing to collaborate with projects

Very interested in feedback from researchers!!

http://flossmetrics.org

http://forge.morfeo-project.org/projects/libresoft-tools