Use case of source code clones detection

Analysis of reused code between to FLOSS projects using FLOSS tools

Luis Cañas-Díaz

lcanas@libresoft.es

Linux Tag 2012, Berlin, May 23rd, 2012

GSyC/LibreSoft

- Research group at Universidad Rey Juan Carlos
- About 20 persons, including students
- Focus on FLOSS (free, libre, open source software)
- One of the main research lines:
    - understanding FLOSS development
    - quantitative, empirical approach
    - based on data retrieval from FLOSS development repositories
- Participating in several R&D projects

Bitergia: an spin-off

- Company starting operations in June 2012
- Building on the experience of LibreSoft
- Offering professional products and services
- Focused on:
    - Metrics about software developent (including community metrics)
    - Specialized support for development forges (including metrics for projects)
- "How to understand risks associated to open source communities" by Daniel Izquierdo on Saturday

`http://bitergia.com`

Introduction

- Provincial Council of A Coruña
- gisEIEL and gvSIG-EIEL , both with similar features

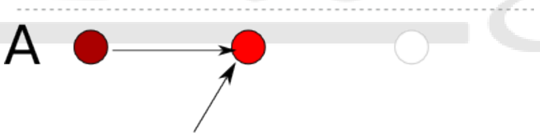
Introduction: gisEIEL and gvSIG-EIEL

- gisEIEL is the geographic information system used by the technical staff of the Provincial Council of A Coruña and the municipalities
- gvSIG-EIELStack includes three gvSIG extensions that provide several functionalities to work with the EIEL (Survey on Infraestructure and Local Facilities)

Introduction: project A and project B

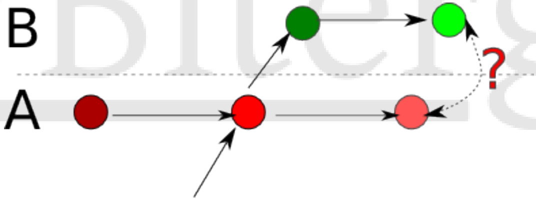
- gisEIEL = project A
- gvSIG-EIEL = project B

Introduction: the history



- gisEIEL (project A):
    - created in 2000 and funded by the Provincial Council of A Coruña
    - was released in 2004 as FLOSS based on gvSIG 1.0

Introduction: the history



- gvSIG-EIEL (project B):
    - years later the Provincial Council of Pontevedra funded the creation of a similar a application (instead of using the project A)
    - project B was released with very similar functionality

Introduction: our client

- Our client was in charge of maintaining the project A
- Interested in:
  - finding out whether a merge is feasible
    - amount of reused code in B
  - how the code is being reused
    - licensing and copyright issues
  - study the functionality

Methodology

- Data analysed is publicly available (replicability)
- Done with FLOSS tools

Methodology

- Retrieval of the source code to be analysed
- Selection of tools to get information from source code
- Process the raw data
- Identification of relevant information

Methodology: sources

- Project A: Snapshot downloaded from 1 SVN repository
- Project B: Snapshots downloaded from 6 Git and 2 SVN repositories
- No feedback from developers

Methodology: CCFinder anc Cloc

CCFinder

- http://www.ccfinder.net/
- CCFinder allows to match similar parts of the code
- Works at token level
- Must be carefully configured

Cloc

- http://cloc.sourceforge.net
- Calculates the SLOC
- Support for 86 programming languages

Methodology: Ninka and grep

Ninka

- http://ninka.turingmachine.org/
- Lightweight license identification tool for source code

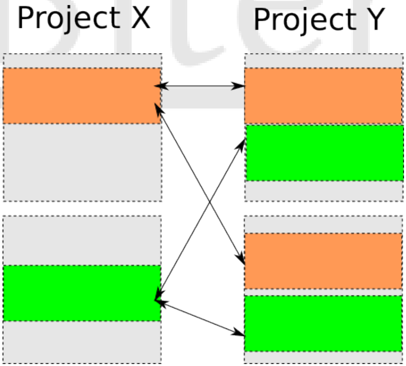
Grep

- Well know command line in the UNIX environment
- Searches text strings using regular expressions

Methodology: Process the raw data from CCFinder

clone id	file id.tokens	file id.tokens
16359	476.1119-1177	2093.644-702
16359	476.1119-1177	2093.749-807
16359	476.1119-1177	2093.889-947
16359	476.1119-1177	2093.1034-1092
16359	476.1119-1177	2093.1181-1239
1207	476.1259-1310	2093.1324-1375
36	476.37-149	2094.37-149
1831	476.260-326	2094.221-287

Methodology: How much code in common?



Results: file by file

- One of the files of the project A:

File name	ExportMapTo.java
Cloned files	3
SLOC	569
License	GPLv2
Copyright	Copyright (C) 2009 Deputación de A Coruña

Results: file by file

- For one file in A we got the clones below in B
- Have a look at the license and copyright!

File name	%	SLOC	license	copyright
ExportSeveralTo.java	43 %	244	*None*	*None*
StopEditingToShp.java	28 %	159	*None*	*None*
ExportTo.java	47 %	267	*None*	*None*

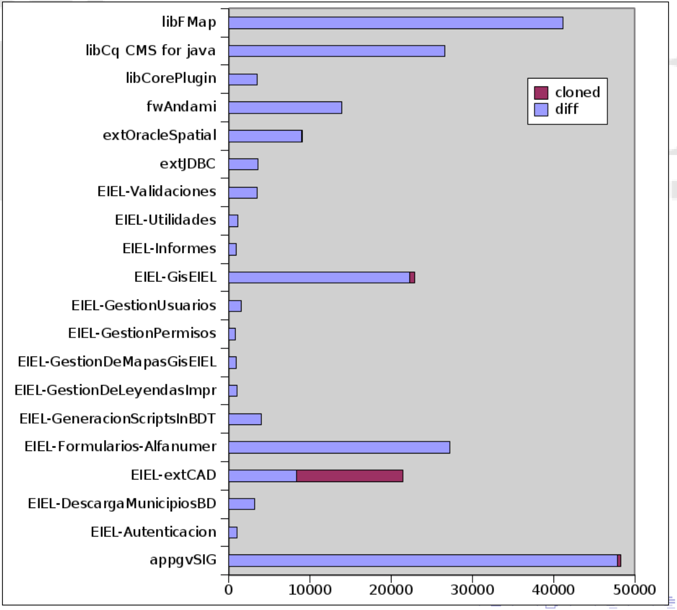
Results: A project vs. B project (1/3)

Module of project A	SLOC	similar SLOC	%
appgvSIG	48279	483	1
EIEL-Autenticacion	1062	0	0
EIEL-DescargaMunicipiosBD	3142	0	0
EIEL-extCAD	21423	13068	61
EIEL-Formularios-Alfanumer	27224	0	0
EIEL-GeneracionScriptsInBDT	3992	0	0
EIEL-GestionDeLeyendasImpr	980	0	0
EIEL-GestionDeMapasGisEIEL	936	0	0
EIEL-GestionPermisos	776	0	0
EIEL-GestionUsuarios	1517	0	0

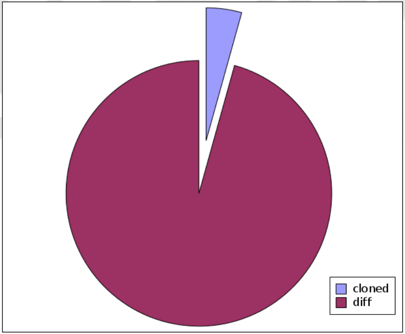
Results: A project vs. B project (2/3)

Module of project A	SLOC	similar SLOC	%
EIEL-GisEIEL	22906	687	3
EIEL-Informes	935	0	0
EIEL-Utilidades	1146	23	2
EIEL-Validaciones	3487	0	0
extJDBC	3600	36	1
extOracleSpatial	9034	90	1
_fwAndami	13886	0	0
libCorePlugin	3510	35	1
libCq CMS for java	26617	0	0
libFMap	41159	0	0

Results: A project vs. B project (3/4)



Results: A project vs. B project (4/4)

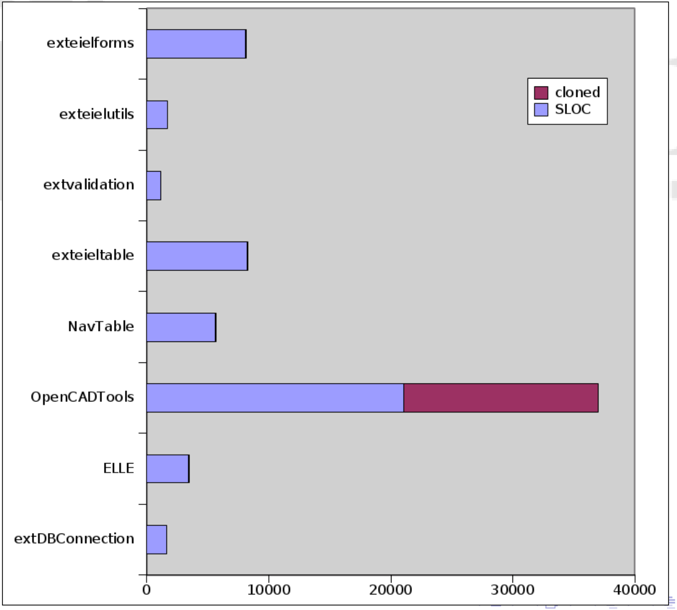


- 6 % of the A's code was reused by project B (14K out of 319K SLOC)

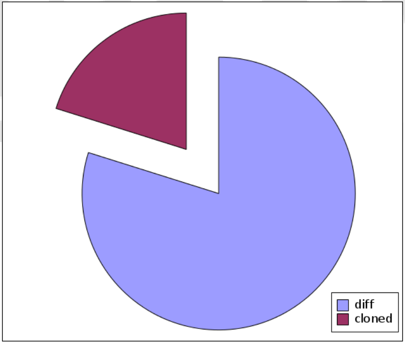
Results: project B vs. project A (1/3)

Module in B	SLOC	SLOC similar	%
extDBConnection	1648	0	0
ELLE	3459	35	1
OpenCADTools	36974	15899	43
NavTable	5685	57	1
exteieltable	8311	83	1
extvalidation	1160	0	0
exteielutils	1711	0	0
exteielforms	8185	82	1

Results: project B vs. project A (2/3)

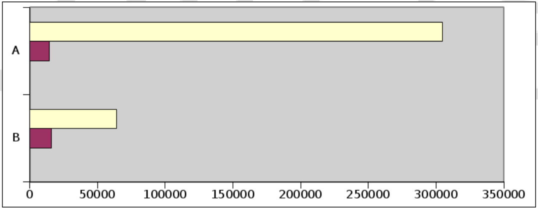


Results: B project vs. A project (3/3)



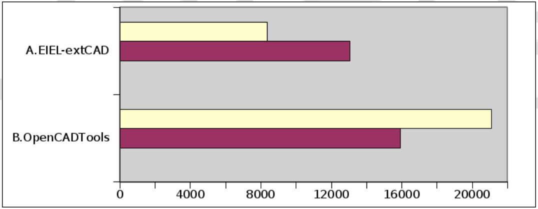
- B reused around 20 % of its code from A (16K out of 80K SLOC)

Final conclusions (1/3)



- 20 % of the code in project B was reused from A
- 6 % of the A's code is reused in project B

Final conclusions (2/3)



- most of the code reused by B is part of a single module (*OpenCADTools*). This module reused 43 % of its code from another module from A called *EIEL-extCAD*

Final conclusions (3/3)

- 91 % of the files reused by B did not contain the original copyright holder
- early versions of A reused code from gvSIG project and they did not contain the original copyright holder either (fixed in latest versions of A)

Your time!

any questions?

Thank you! / ¡Gracias!

contact me at lcanas@libresoft.es